Appendix C to the 2015 Utah 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)

November 2015

1 Introduction

In the 2013-2014 academic year, Utah transitioned from its previous Utah Criterion Referenced Tests (CRT) to the Student Assessment of Growth and Excellence (SAGE). 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 80th percentile because they scored higher than 80% of their peers. This is somewhat misleading however, because their score was *equal to* or greater than 100% of their peers and so could be also defined as achieving at the 99.99^{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 50th 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. End-of-grade test (EOGT) content area plots are disaggregated by grade, while end-of-course test (EOCT) subject plots are not.

2.1 Marginal Distributions

Generally there is little evidence of potential issues with the Utah SAGE tests in either the prior (2014, shown in blue) and current (2015, shown in red) scale scores for the end of grade tests (EOGT). There are some indications of minor floor effects in the Mathematics assessments, particularly in the 2014 transition year. The end of course test (EOCT) distributions are not concerning for the content areas within the sciences domain, but there is evidence of a floor effect in the Secondary Mathematics assessments. These effects (which appear as spikes at the extreme left side of the distributions) are more prominent in the prior scores, but are notable in the current year as well. This conforms to what was observed first in the SGP model goodness of fit plots (see 2015 technical report Appendix A), and conditional density and SGP distribution box plots in subsequent sections confirm this floor effect.

2.1.1 ELA

Fig. C.1: Marginal distributions of prior scale scores: EOGT ELA.

Prior Year Scale Scores

Fig. C.2: Marginal distributions of current scale scores: EOGT ELA.
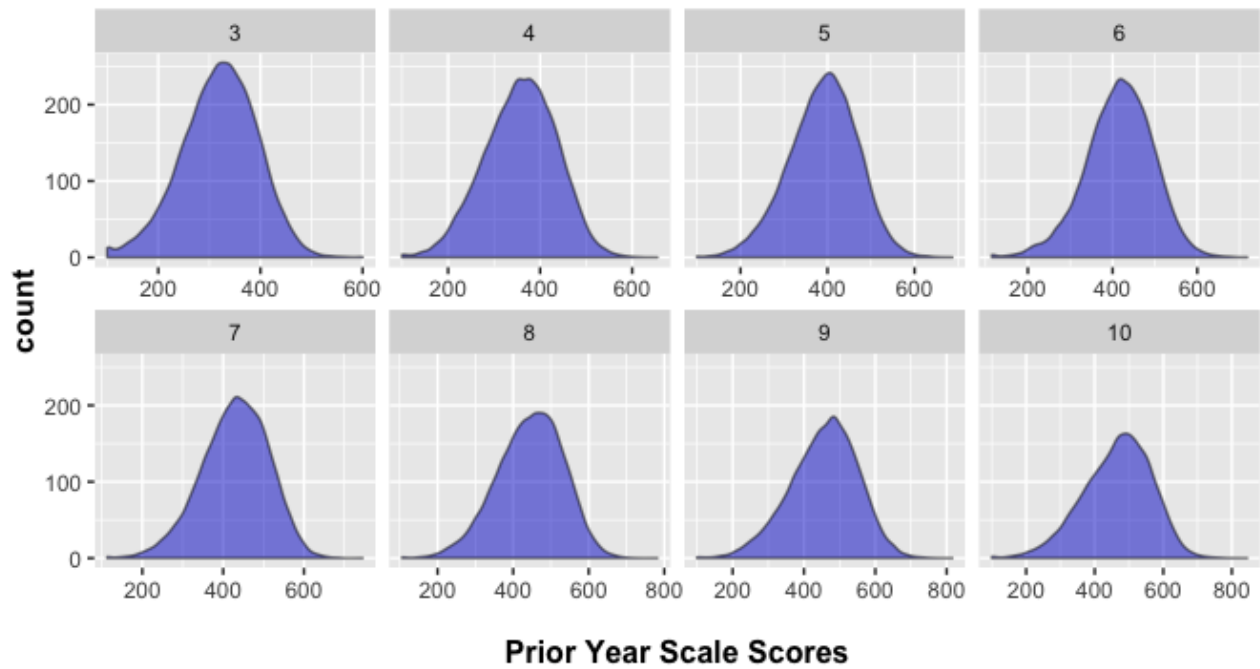
Current Year Scale Scores
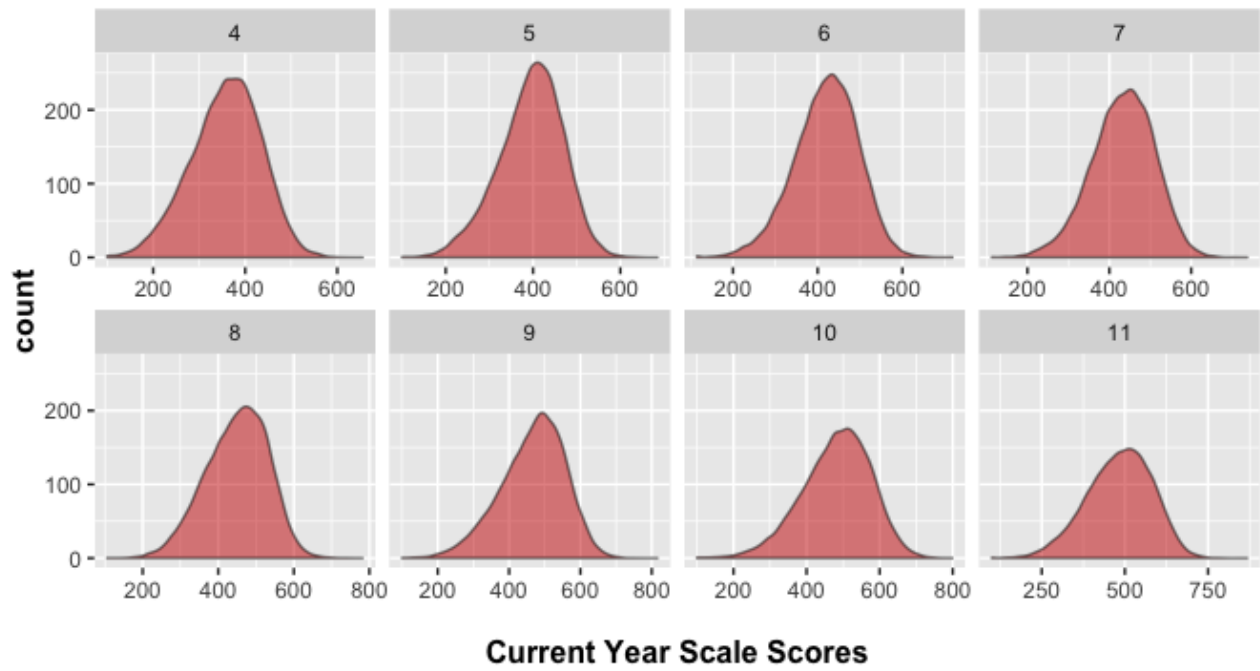
2.1.2 Mathematics

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

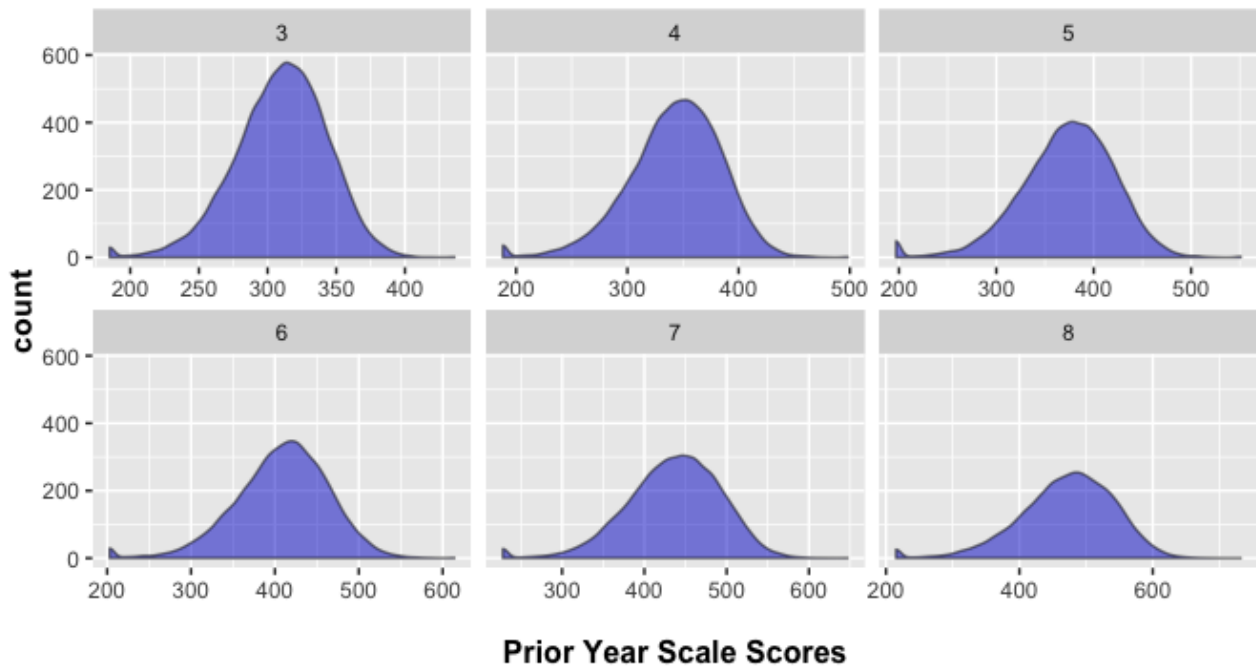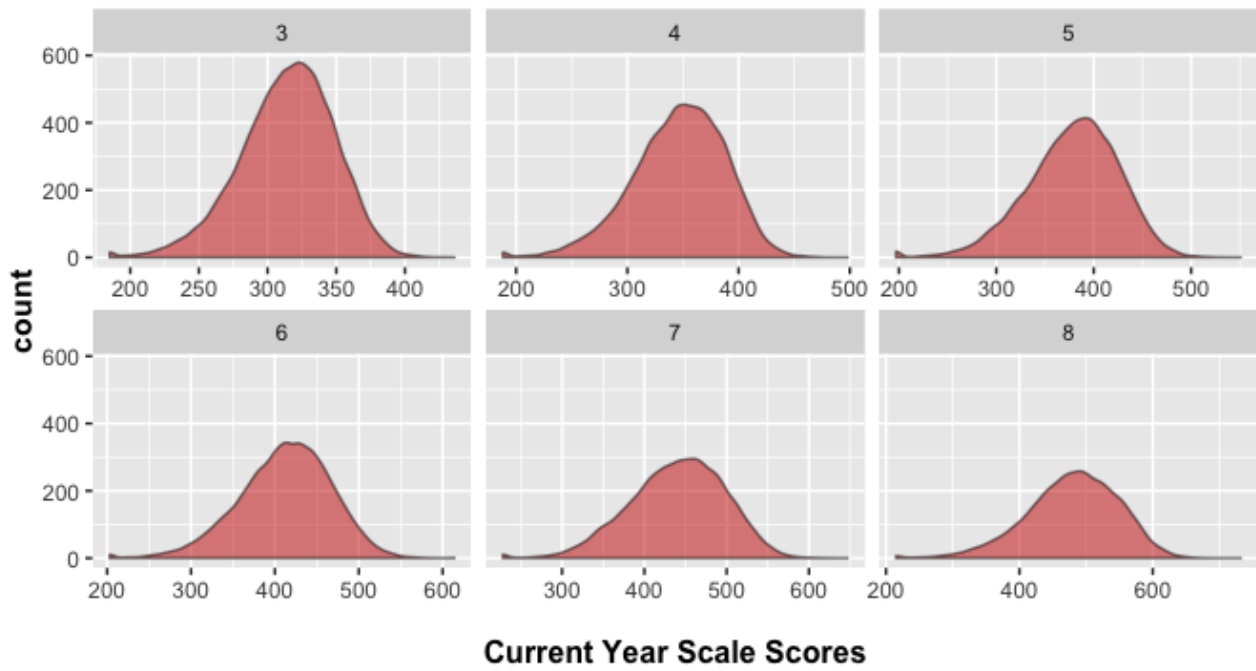**Fig. C.4:** Marginal distributions of current scale scores: EOGT Mathematics.

2.1.3 Science

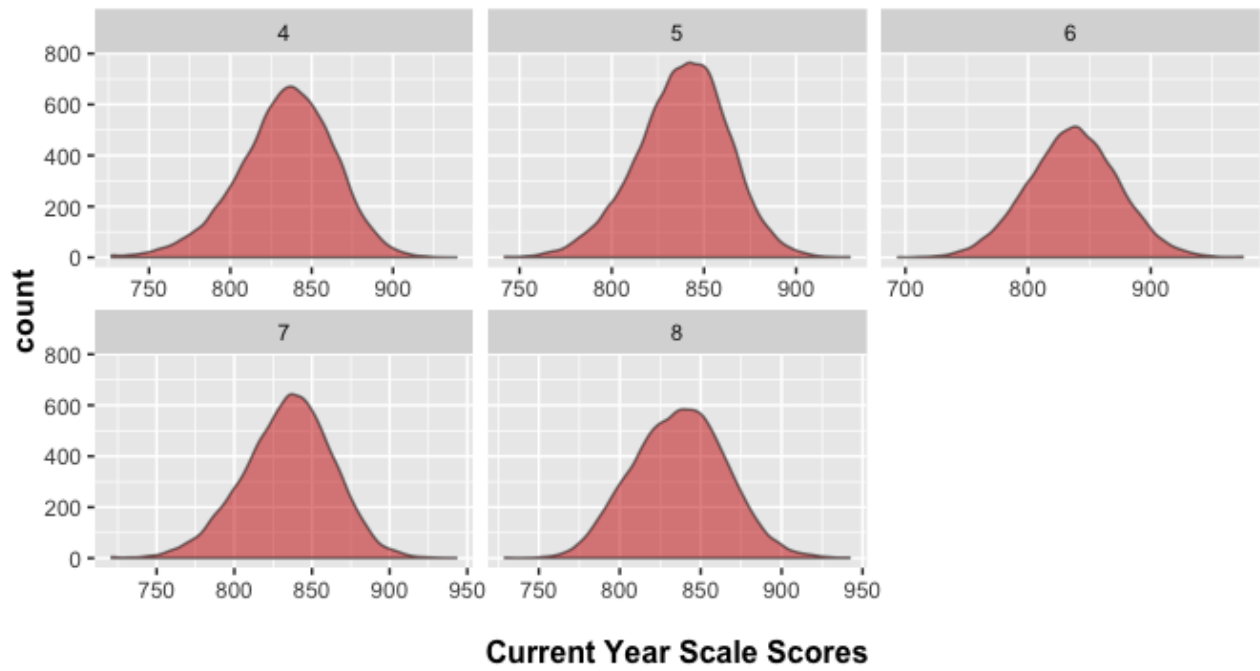**Fig. C.5:** Marginal distributions of prior scale scores: EOGT Science.

Prior Year Scale Scores

Fig. C.6: Marginal distributions of current scale scores: EOGT Science.
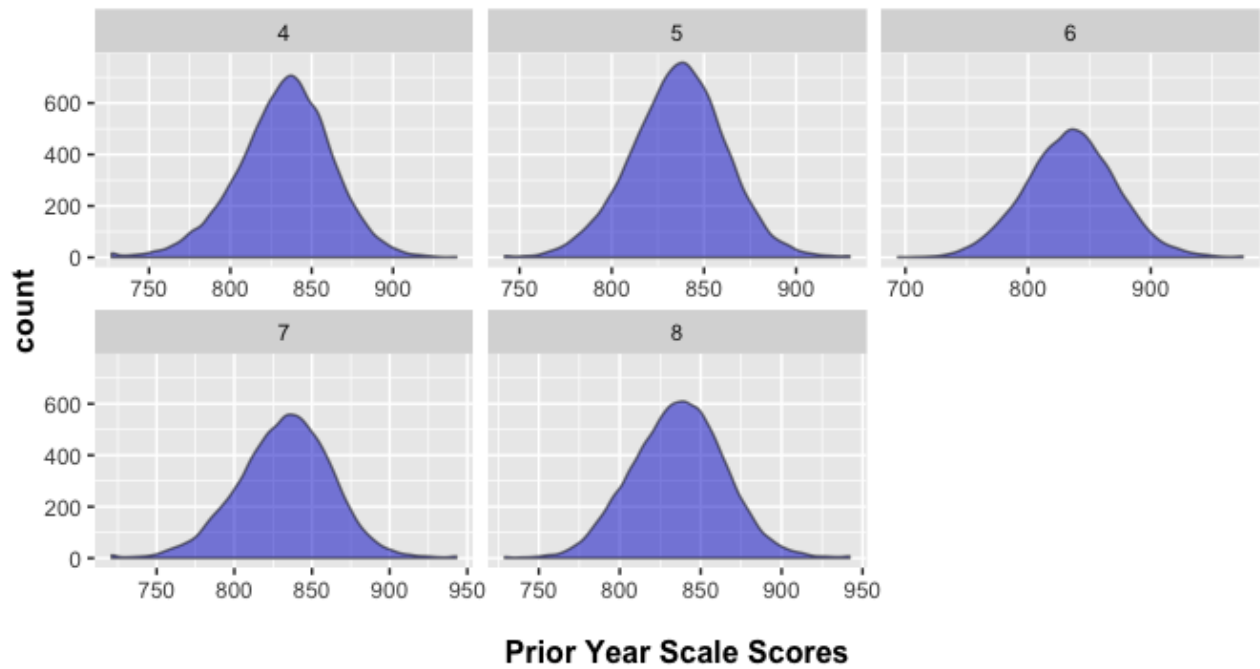
Current Year Scale Scores

2.1.4 EOCT Content Areas

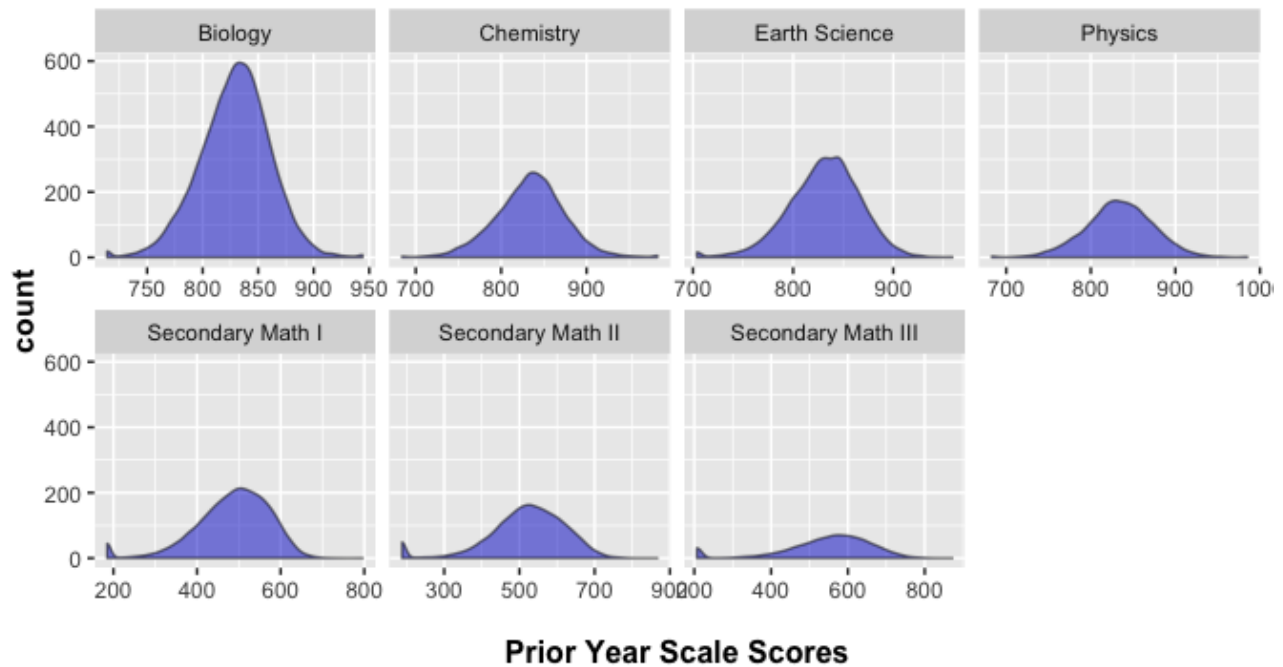**Fig. C.5:** Marginal distributions of prior scale scores: EOCT Content Areas.
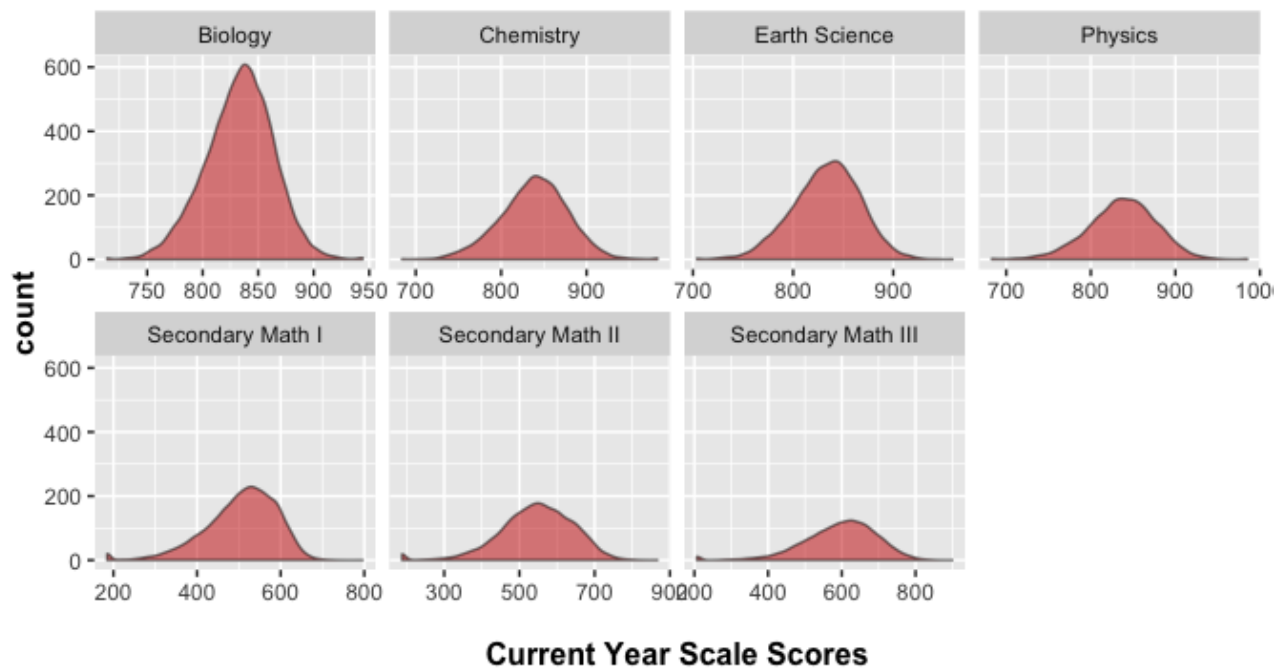

Fig. C.6: Marginal distributions of current scale scores: EOCT Content Areas.

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 Utah data, we again see very few issues in all content area and grade combinations, with the exception of the EOCT Secondary Mathematics courses. Whereas floor effects appear in single years in other places, these plots (Figure C.12) show consistent floor effects in both years and the prevalence of students scoring the lowest observable score for (at least) two years in a row.

2.2.1   Mathematics

Fig. C.7: Conditional distribution(s) of current and prior scale scores: EOGT Mathematics.
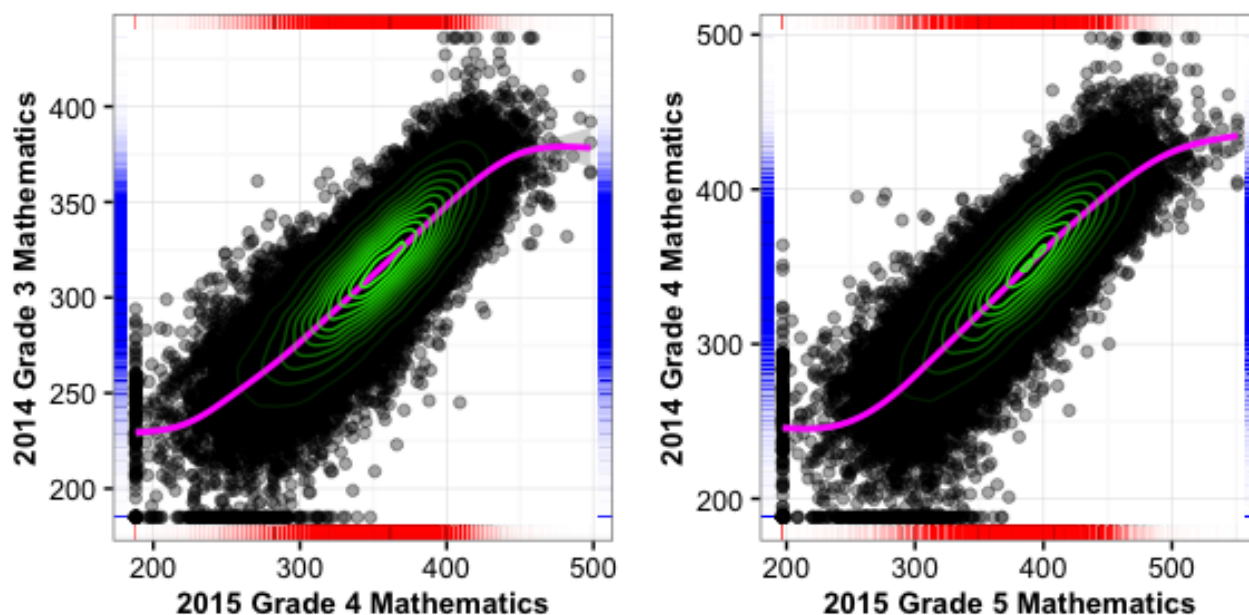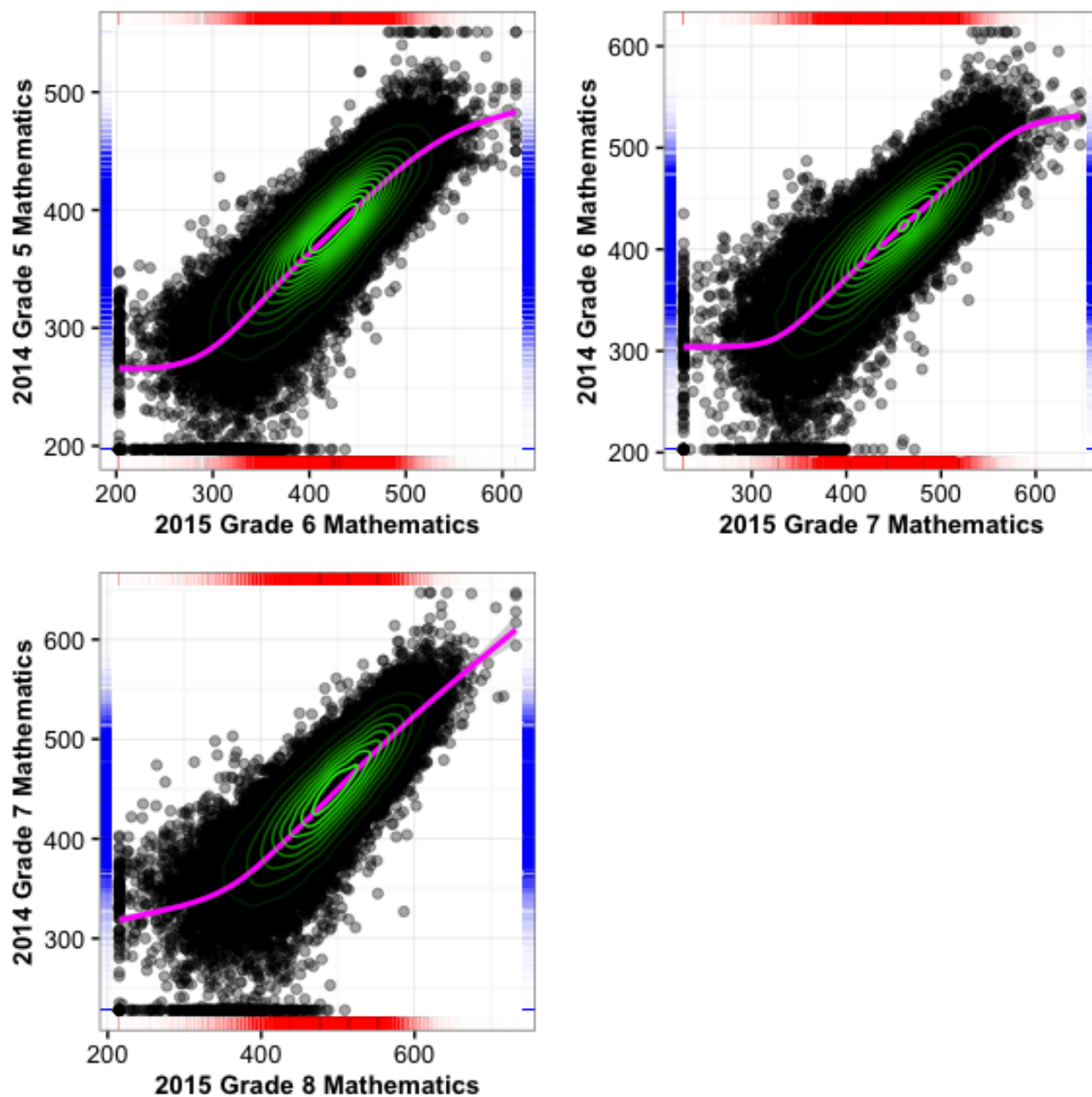
Fig. C.8: Conditional distribution(s) of current and prior scale scores: EOGT Mathematics *Continued.*
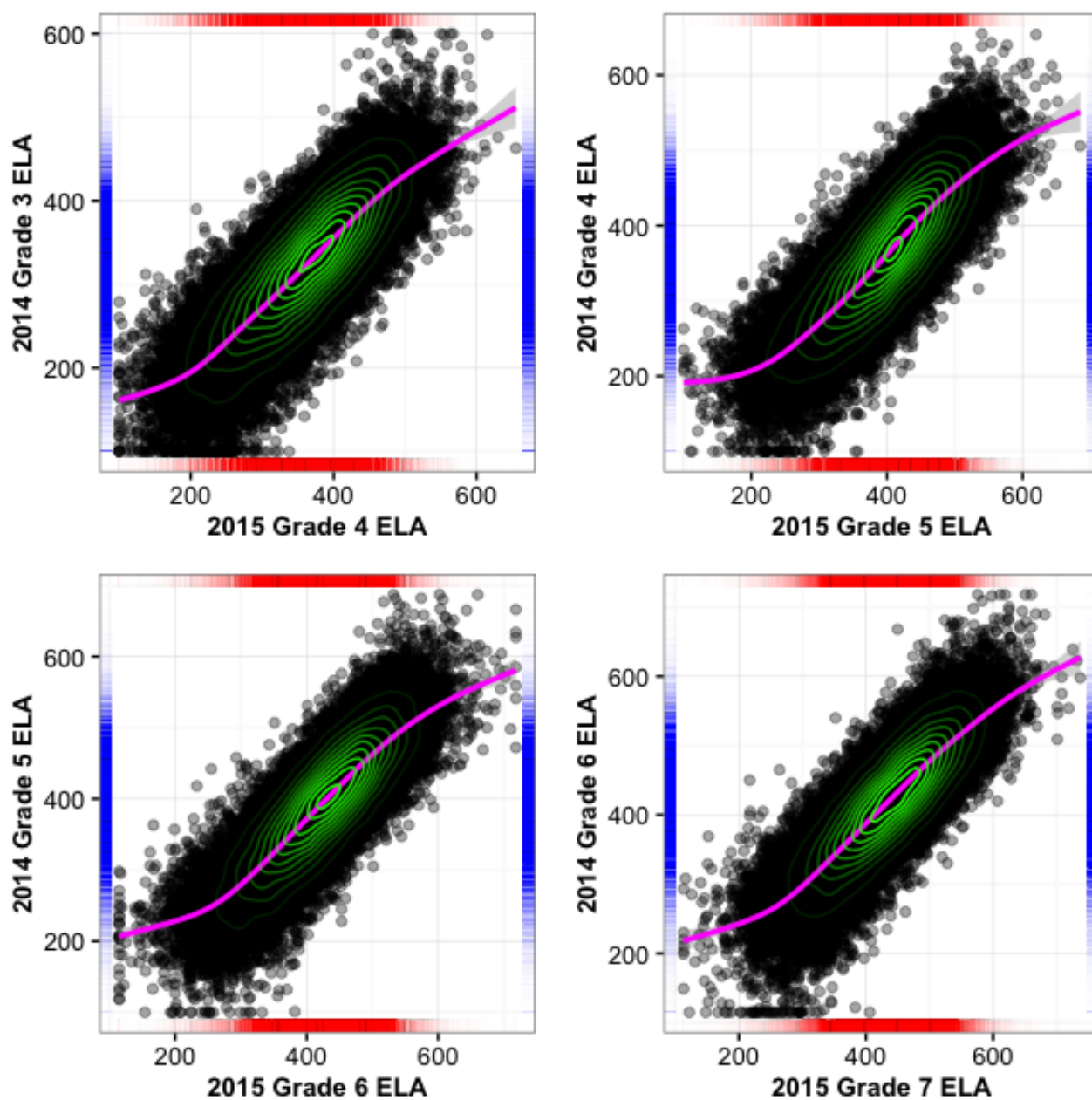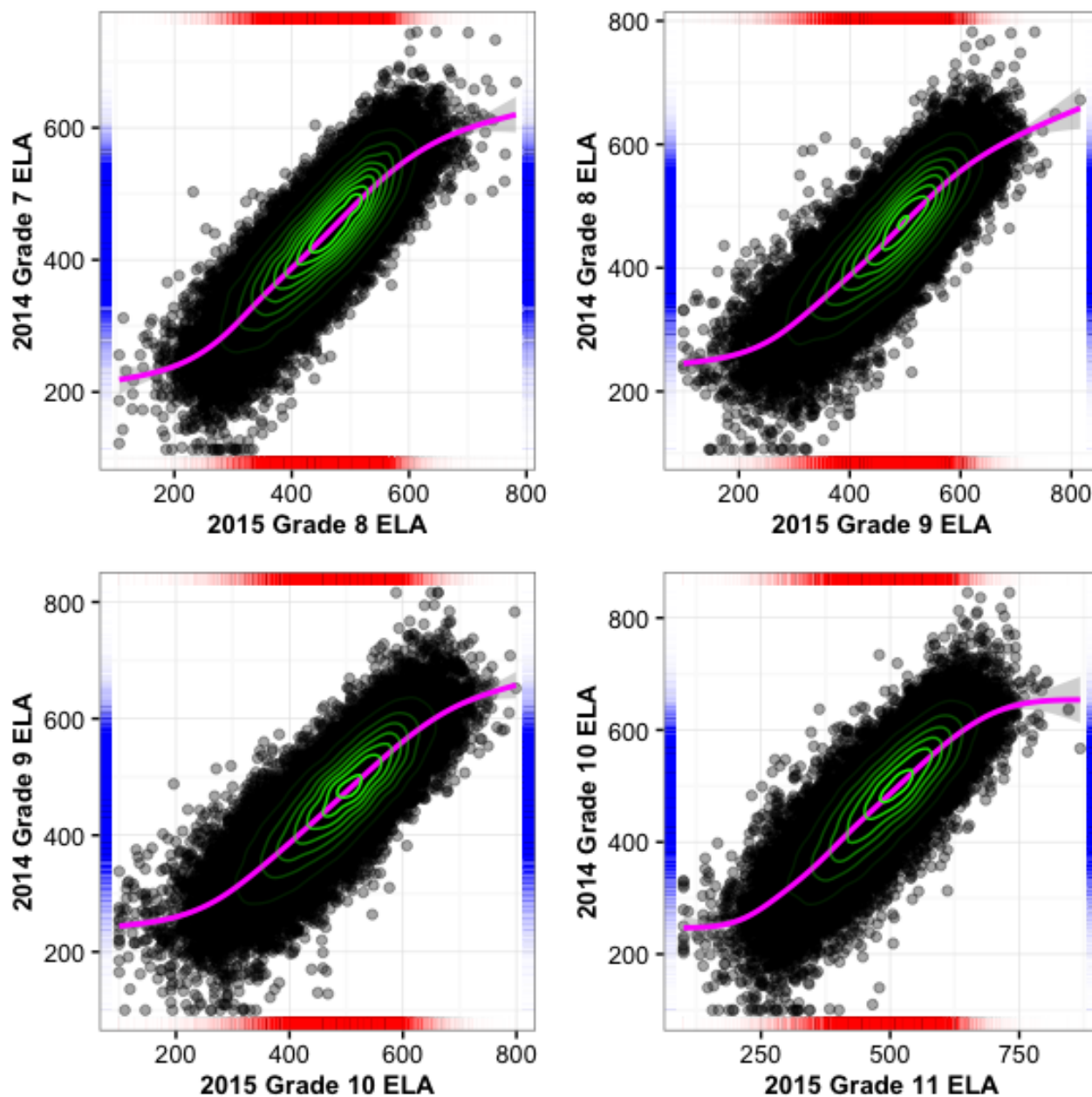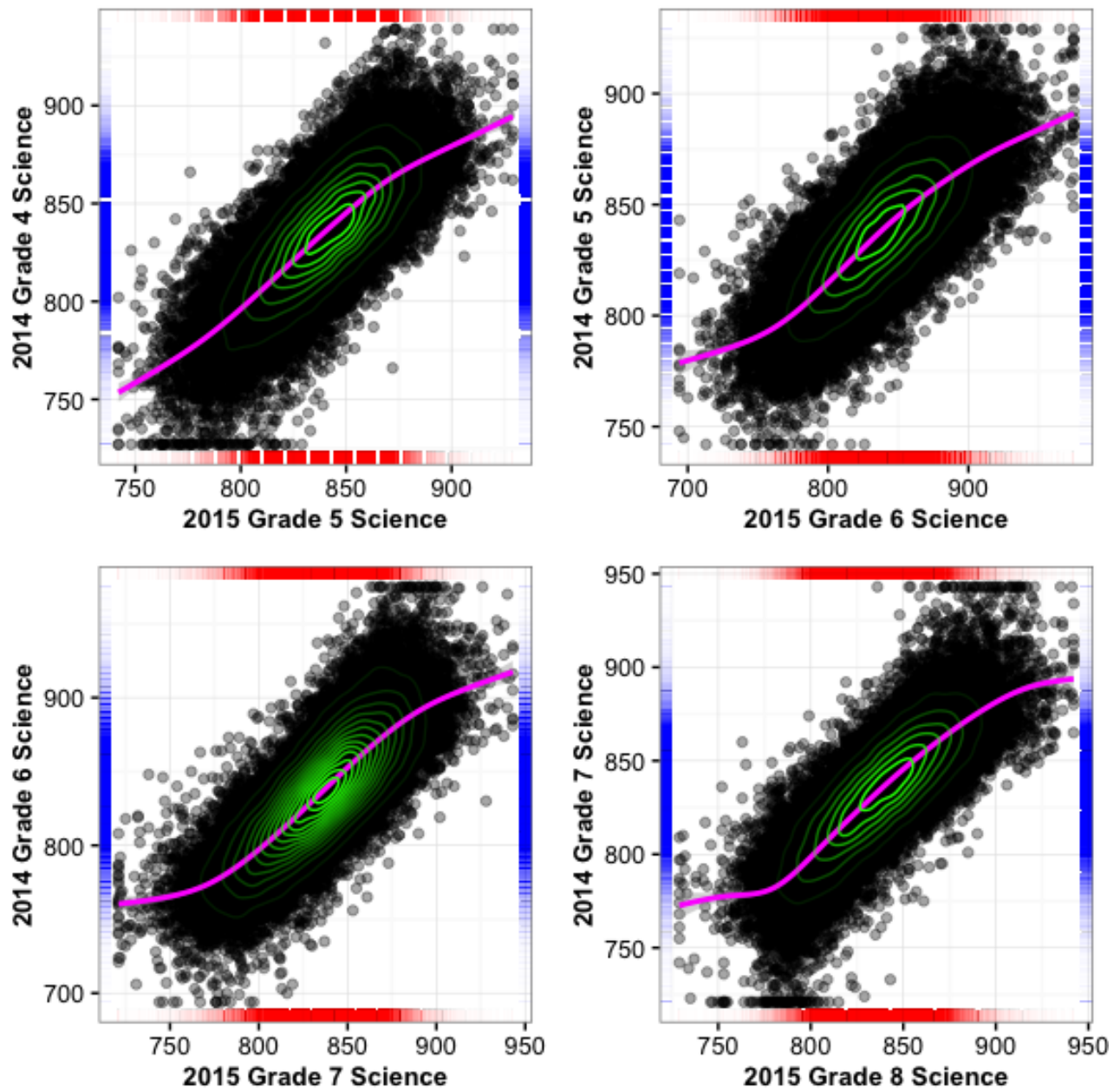
2.2.2   ELA

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

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

2.2.3 Science

Fig. C.11: Conditional distribution(s) of current and prior scale scores: EOGT Science.

2.2.4   EOCT Content Areas

The Secondary Mathematics courses for both the 2014 and 2015 Utah EOCT data[1] have floor effects, including students scoring the lowest observable score for both years (visible as the dark shaded point in the extreme bottom left corner of the plots in Figure C.12). This suggests that staying at the floor from year to year is not uncommon, which may lead to inflated growth estimates for these chronically low achieving students.

Fig. C.12: Conditional distribution(s) of current and prior scale scores: EOCT Mathematics.

---
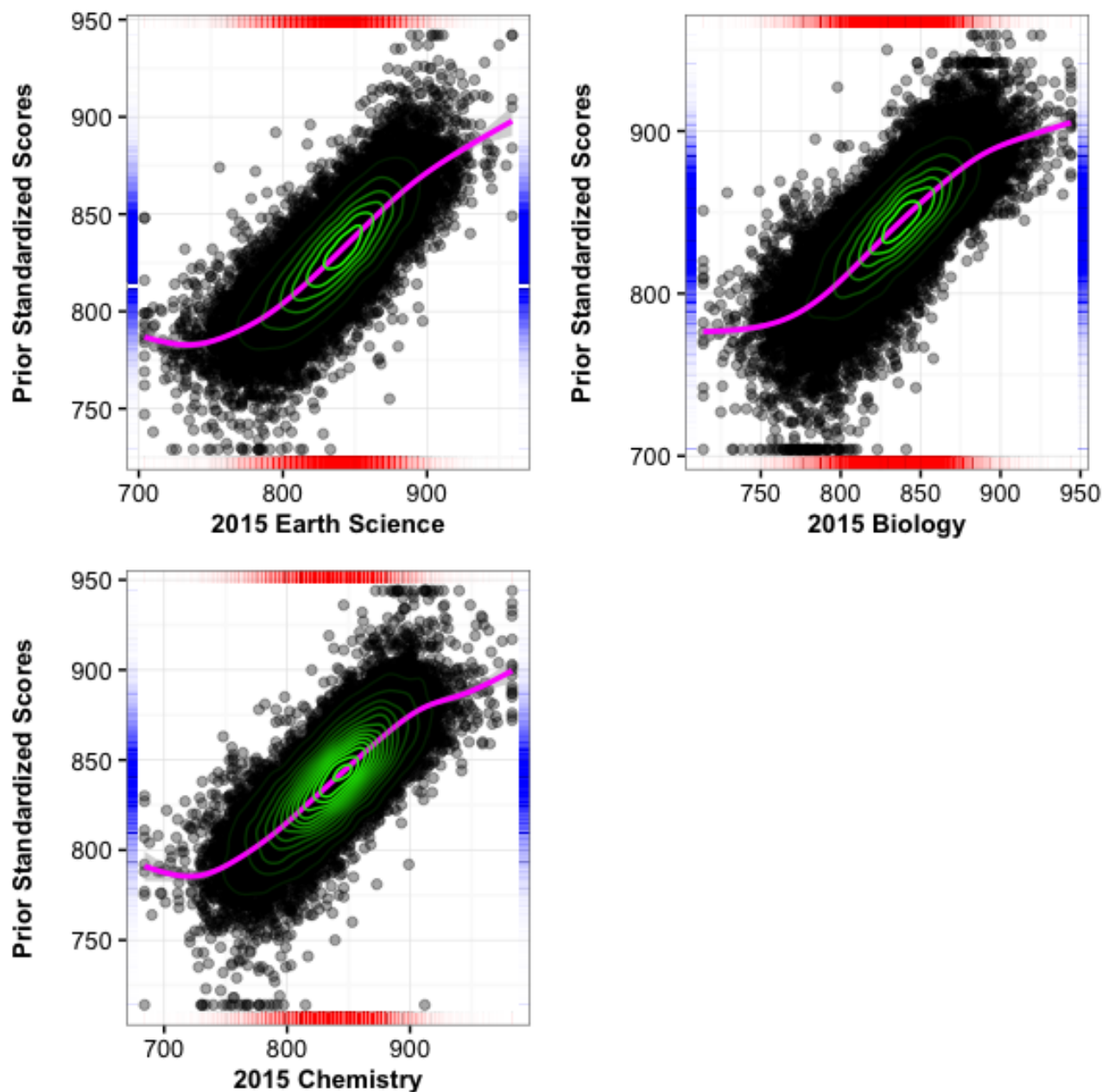
[1]Note that all of these plots use the standardized prior year scale scores. This is due to the use of multiple content area progressions in the EOCT analyses, so the standardization puts the prior scores on a more comparable scale.

Some more minor floor effects are apparent for the 2014 science related EOCT scores, and there may be some evidence of a ceiling effect in Physics. Also of note is the dark shaded point in the extreme top right corner of the Earth Science plot. Although this does not look particularly concerning here, this is evidence of a minor ceiling effect and is corroborated in the following section on the SGP distribution for the 2015 highest scoring students.

Fig. C.13: Conditional distribution(s) of current and prior scale scores: EOCT Sciences.
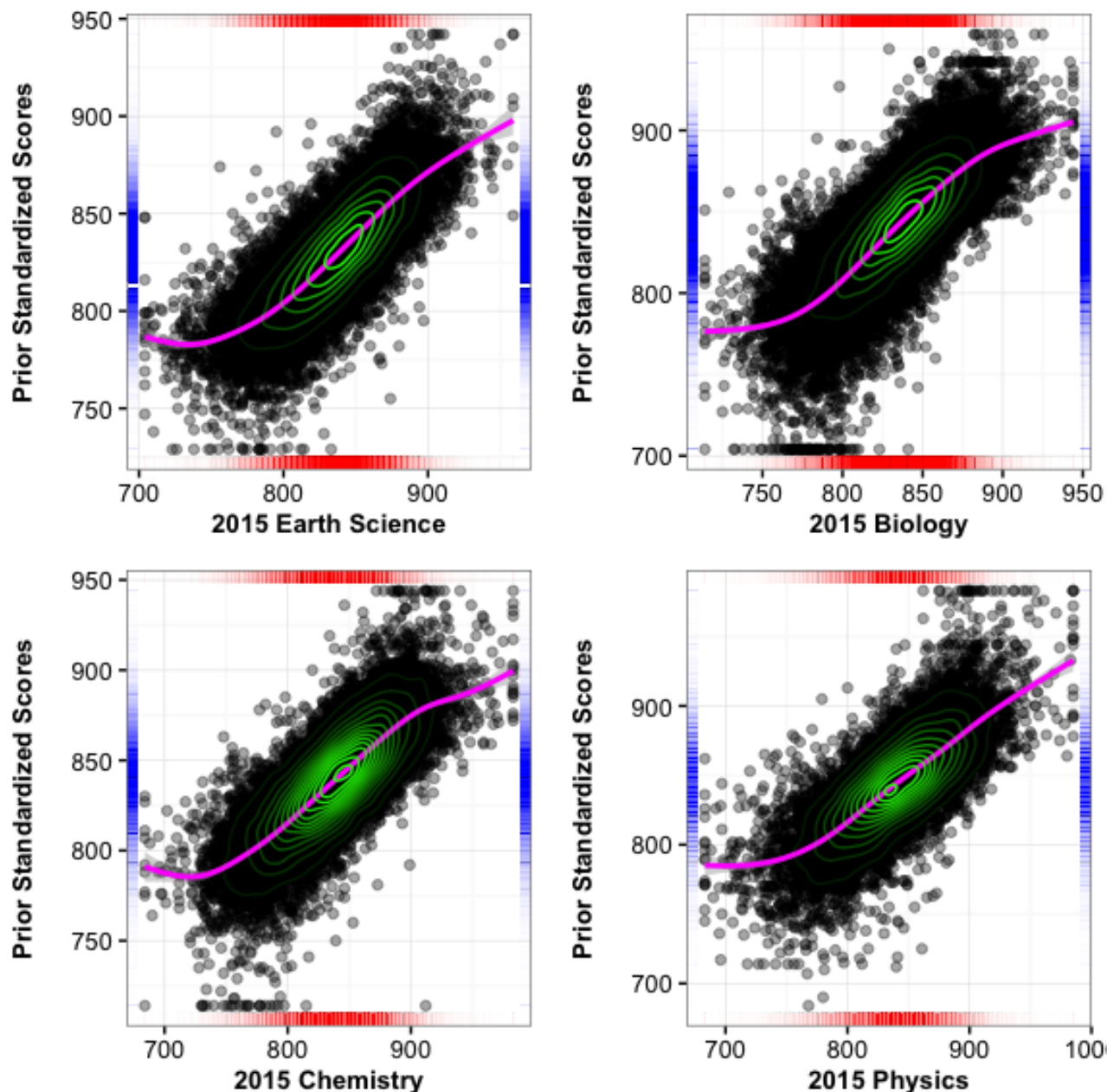
3 SGP Ranges for the Highest and Lowest Achieving Students

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 in the EOGT and EOCT subsections 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.

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 25th and 75th percentiles). The upper whisker extends from the hinge to the highest value that is within $1.5 \times$ 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 $1.5 \times$ 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 Utah EOGT SGP analyses, we see a slight ceiling effect in 8th grade math and some floor effects in grade level math as well (the 5th grade results are the most concerning). EOCT Secondary Mathematics SGPs have a pronounced floor effect.

3.1   EOGT Content Areas

Fig. C.14: SGP distributions for highest and lowest 0.5 percent of scale scores by content area and grade level.
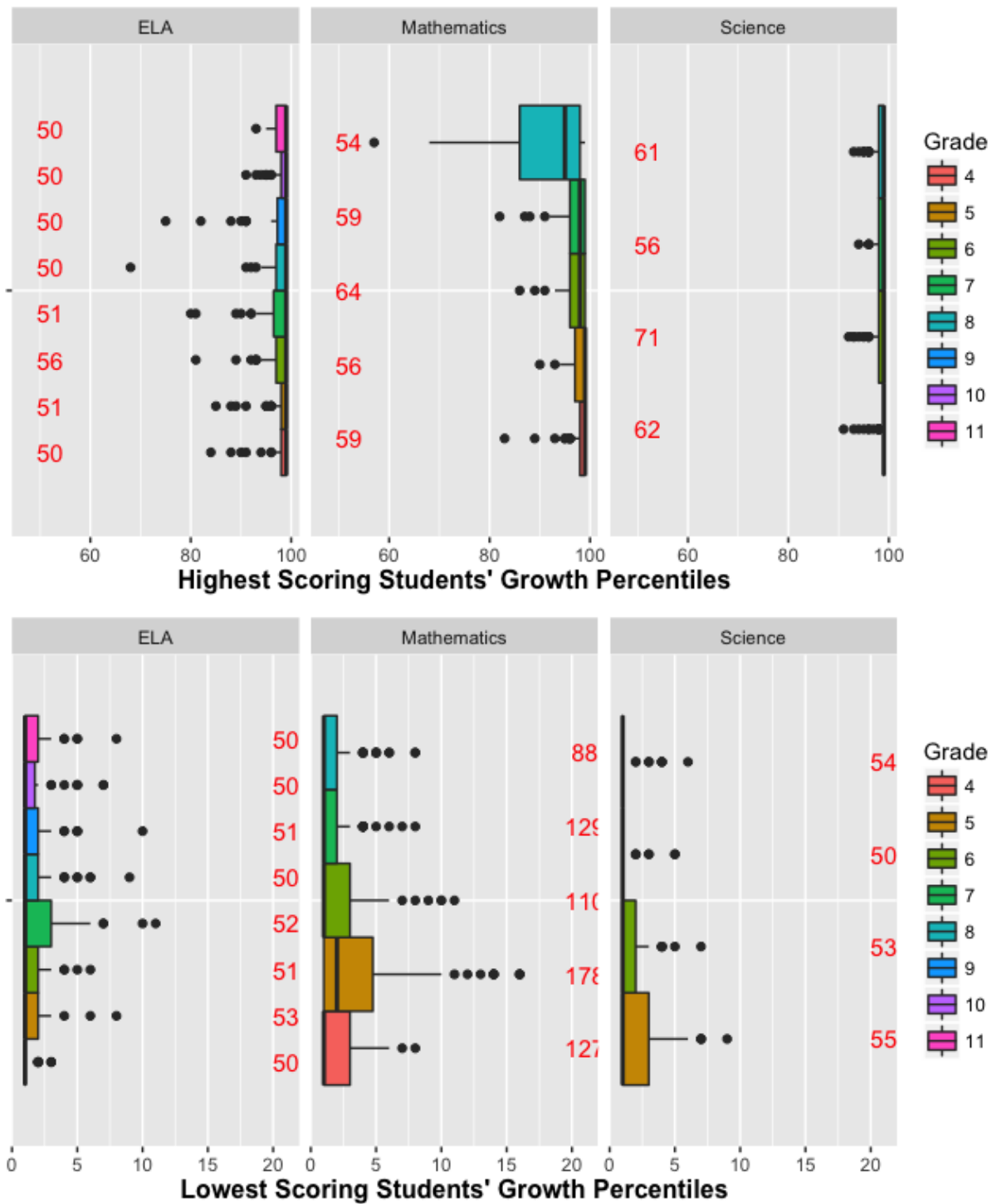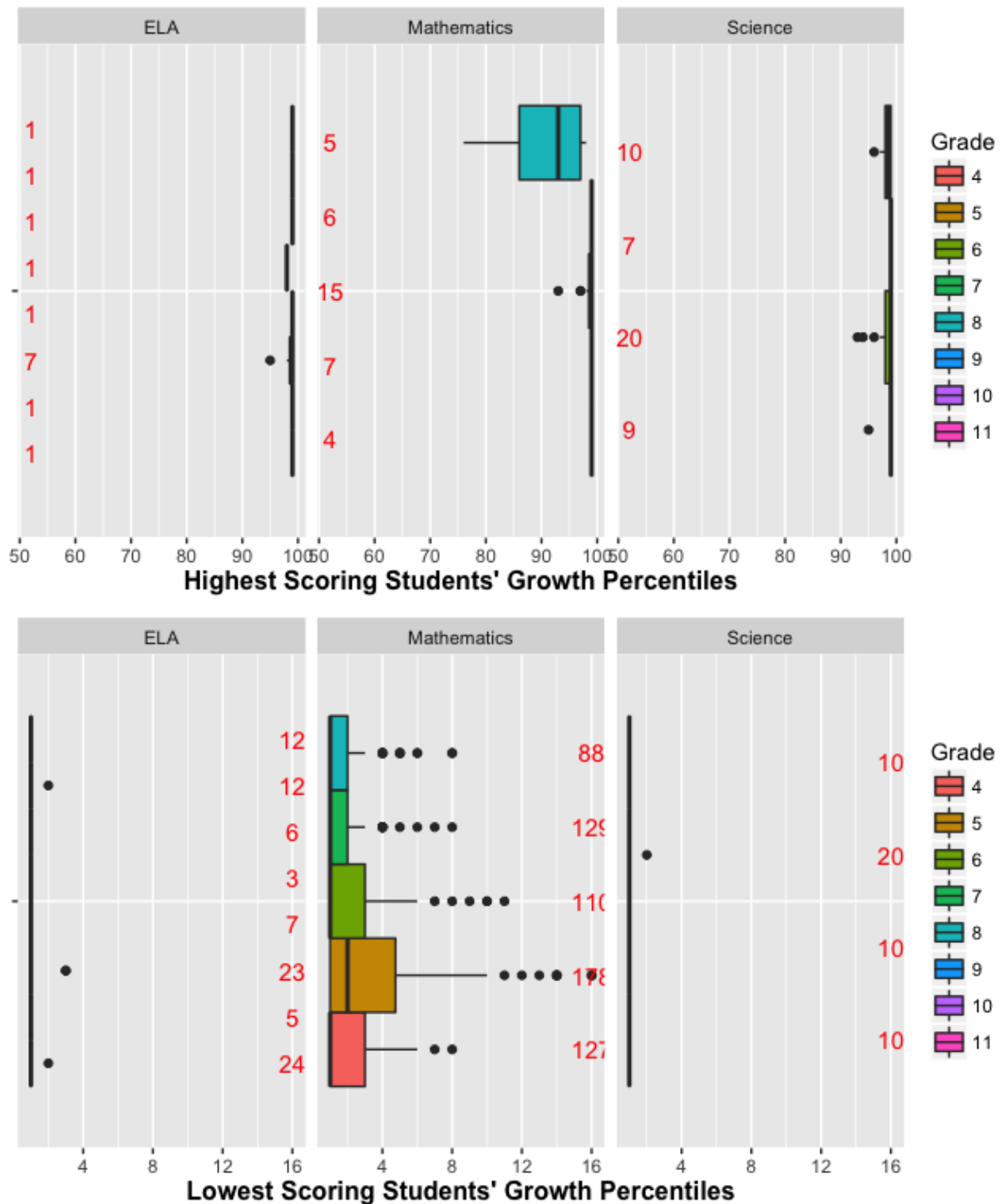
Fig. C.15: SGP distributions for the HOSS and LOSS scores by content area and grade level.

3.2 EOCT Content Areas

Fig. C.16: SGP distributions for highest and lowest 0.5 percent of scale scores by content area.
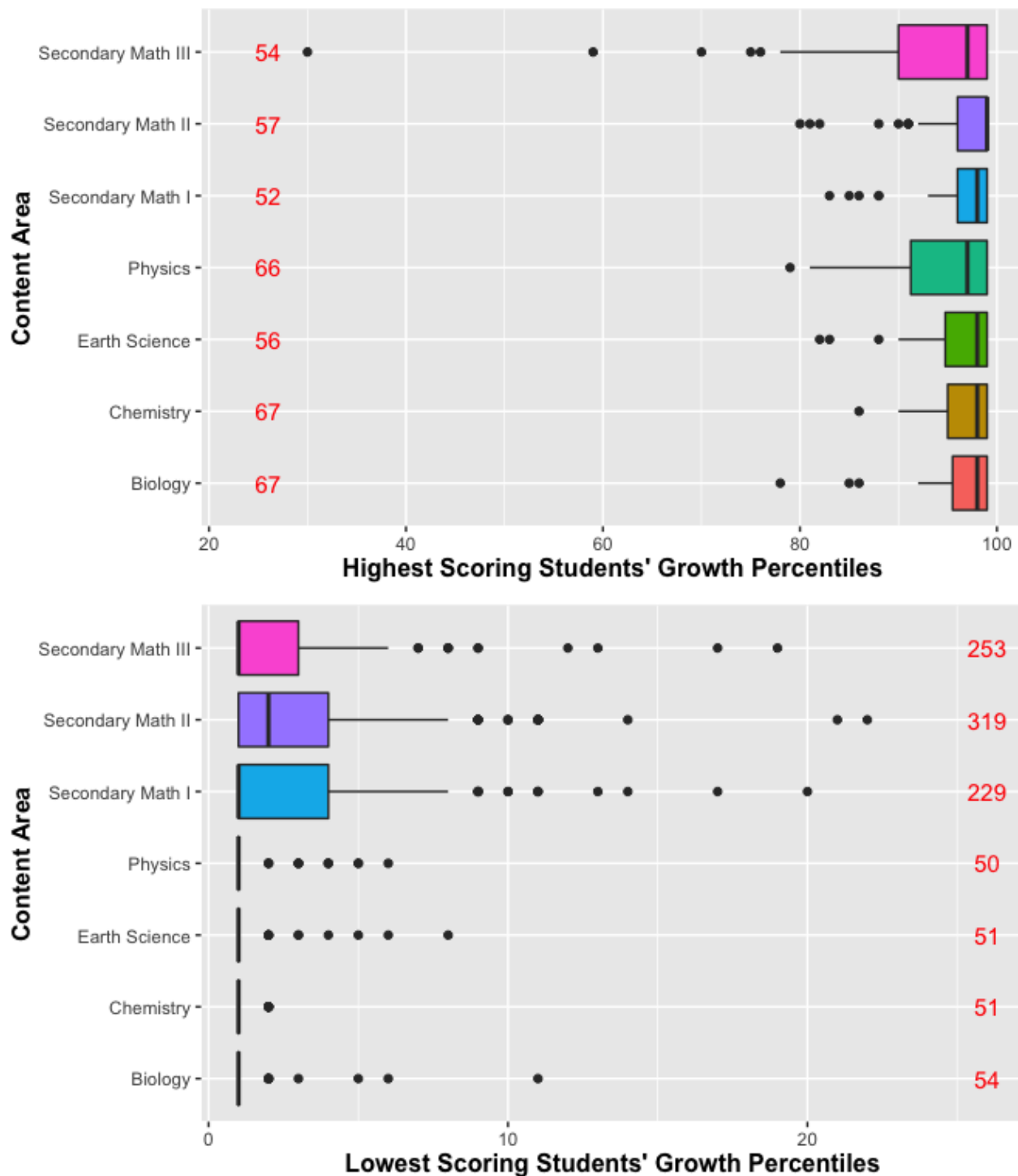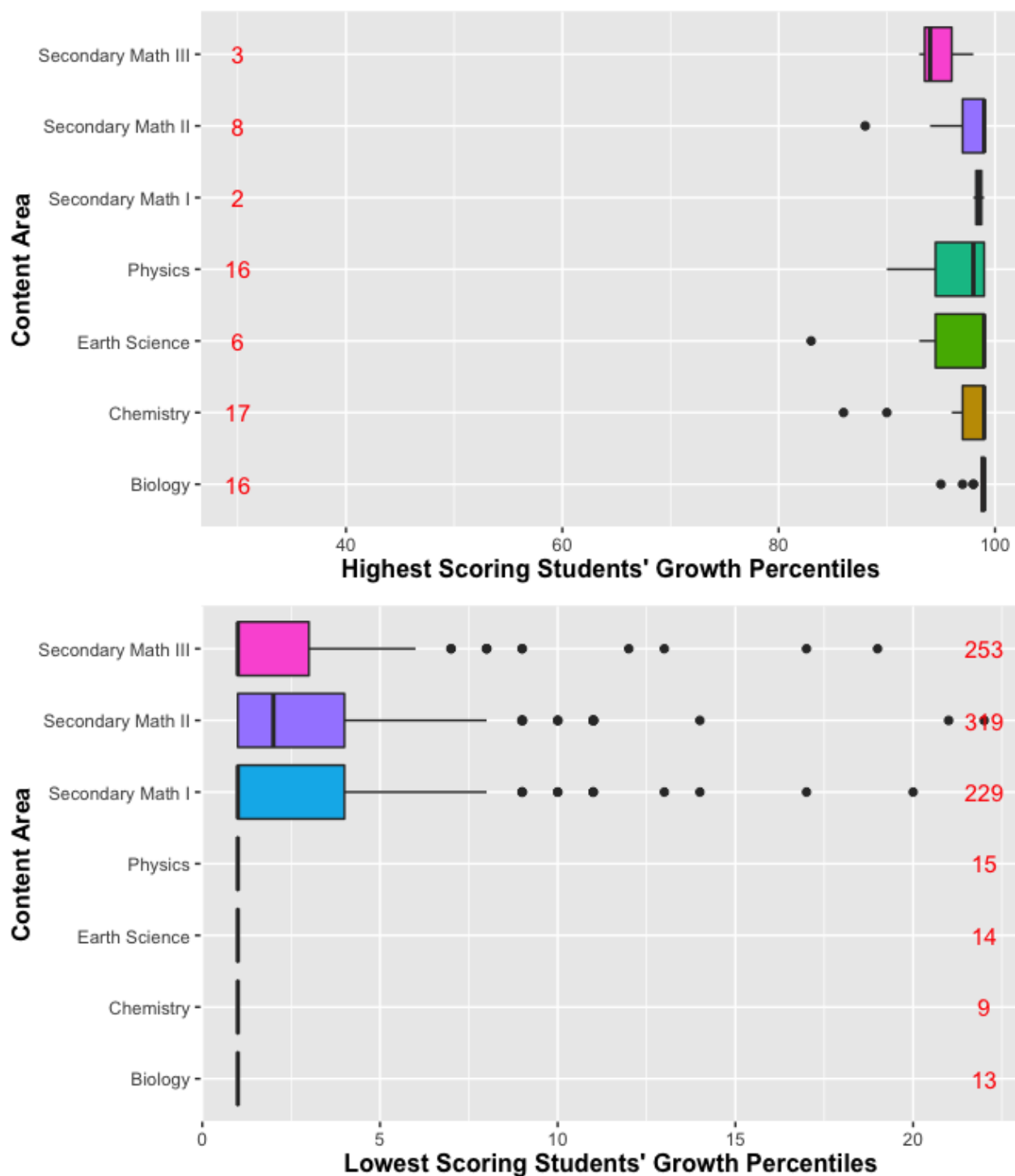
Fig. C.17: SGP distributions for the HOSS and LOSS scores by content area.

4   Discussion

Overall there is little evidence of major ceiling effects in the 2015 Utah SGP analyses. There might be some room for concern about floor effects in the Secondary Mathematics analyses, although the consequences for the floor effects may be less problematic (i.e. underestimating how low these students growth actually is).

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.