

**INVESTIGATION OF PERFORMANCE DECREASES ON THE
ENGLISH LANGUAGE ARTS CALIFORNIA STANDARDS TEST**

Jeffrey A. White, M.A.

Los Angeles Unified School District
Program Evaluation and Research Branch
Planning, Assessment, and Research Division Publication No. 227

November 15, 2004

In late August, a Board Member requested that the Program Evaluation and Research Branch investigate why test scores declined in the district. This brief attempts an initial response to this question and presents two types of explanations. The first considers potential technical reasons for declines and the second considers issues of instructional focus.

Background Information on the California Standards Tests (CSTs)

Before exploring CST scores in depth, a short description of the construction and use of these tests in California is warranted. The California Standards Tests are administered statewide as a measure of student proficiency on California grade-level standards. From 1999 to 2004, these tests have evolved from existing and augmented items extracted from the Stanford Achievement Test (9th Edition), to stand-alone tests with items developed by Harcourt Brace, to stand-alone tests that incorporate items developed by both Harcourt Brace and the Educational Testing Service (ETS).

Within grade levels, each CST contains the same number of items that measure particular areas of the California Standards. The number of items a student answers correctly are reported as a raw score. Each year, raw scores are translated to scale scores ranging from 150 to 600 for each grade level. Once scale scores are determined, they are translated into performance level scores (far below basic, below basic, basic, proficient, and advanced). Performance levels, as well as performance at or above certain performance levels, have been used as the primary metric of reporting for the Standardized Testing and Reporting system since spring 2001 for English Language Arts and spring 2002 for Mathematics. CST scores are presently the most heavily weighted component in the API of the Public Schools Accountability Act and in AYP calculations required by the No Child Left Behind legislation.

Technical Reasons for Declines in CST Scores

A natural first area of exploration regarding CST performance declines concerns the technical aspects of the test. A first step in this process is to determine whether LAUSD performance patterns differ from those of all schools in California. Figures 1 through 6 present California and LAUSD mean scale score performance on the English Language Arts CST by grade level from 2002 to 2004. California charts are on the left with gray backgrounds and LAUSD charts are on the right with white backgrounds.

Overall, patterns of performance for California and LAUSD are similar. Performance increases and decreases were equivalent in all grade levels except grade 10 where LAUSD demonstrated a slight decline while California schools demonstrated an increase in scale score performance. The finding that

performance patterns for LAUSD and California are so similar would seem to indicate that technical properties in the construction of the CSTs might, in part, explain performance decreases in LAUSD.

Figure 1. Mean ELA Scale Score: California Elementary Schools

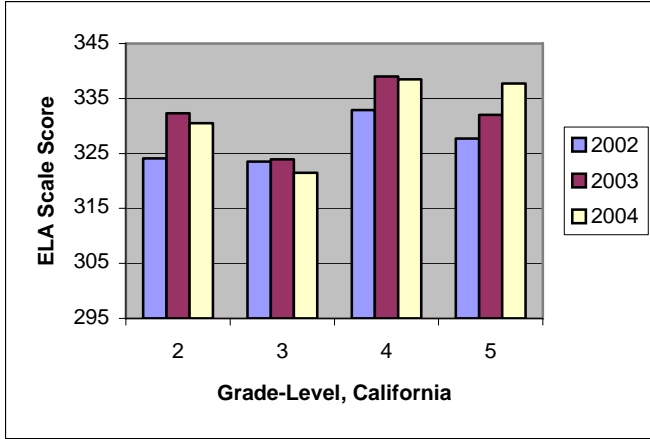


Figure 2. Mean ELA Scale Score: LAUSD Elementary Schools

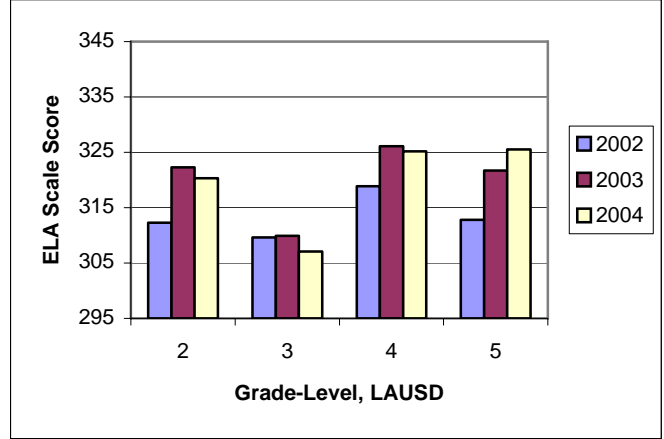


Figure 3. Mean ELA Scale Score: California Middle Schools

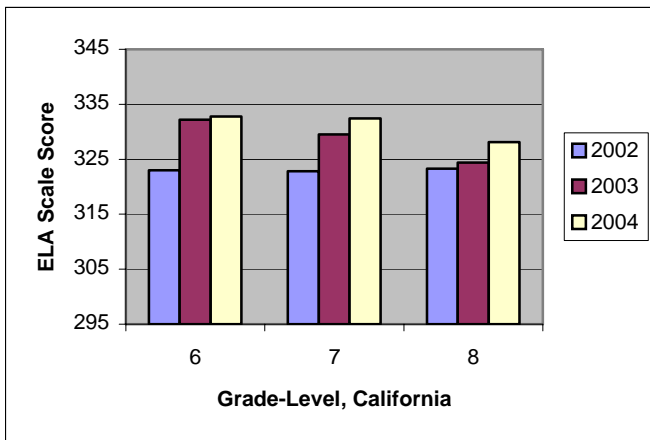


Figure 4. Mean ELA Scale Score: LAUSD Middle Schools

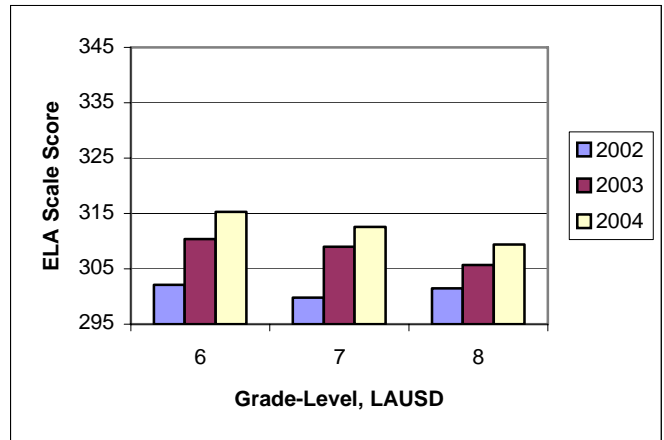


Figure 5. Mean ELA Scale Score: California Senior High Schools

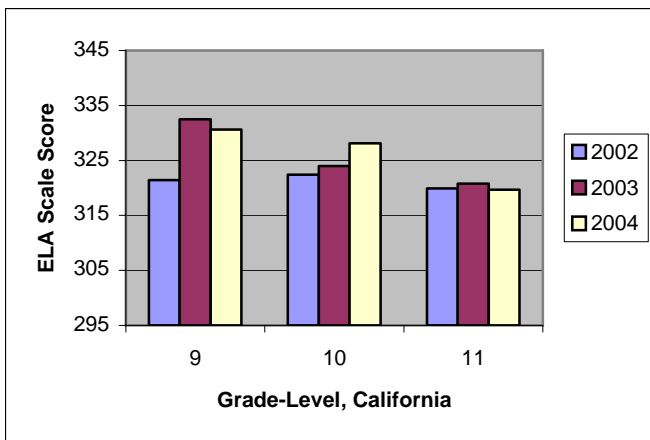
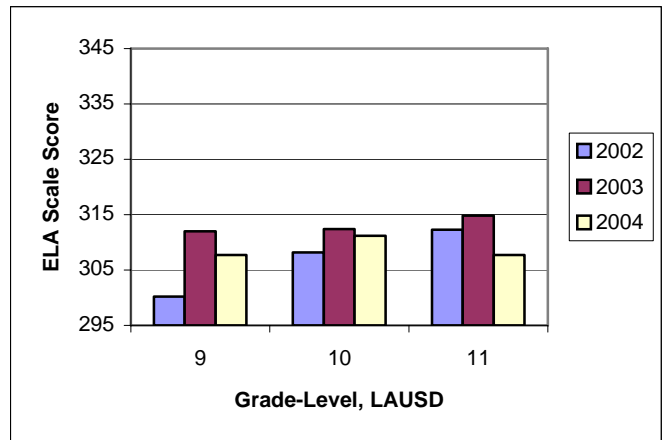


Figure 6. Mean ELA Scale Score: LAUSD Senior High Schools

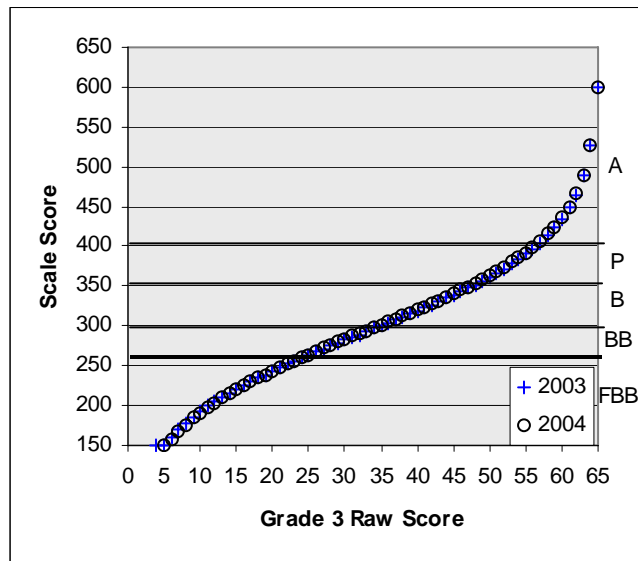


Conversations with educational researchers and evaluators with substantial technical understanding of the psychometric properties of the California Standards Test support the possibility that technical reasons may be responsible for declines in the CST. As discussed earlier, the CSTs have changed from their initial form in 1999. While the number of items and the content standards addressed within grade levels do not change, a percentage of items is replaced each year. Between the 2003 and 2004 school years, the new test vendor, ETS, reportedly replaced a greater percentage of existing Harcourt Brace developed test items with ETS developed test items. It is possible that these items may not match the difficulty levels or assess the standards in the same ways as those of Harcourt Brace. This could have resulted in lower or higher scores in some grade levels. These differences may limit the comparability of CST scores from one year to another. Unfortunately, LAUSD and other school districts in California are not provided with the item level data or the specific questions asked in order to adequately address this issue.

Another possible technical explanation is in the translation of raw scores to scale scores on the CST. Figure 7 presents the relationship between CST raw scores and scale scores in spring 2003 and spring 2004. Although Figure 7 presents the relationship for third grade students only, it is similar to that of other grade levels. This raw score to scale score relationship is represented with a plus sign in 2003 and a circle in 2004. The darker gridlines on the chart are the cut points of CST performance levels and provide a reference point for scale scores. In general, as CST raw scores increase, scale scores increase at a higher rate. This relationship is intended to take into account the difficulty of answering an increasing number of items correctly. As one can see, the distance between the highest raw score (65) and the second highest raw score (64) is much smaller than the distance in the respective scale score units (600 and 528).

If a perfect, one-to-one, relationship between 2003 and 2004 raw and scale scores existed, the plus sign would be directly in the middle of each circle. While this is not the case, the discrepancy between the two years does not appear to be large enough to expect major scaling differences in the two tests. Thus, it does not appear that scaling differences between 2003 and 2004 would explain performance declines.

Figure 7. Relationship between ELA CST Raw and Scale Scores: Grade 3



Substantive Reasons for Declines in CST Scores

While technical reasons for declines in CST performance are plausible, substantive reasons related to instruction may also explain some of the declines. The English Language Arts California Standards Test is comprised of five reporting clusters aligned with the California Content Standards. Each reporting cluster has a differential demand for critical thinking and a different relative influence on the total raw score at different grade levels. *Word Analysis, Fluency, and Systematic Vocabulary* requires that students understand the basic features of reading, select letter patterns and know how to translate them into spoken language by using phonics, syllabication, and word parts, and thus be able to achieve fluent oral and silent reading. These items are among the least demanding ELA reporting clusters and constitute the highest proportion of items at lower grades and the lowest proportion in higher-grade levels.

Written and Oral English Language Conventions requires students to write and speak with a command of standard English conventions appropriate to the grade level. This is the second least demanding cluster and proportionally constitutes the second highest percentage of items at lower grades and decreases as grade levels increase.

Reading Comprehension requires students to read and understand grade-level appropriate material; drawing upon a variety of comprehension strategies as needed (e.g., generating and responding to essential questions, making predictions, comparing information from several sources). This section is a

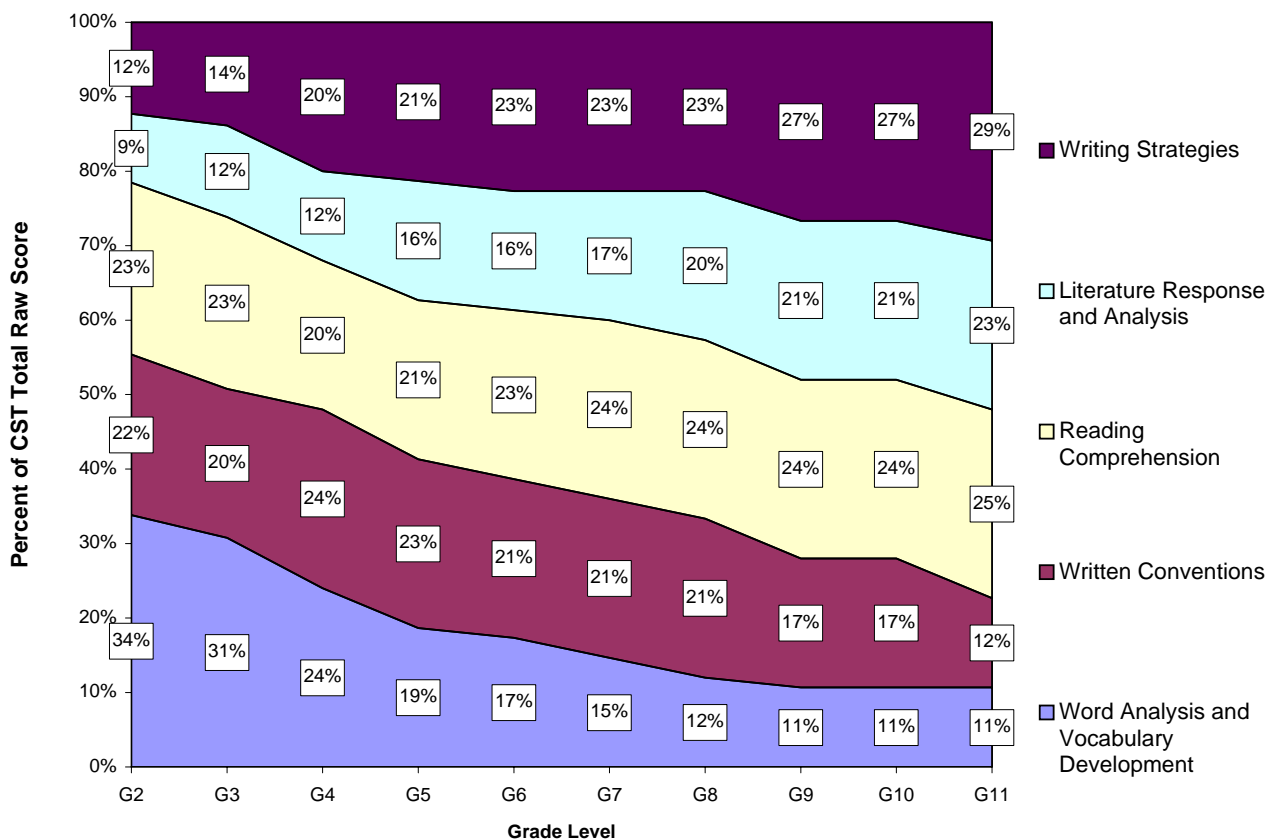
more demanding type of item on the ELA CST and makes up approximately 20 to 25 percent of the items on the ELA CST.

Literature Response and Analysis requires students to read and respond to a wide variety of significant works of children’s literature such that they can distinguish between structural features of the text and literacy terms and elements (e.g., theme, plot, setting, and characters). This is also among the more demanding CST sections and increases from 9% in grade 2 to 23% in grade 11.

Finally, *Writing Strategies* requires students to write clear sentences and paragraphs that develops a central idea, considering their audience and purpose, progressing through the stages of the writing process (e.g., pre-writing, drafting, revising, and editing successive versions).

Figure 8 presents the relative influence of each reporting cluster on the total raw score at each grade level.

Figure 1. Percent of English Language Arts Raw Score by Grade Level



Investigating raw scores performance on each reporting cluster is helpful in determining areas of student strength and weakness that may be associated with instructional changes from one year to the next.

Tables 1 and 2 present raw score data by reporting cluster and grade level for spring 2003 and 2004. Table 1 presents the two reporting clusters with a decreasing number of items as grade levels increase. In general, these items are less demanding in terms of critical thinking skills. Table 2 presents the reporting clusters with an increasing number of items as grade levels increase. In general, these items are more demanding in terms of critical thinking skills. The bold-faced type indicates a decrease in mean raw score whereas the regular-faced type indicates no change or an increase in mean raw score from 2003 to 2004.

Table 1 shows that student performance on the reporting clusters that demand less critical thinking and decrease in relative importance to the total test score was more likely to improve, while Table 2 indicates that student performance on the reporting clusters that demand more critical thinking and increase in relative importance to the total test score was more likely to decline.

These findings are consistent with interview and observation data from the 2003-04 Reading First Evaluation Report (forthcoming) related to the use of the Open Court unit assessment in Reading. It is possible that student drilling may have been used to influence the results of unit assessment data. The 2002-03 school year was the first school year where Open Court unit assessments were administered. These tests focus on a subset of the identified grade-level standards. The 2003-04 school year was the first time that teachers were required to use the Open Court unit assessments to guide instruction. Additionally, schools were required to submit the results of this assessment to their local district superintendents. While Open Court unit assessments are not psychometrically linked and are not intended for high stakes purposes, many teachers believed they would be held accountable for student success on the assessments. Additionally, teachers were instructed to plan their lessons using the unit assessment data. A common reported teacher response to this policy was “if nothing else, make sure you cover the material to be tested in the unit assessment.” Many teachers may have narrowed or truncated the Open Court curriculum to emphasize content that would be assessed, rather than focusing on broader skills and strategies used to build proficiency.

Consider an example regarding the skill of inference, tested in one of the unit assessments. If the assessment items asked students to read a paragraph and draw inferences, some teachers might have spent the next 6 weeks focusing and drilling on inference, excluding instruction in five other skills areas that should spiral through the curriculum but were not assessed. This practice is inconsistent with the framework of Open Court, limits exposure to other content areas covered on the CST, and would be consistent with lower scores on the more critically demanding areas of the standards.

While it is impossible to tell whether these findings are pervasive enough to explain the lower grade level decline in CST performance in LAUSD, it certainly deserves mention. Not only would this approach to instruction have a detrimental effect upon the current scores, future scores would also be likely to diminish due to the increased emphasis on higher critical thinking on the ELA CSTs as grade levels increase.

Conclusion

This brief was intended to shed some light on why test scores may have decreased. Given our limited access to item-level data and specific CST questions, it is difficult to definitively state whether performance decreases are associated with changes in the CST. The similar performance patterns exhibited by LAUSD and California suggest this as a possibility. On the other hand, interview and observation data from our ongoing evaluation of Reading First indicate a potential misuse of Open Court unit assessments that corresponds with reporting cluster declines, suggesting that instructional changes may also be factor. The Program Evaluation and Research Branch will continue to investigate these issues and share any additional findings or explanations as they become available.

Table 1. CST ELA Raw Score Performance by Grade Level and Reporting Cluster

Reporting Cluster	Grade	N of items	% of items	2003 N	2004 N	2003 Mean	2004 Mean	Change	% Change	2003 SD	2004 SD
Word Analysis and Vocabulary Development	2	22	34%	63,279	61,344	13.1	12.1	-1.0	-7.6%	4.9	4.4
	3	20	31%	63,642	62,222	11.9	12.2	0.3	2.6%	4.3	4.5
	4	18	22%	59,857	62,998	9.6	10.4	0.7	7.7%	3.8	4.0
	5	14	19%	54,562	60,671	7.2	7.6	0.4	6.1%	3.0	3.4
	6	13	17%	55,853	53,211	6.1	7.1	0.9	15.4%	2.8	2.9
	7	11	13%	53,696	56,432	6.4	5.9	-0.5	-8.4%	2.6	2.5
	8	9	12%	49,933	54,697	4.8	4.9	0.1	1.5%	2.0	1.9
	9	8	11%	56,900	58,588	3.5	4.1	0.6	16.2%	1.7	2.0
	10	8	11%	39,571	44,222	4.1	4.6	0.5	13.3%	2.1	1.9
	11	8	11%	27,900	31,939	4.4	4.6	0.1	3.3%	2.0	2.0
	Written Conventions	2	14	22%	63,279	61,344	8.8	8.4	-0.5	-5.2%	3.4
3		13	20%	63,642	62,222	7.1	7.3	0.2	3.3%	3.0	2.8
4		18	22%	59,847	62,998	9.1	9.6	0.5	5.9%	3.6	3.8
5		17	23%	54,553	60,671	8.2	10.1	1.9	23.0%	3.0	3.5
6		16	21%	55,808	53,211	8.3	9.6	1.2	14.6%	3.0	3.6
7		16	19%	53,670	56,432	7.2	8.4	1.1	15.8%	3.1	3.4
8		16	21%	49,911	54,697	7.2	7.6	0.4	6.3%	3.0	3.1
9		13	17%	56,805	58,588	6.7	7.0	0.3	4.3%	2.7	2.7
10		13	17%	39,531	44,222	6.1	7.1	1.1	17.6%	2.4	2.9
11		9	12%	27,998	31,939	4.5	5.2	0.6	14.3%	2.0	2.2

Table 2. CST ELA Raw Score Performance by Grade Level and Reporting Cluster

Reporting Cluster	Grade	N of items	% of items	2003 N	2004 N	2003 Mean	2004 Mean	Change	% Change	2003 SD	2004 SD
Reading Comprehension	2	15	23%	63,279	61,344	7.3	7.4	0.1	1.3%	3.3	3.6
	3	15	23%	63,641	62,222	7.9	7.3	-0.7	-8.3%	3.5	3.3
	4	15	18%	59,852	62,998	7.2	7.1	-0.2	-2.2%	3.3	3.6
	5	16	21%	54,556	60,671	7.3	7.8	0.5	7.0%	3.1	3.1
	6	17	23%	55,839	53,211	7.7	7.6	-0.2	-2.0%	3.3	3.2
	7	18	22%	53,683	56,432	9.3	9.4	0.1	1.0%	3.9	3.9
	8	18	24%	49,905	54,697	8.9	8.7	-0.2	-2.4%	3.7	3.5
	9	18	24%	56,883	58,588	8.9	8.5	-0.4	-4.5%	3.8	3.6
	10	18	24%	39,565	44,222	8.7	9.1	0.5	5.4%	3.7	4.1
	11	19	25%	28,024	31,939	10.3	10.6	0.3	3.2%	4.0	3.9
Writing Strategies	2	6	9%	63,271	61,344	3.5	4.0	0.5	15.8%	1.8	2.0
	3	8	12%	63,607	62,222	4.4	4.0	-0.4	-8.3%	2.1	2.0
	4	9	11%	59,856	62,998	7.0	7.3	0.3	4.5%	3.0	3.1
	5	12	16%	54,562	60,671	7.6	7.8	0.3	3.5%	3.2	3.5
	6	12	16%	55,852	53,211	8.2	7.1	-1.1	-13.8%	3.4	3.5
	7	13	16%	53,697	56,432	7.2	7.8	0.6	9.0%	3.2	3.5
	8	15	20%	49,931	54,697	8.0	8.9	0.9	11.3%	3.2	3.7
	9	16	21%	56,912	58,588	9.5	8.9	-0.6	-6.5%	4.0	4.1
	10	16	21%	39,573	44,222	10.9	10.9	0.0	-0.3%	4.6	4.4
	11	17	23%	28,039	31,939	13.0	11.4	-1.6	-12.4%	4.9	4.6
Literature Response and Analysis	2	8	12%	63,264	61,344	3.1	3.1	0.0	-1.4%	1.7	1.7
	3	9	14%	63,642	62,222	5.3	4.8	-0.6	-10.4%	2.0	2.1
	4	15	18%	59,853	62,998	4.9	4.1	-0.8	-16.5%	2.1	2.0
	5	16	21%	54,559	60,671	6.5	7.2	0.6	9.6%	2.7	2.9
	6	17	23%	55,842	53,211	5.7	5.9	0.2	3.2%	2.9	2.6
	7	17	20%	53,690	56,432	6.7	6.0	-0.8	-11.4%	2.6	2.7
	8	17	23%	49,920	54,697	7.1	7.1	0.0	-0.2%	2.9	3.1
	9	20	27%	56,905	58,588	7.6	6.8	-0.8	-10.1%	3.4	3.2
	10	20	27%	39,552	44,222	7.1	6.7	-0.3	-4.5%	3.5	3.3
	11	22	29%	28,039	31,939	7.2	7.8	0.6	7.9%	3.2	3.4