Appendix C to the 2015 Hawaii Growth Model Report
Investigation of Potential Ceiling and Floor Effects.

Adam R. VanIwaarden
Damian W. Betebenner
National Center for the Improvement of Educational Assessment (NCIEA)

December 2015

1 Introduction

In the 2014-2015 academic year, Hawaii transitioned from the Hawaii State Assessment (HSA) to the Smarter Balanced Assessment (SBA). The transition included numerous changes to the assessment system including the incorporation of new performance standards and moving to a vertical scale. As other states have gone through similar assessment transitions in 2014-2015, many have observed ceiling and floor effects in the new assessments (i.e. a relatively large proportion of students scoring at/near the scale extremes). This has occurred despite purported improvements in assessment qualities that should prevent these effects (e.g. adaptive tests). Regardless of the source of assessment ceilings/floors, they can make the Student Growth Percentile (SGP) estimates questionable.

Although very similar in nature, ceiling effects are somewhat more problematic than floor effects because consistently highest achieving students receive lower than expected growth percentiles and therefore the students are negatively impacted. Conversely, consistently lowest achieving students have higher estimated SGPs than would be expected. Although this could be interpreted as a positive impact on these students by giving them higher SGPs, it can also conceals unacceptably low growth.

Essentially these problems are caused by the way in which a "percentile" is defined to begin with, and the inability of the assessments (and therefore the SGP model) to make granular distinctions between kids who score at the extremes of the test year after year. As an example, if a group of students were tested with a relatively easy test and 20% of the students had a perfect score, these students would be defined as being in the 80^{th} percentile because they scored *higher than* 80% of their peers. This is somewhat misleading however, because their score was also *equal to* or greater than 100% of their peers and so could potentially be defined as achieving at the 99^{th} percentile. To extend this heuristic from achievement to growth, if 50% of those top scorers also scored perfectly on the next test, we might estimate that they had 50^{th} percentile growth. Although there is nothing *technically* incorrect about this estimate because their growth is fairly typical for their specific norm group, it is an inadequate or unsatisfactory assessment of their growth because they have consistently attained at the highest level.

Given these impacts and the difficulty in detecting them given traditional SGP diagnostic tools, the Center for Assessment has added "Ceiling/Test Effects" indicators to the SGP model goodness of fit plots, as well as providing all clients even more rigorous diagnostics through this appendix to the annual technical report. This report includes:

1. Plots of the scale score distributions for the current and prior years, which may provide an indication of whether a ceiling or floor is present in either (or both) the current or historical data.
2. Box plots showing the range and distribution of SGPs for *only* the highest and lowest achieving students in the current year.

2 Prior and Current Year Score Distributions

The marginal and conditional distributions of scale scores can serve as a preliminary indicator of potential ceiling or floor effects in the calculation of student growth percentiles. Some minor problems could present themselves if these characteristics are present in either prior or current year scores, and are particularly likely when present in both. The plots below depict distributions for the current year and the most recent prior year used in the SGP calculations. The marginal (individual or univariate) distributions for each year are shown in the first subsections below, followed by the conditional (joint) distributions.

2.1 Marginal Distributions

Generally there is evidence of ceiling effects across all grades and in both subjects (although it is more prominent in the reading tests). These effects, which appear as fat, truncated tails at the extreme right side of the distributions, are observed primarily in the current (SBA) scores. The prior (HSA) score distributions have long tails at the upper end of the distributions, suggesting that there may be issues with students consistently achieving maximum scores, which would then translate to ceiling effects in growth analyses. This conforms to what was observed first in the SGP model goodness of fit plots (see 2015 technical report Appendix A), and confirmed again in the conditional density and SGP distribution box plots in subsequent sections.

2.1.1 Reading

Fig. C.1: Marginal distributions of prior scale scores: Reading.

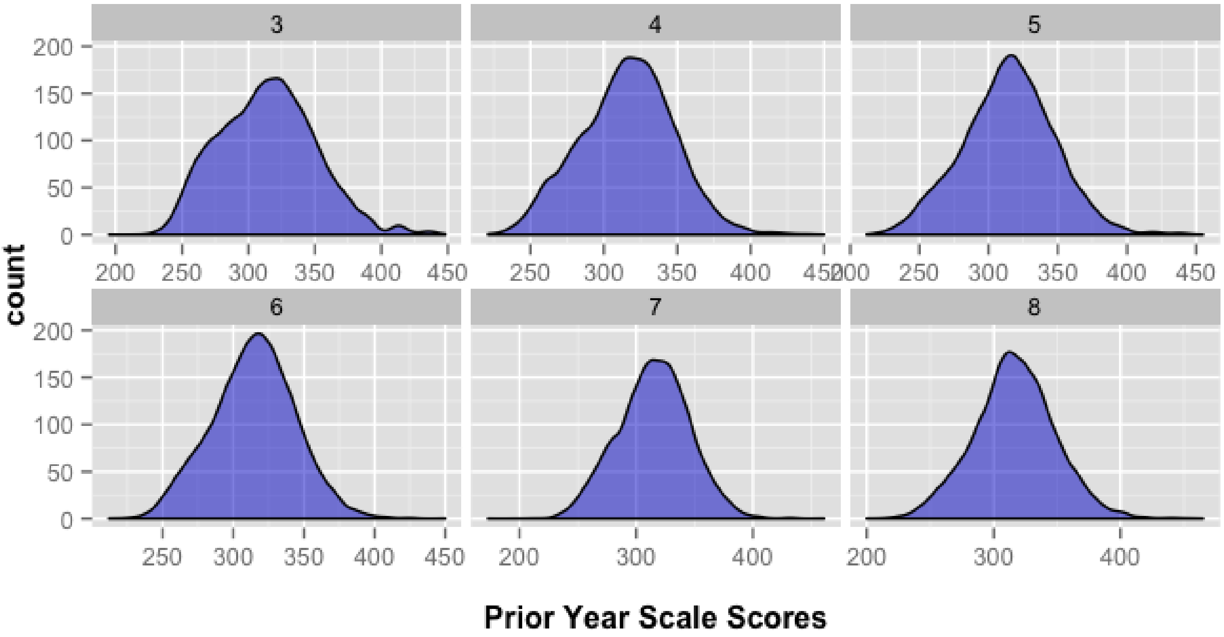
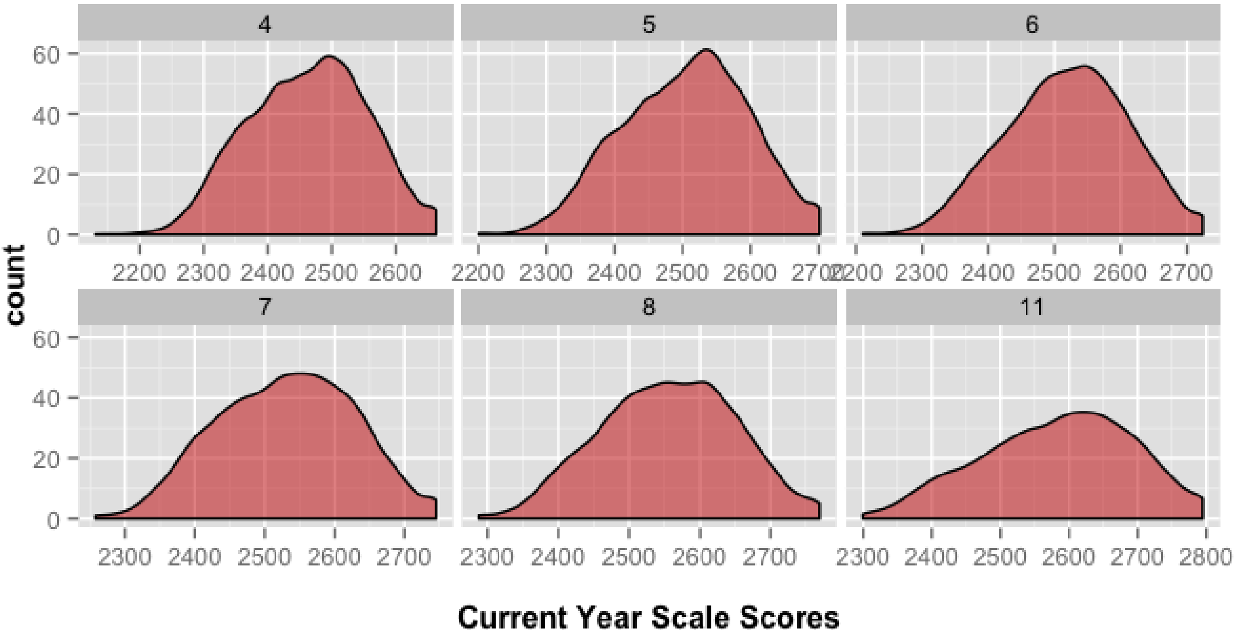


Fig. C.2: Marginal distributions of current scale scores: Reading.



2.1.2 Mathematics

Fig. C.3: Marginal distributions of prior scale scores: Mathematics.

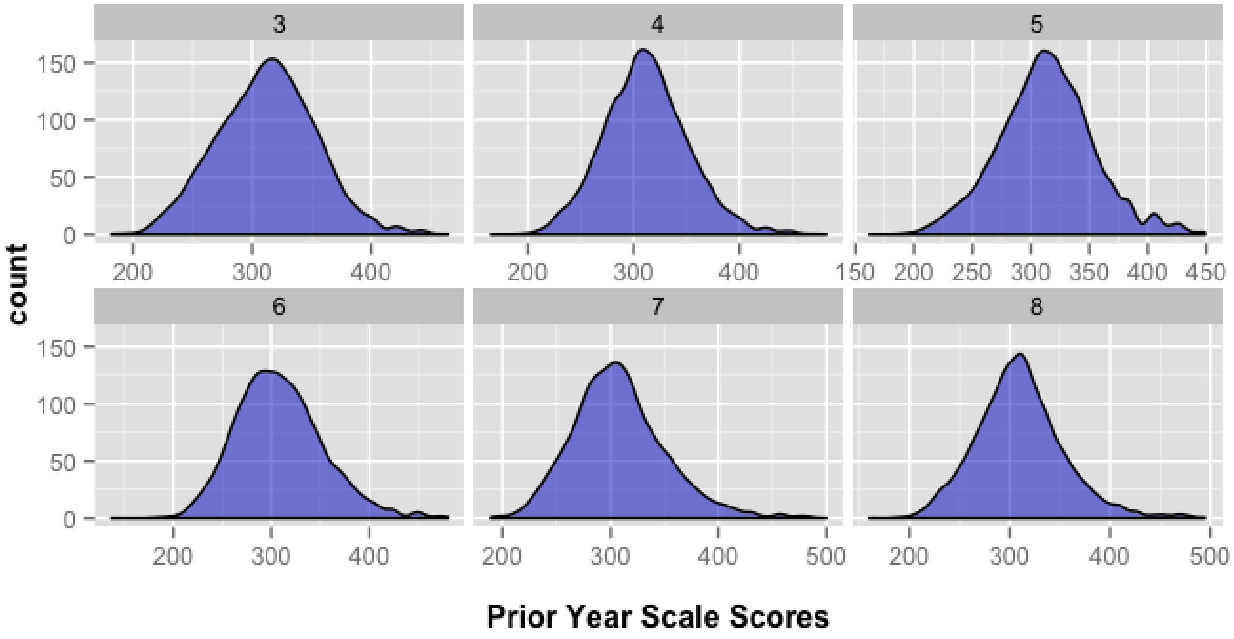
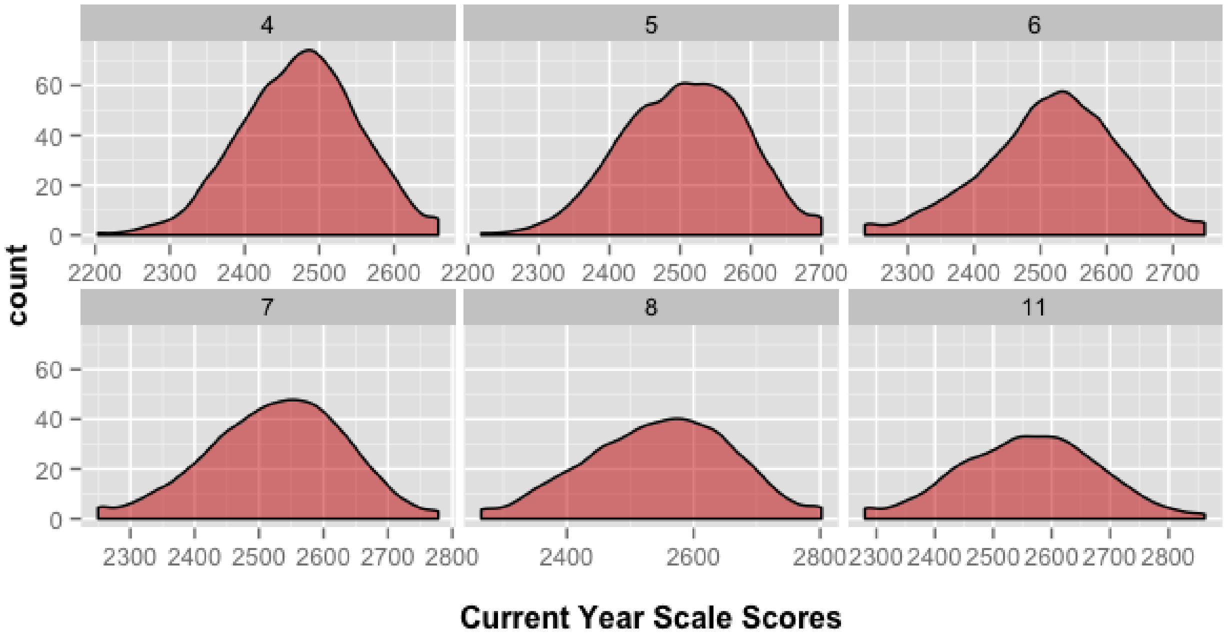


Fig. C.4: Marginal distributions of current scale scores: Mathematics.



2.2 Conditional Distributions

The marginal density plots provide a limited amount of information, particularly for the potential for ceiling/floor effects in the calculation of ***growth***. In order to provide a more nuanced view of the relationship between the prior and current scale scores, the following plots depict the conditional (joint) distributions for each content area and grade level. These plots start with a basic scatter plot of each student's scores, and on top of this is layered 1) **green contour lines** to provide a sense of joint density, 2) a **magenta non-linear line** identifying the bivariate relationship between prior and current scores, and 3) **rug plots** that describe the marginal distributions (as above, the prior scores are blue and current scores are red).

For the 2014 and 2015 Hawaii data, we again see cause for concern for ceiling effects in nearly all content area and grade combinations. Additionally, we also see some concern for floor effects in the middle-grade Mathematics analyses. These effects present themselves as dark shaded points in the extreme top-right or bottom-left corners of the plots. This suggests that staying at the extremes from year to year is not uncommon, which may lead to odd growth estimates for these chronically high/low achieving students. The 6^{th} to 8^{th} Grade Mathematics plots in Figure C.8 shows indications consistent with both ceiling and floor effects.

2.2.1 Mathematics

Fig. C.7: Conditional distribution(s) of current and prior scale scores: Mathematics.

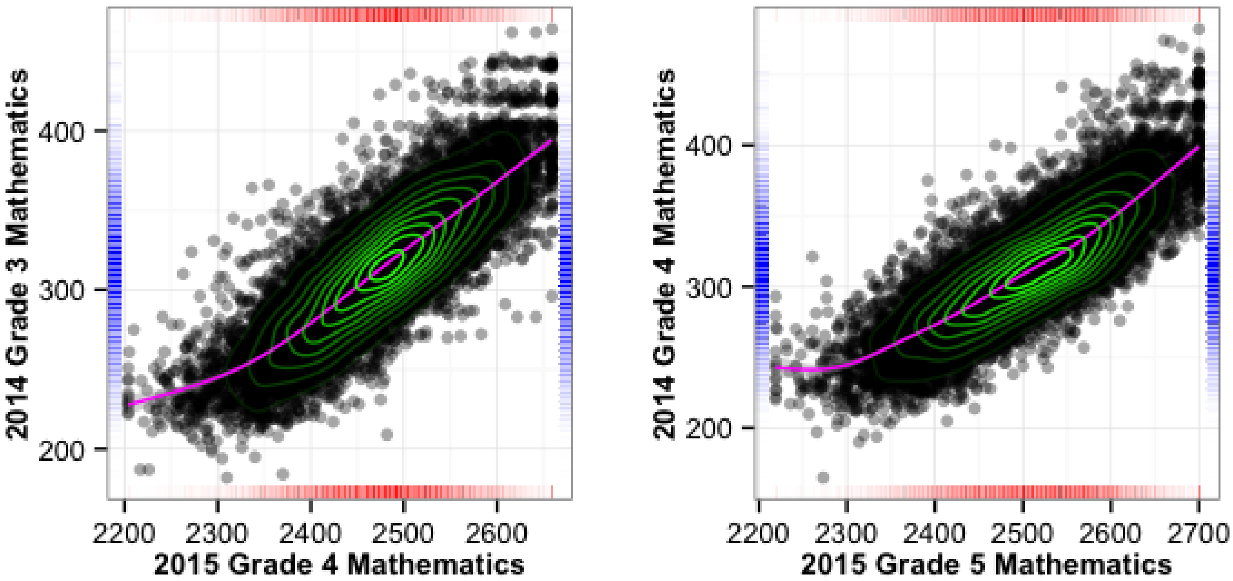
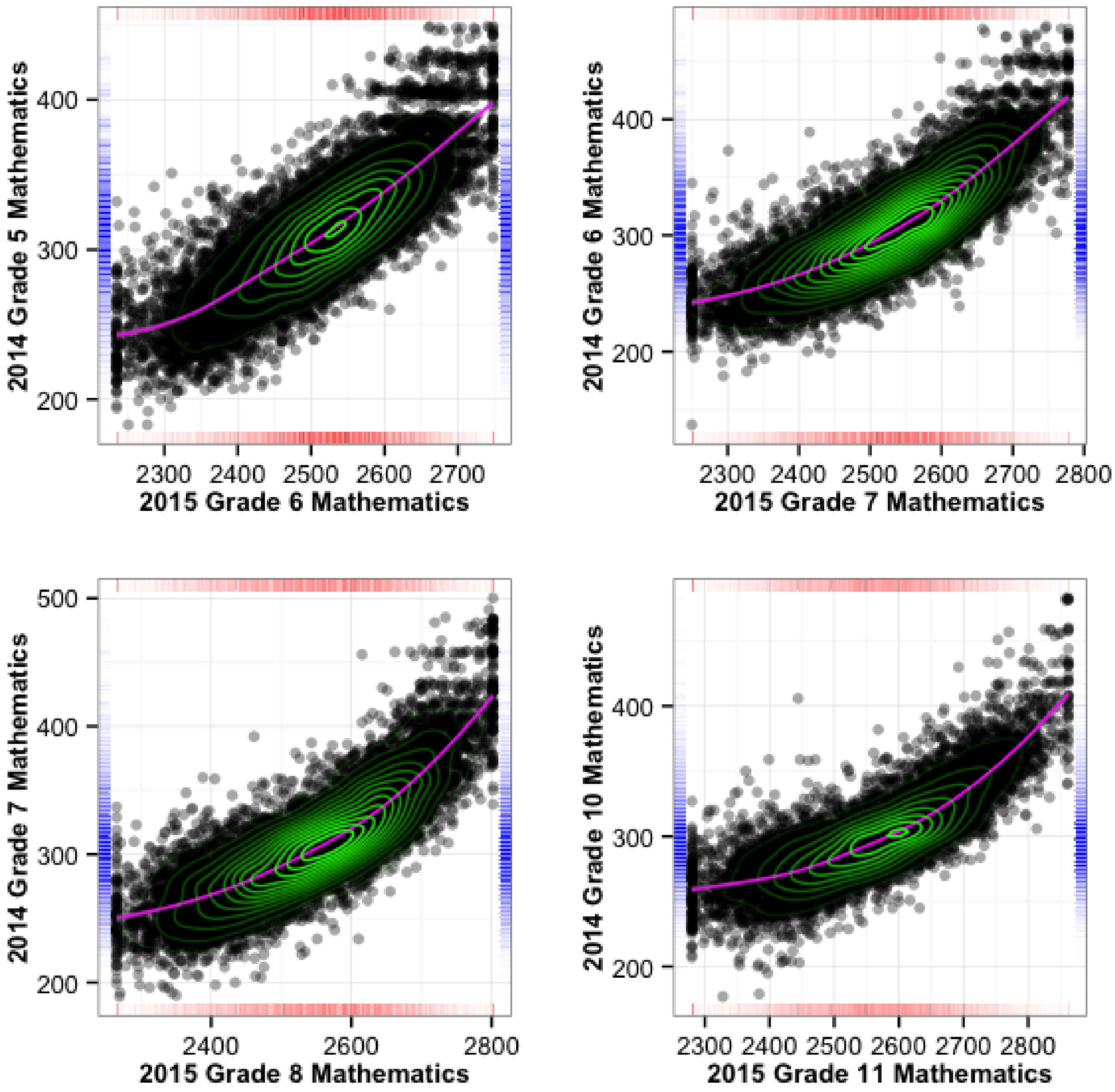


Fig. C.8: Conditional distribution(s) of current and prior scale scores: Mathematics *Continued.*



2.2.2 Reading

Fig. C.9: Conditional distribution(s) of current and prior scale scores: Reading.

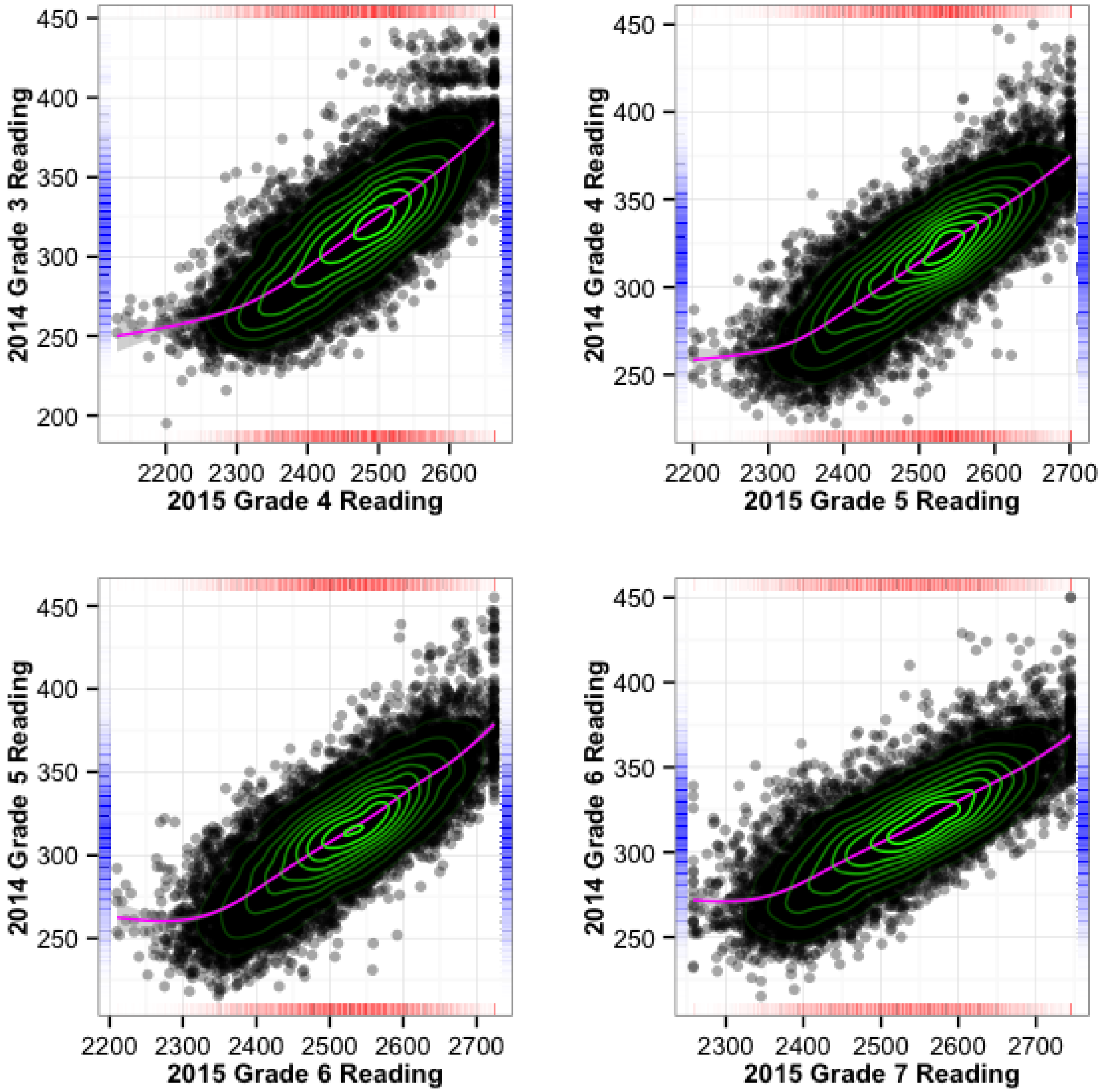
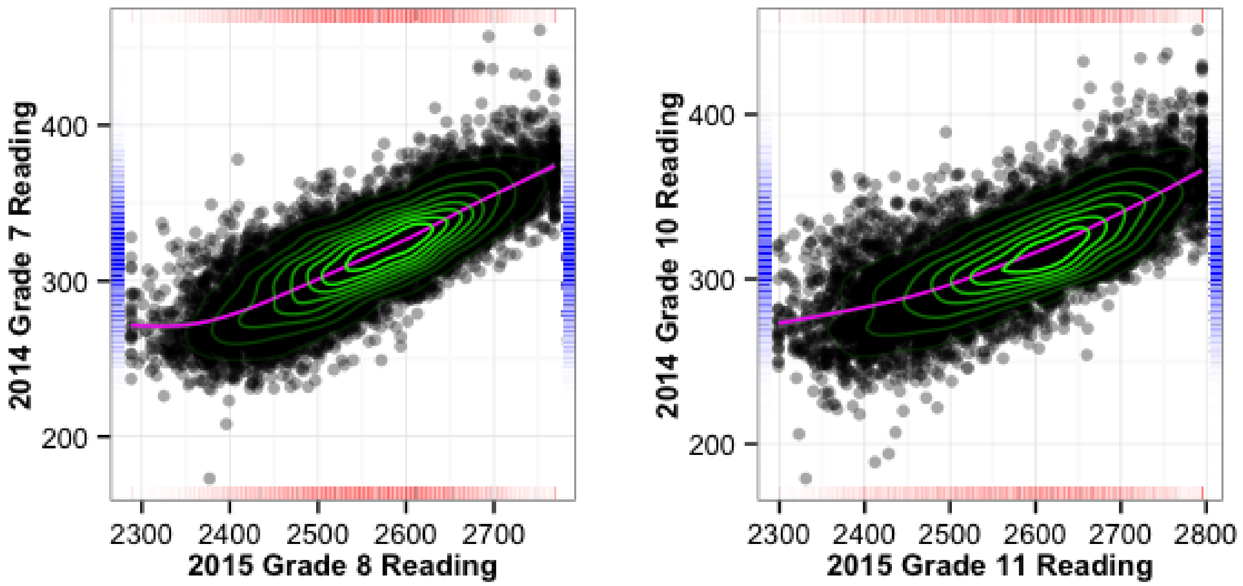


Fig. C.10: Conditional distribution(s) of current and prior scale scores: Reading *Continued.*



3 SGP Ranges for the Highest and Lowest Achieving Students

In order to isolate the impact of assessment ceilings/floors on student growth percentile (SGP) calculations, the following section provides box plots of the distribution of SGPs for the highest and lowest achieving students. We are specifically interested in the SGPs for students scoring at the highest/lowest obtainable scale score (HOSS/LOSS - i.e. the actual ceiling/floor). However, in order to assure that an adequate number of students are included in these plots, the first set of plots in each subsection uses, at a *minimum*, the highest/lowest 50 scores. Note that this roughly corresponds to the number of students used in the SGP model goodness of fit plots, and this is why these plots are provided as a starting point for this part of the investigation. If all 50 students have only a single scale score value (i.e. the HOSS/LOSS), then **all** students with this score are included. Consequently, the number of students included in each box plot may be greater than 50 (the exact number is shown at the margins in red text).

The second set of box plots isolate ***only*** those students scoring the HOSS/LOSS. These plots may then incorporate a varying number of students depending on the prevalence of a ceiling/floor in the current year test scores.

The box plots provide several descriptive statistics. The dark line within the box marks the *median* SGP, while the ends ("hinges") of the boxes correspond to the first and third quartiles (the 25^{th} and 75^{th} percentiles). The upper whisker extends from the hinge to the highest value that is within 1.5 × IQR of the hinge, where IQR is the inter-quartile range, or distance between the first and third quartiles. The lower whisker extends from the hinge to the lowest value within 1.5 × IQR of the hinge. Data beyond the end of the whiskers are outliers and plotted as points. Evidence of a *lack* of either a ceiling or floor effect would be to have all high achieving students with SGPs near 99 and all low achieving students with SGPs near 1. That is, the desired visual evidence is a solid line at SGP = 99/1.

In the 2015 Hawaii SGP analyses, we see ceiling effects in all grades and subjects. There is also evidence of floor effects in the middle-grade mathematics analyses as well.

Fig. C.14: SGP distributions for highest and lowest 0.5 percent of scale scores by content area and grade level.

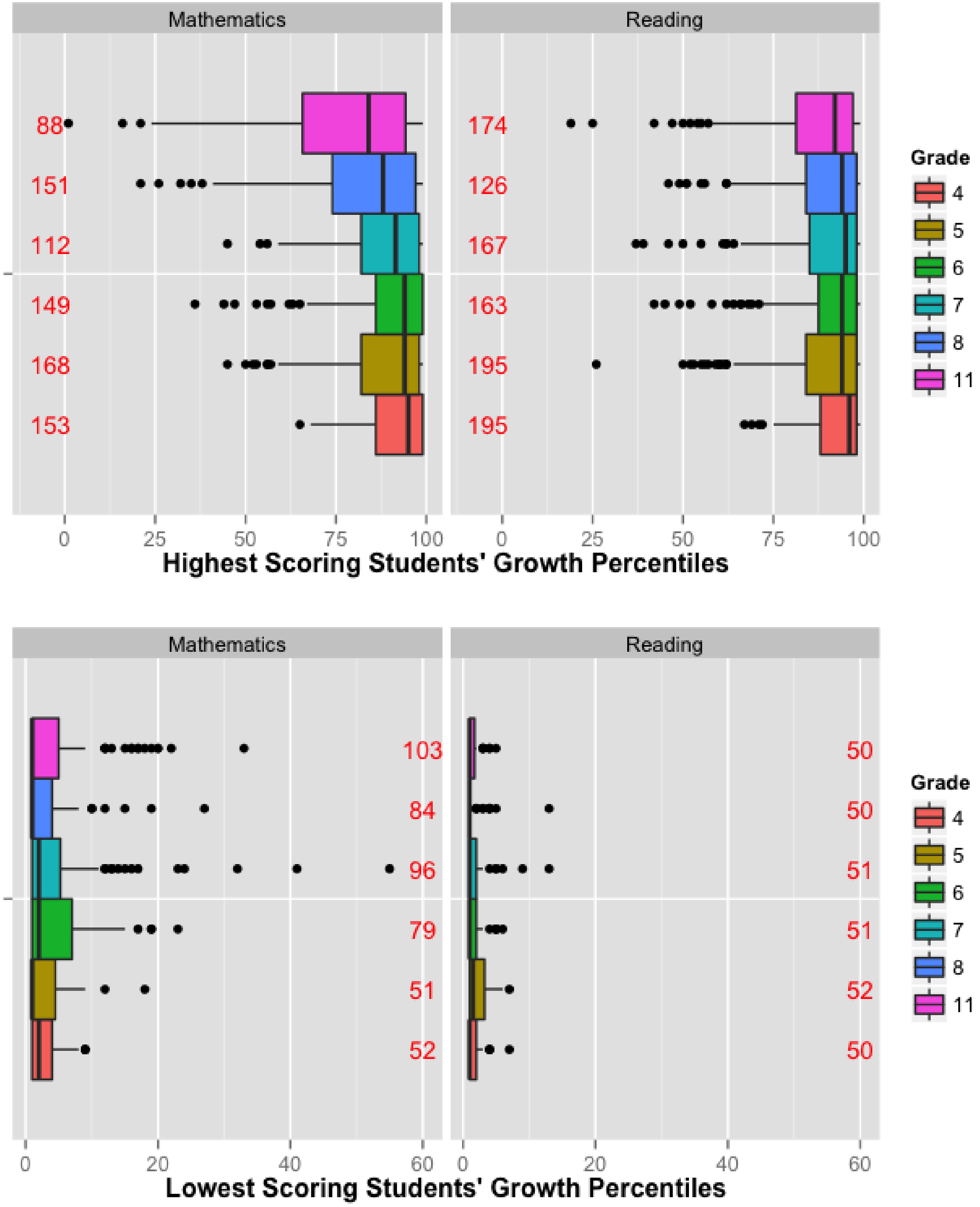
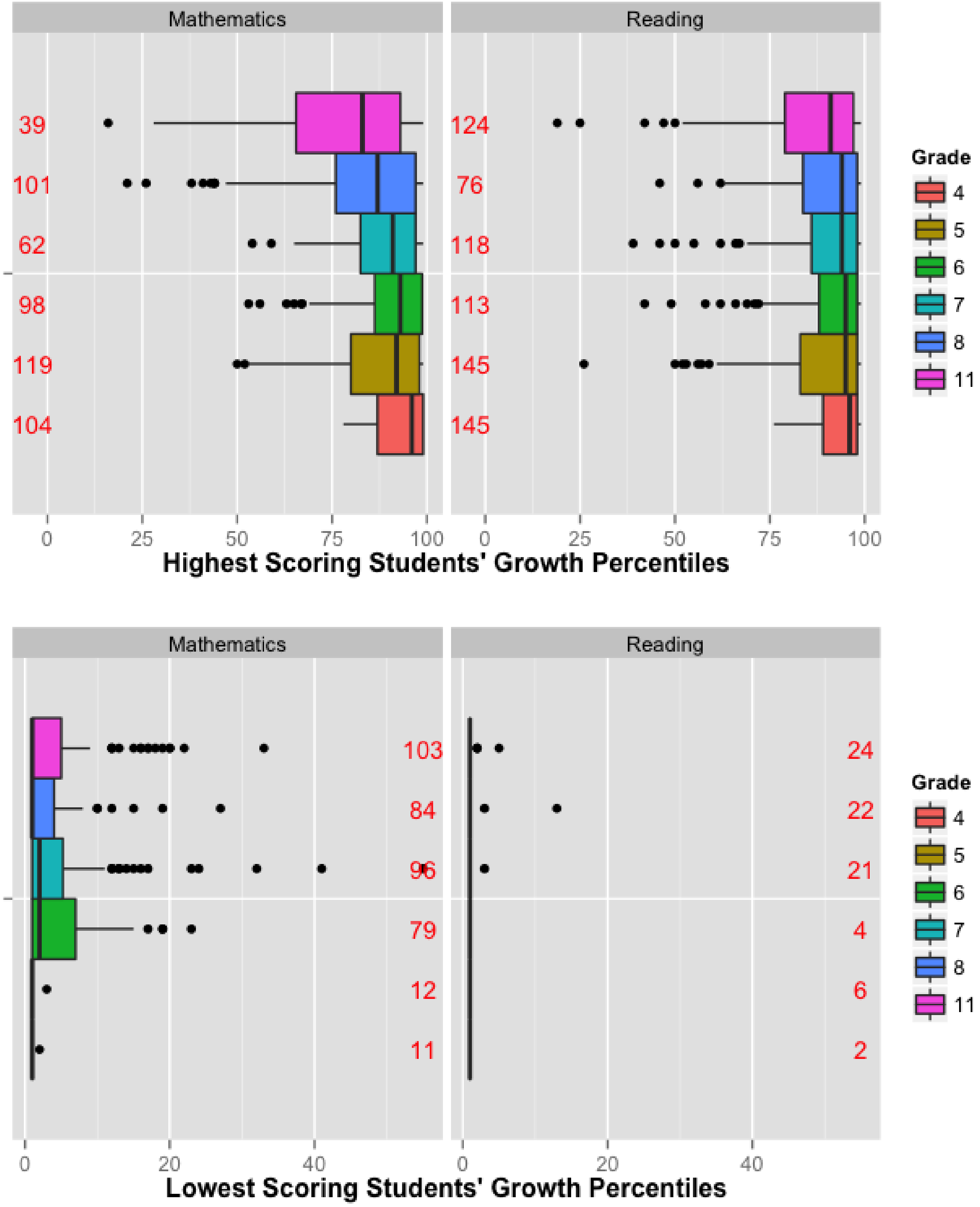


Fig. C.15: SGP distributions for the HOSS and LOSS scores by content area and grade level.



4 Discussion

Overall there is evidence of ceiling effects in the 2015 Hawaii SGP analyses in both Mathematics and Reading across all grades. There is also evidence of more minor floor effects in the middle-grade Mathematics analyses. When ceiling or floor effects are encountered, there are several ways in which they can be "corrected" manually or analytically. These include (but not limited to):

1. Convert all students scoring at the HOSS (LOSS) to 99 (1).
2. Run SGP analyses with more granular scores. For example, many tests that use Item Response Theory (IRT) to analyse test results provide scaled scores that enforce an artificial ceiling (floor), but also have more granular achievement scores available (IRT θ estimates).
3. Leave the results without a correction.