The California Standards Test

Scientific Learning Corporation
Innovation and Research Department
Scientific Learning: Research Reports 14(3): 1-15

Executive Summary

The California Standards Tests (CSTs) are used to annually assess the reading, English, and language arts abilities of California students in grades 2-11. However, the design of the CSTs makes it challenging to use them to evaluate the impact of interventions. According to the California Department of Education, **CST scaled scores cannot be compared across grades**. This means that these scores are unsuitable for analyzing growth in student reading performance. Fortunately, CST **proficiency level scores** (see Figure 1) *can* be analyzed between grades in a valid way.

Figure 1. Conceptual diagram of CST proficiency levels. Note: Not necessarily to scale.

While proficiency level scores are our only valid window into the CST performance gains made by students and groups, these scores are limited by ordinality, low resolution, and ceiling and floor effects.

Optimally, studies investigating how interventions impact students' learning trajectories will utilize **alternative assessments** that are better suited to serve as outcomes measures. When CST proficiency levels are the only available information for analyzing student gains and the sample size is large enough (50 students or more), the strongest statistical analysis can be done using a **Monte Carlo implementation of a Non-Parametric Randomization Test** (McNPR test).

This paper outlines the sizable challenges inherent in analysis of CST results, provides an extensive list of suitable alternative assessments, and describes the McNPR test in detail.

Version 2010.04.08

Analysis of the California Standards Tests

The California Standards Tests (CSTs) are used within the Standardized Testing and Reporting (STAR) Program to annually assess the reading, English, and language arts abilities of California students in grades 2-11. The CSTs evaluate student performance relative to the California content standards for each grade and subject area, and they are a central component of the state's accountability system for schools and districts. The CSTs are well suited for comparing the performance of different schools or districts within a given school year, as long as these comparisons are restricted to a particular grade and subject area.

The design of the CSTs makes them less suitable for evaluating a student's progress over time or measuring the effectiveness of specific interventions. The California Department of Education's report titled *Explaining 2009 STAR Program Summary Results to the Public* states:

"STAR Program Test results can be compared within the same grade and subject… Comparisons should not be made between grades or subjects."

This passage clearly states that CST scores were not psychometrically designed for comparative analysis between grades. For example, a student's 4th and 5th grade CST scaled scores cannot be compared to see if their reading ability improved. This means that year-to-year changes in a student's scaled score cannot answer the question of how much individual reading growth that student experienced in one year of schooling. It also indicates that year-to-year changes in scaled scores for groups of students between grades cannot answer the question of how much reading growth has occurred between grades.

Fortunately, in addition to the CST scaled score, each student receives a proficiency level score. Unlike the scaled scores, changes in these proficiency levels can be compared year to year. There are five levels:

- 1 – Far Below Basic
- 2 – Below Basic
- 3 – Basic
- 4 – Proficient
- 5 – Advanced

Challenges of Analyzing CST Proficiency Level Scores

The existence of proficiency levels makes it possible to use CST results to look at student progress. However, there are several issues and limitations to keep in mind when considering these scores.

1) Ordinality

Because proficiency levels are ordinal scores, the categories may be of different sizes (e.g. the 'Basic' category may encompass a wider range of CST scores than the 'Proficient' category). Furthermore, it is inappropriate to conduct arithmetic operations on these scores, such as calculating the "average level" for a group. The only thing we know for sure about these scores is that a '5' is greater than a '4', a '4' is greater than a '3', etc. One possible orientation for CST proficiency levels is shown in Figure 2, below.

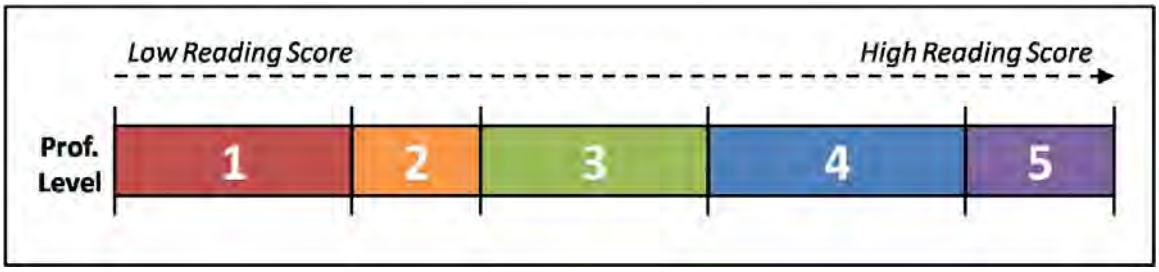

Figure 2. Conceptual diagram of CST proficiency levels. Note: Not necessarily to scale.

2) Low Resolution

Analysis of these ordinal scores is further complicated by the fact that there are only five levels. This provides a very *low-resolution* view of student growth. Students might make significant reading gains, but those gains might not have moved them across a border between proficiency thresholds. **Gains of this nature may be significant and meaningful, but they cannot be captured by looking at proficiency level changes.**

When an individual student moves up or down a proficiency level, that information doesn't tell us whether the gain/loss is statistically significant – the change could be due to random fluctuation in test performance between administrations. Additionally, it can be misleading to count the number of proficiency levels gained or lost between tests – some two-level gains might actually be *smaller* than some one-level gains. See Figure 3 for an example.

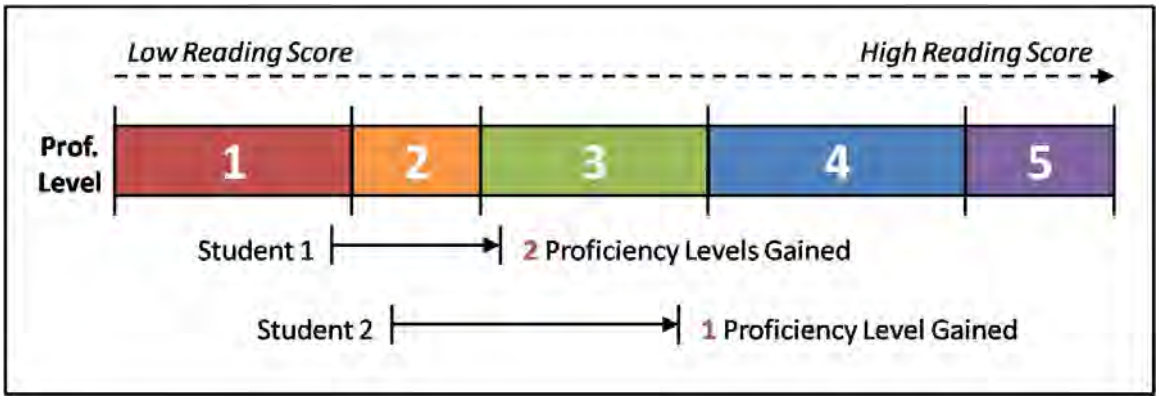

Figure 3. Measuring the number of levels gained may be misleading

3) Ceiling and Floor Effects

The categorical scale bottoms out at 1 and tops out at 5. Students who are in the highest category cannot score any higher (ceiling effect); those who are in the lowest category cannot score any lower (floor effect). These limitations do not invalidate analyses of CST proficiency levels, but it is important to keep them in mind, especially for those students who start on the high or low end of the CST proficiency spectrum.

Testing Variability Issues

When students take a test, their performance may not reflect their true ability level. On any given day, there is a good chance that their performance will be close to their true ability, and a small chance their performance will be far from their true ability. Actual performance is influenced by testing conditions, environmental factors, student preparation and mindset, and other factors (e.g., a child is coming down with the flu or spent the previous night at a slumber party). On average, a student's performance will reflect their true ability, but any individual test performance is variable. Figures 4 and 5 show two conceptual examples of testing variability.

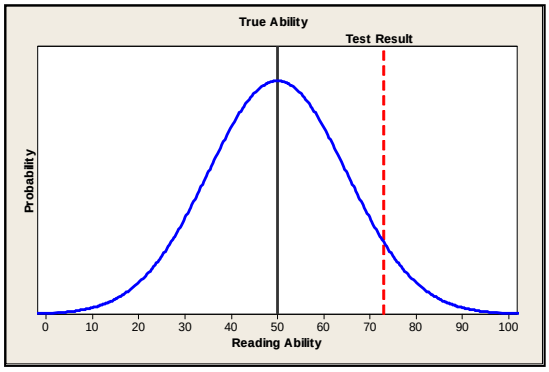

Figure 4. Example of a student whose test result has exceeded her true ability.

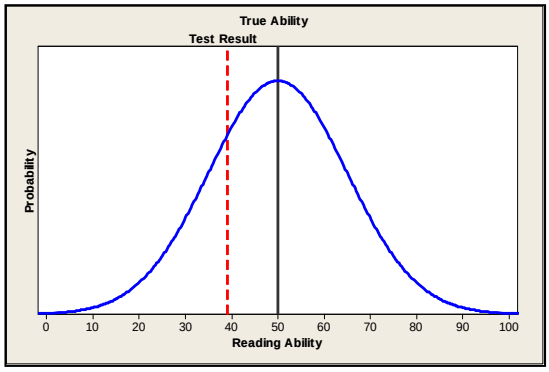

Figure 5. Example of a student whose test result has fallen short of her true ability.

Whenever groups of students are pre- and post-tested, even in the absence of an intervention, some students will show increases in their scores and some will show decreases. These changes may be due to the variability of the testing process, not to any real change in the student's true ability. As Figures 3 and 4 imply, our assumed model of test-taking variability is that, on average, students are as likely to over-perform as under-perform. Our general assumption is that test performance is symmetrically distributed around a student's true ability.

Analysis Questions

In evaluating the effectiveness of an intervention, the key question is whether student performance has increased more than would be expected due to testing variability. The following recommendations provide two alternatives for answering this question in light of the limitations of the CSTs.

Recommendations

Recommendation 1: Alternative Assessments

Because the CST only permits categorical analysis of reading levels year-to-year, it cannot provide a nuanced, high-resolution view of individual student growth. Scientific Learning has compiled a list of assessments that are appropriate for a variety of grade levels and Fast ForWord sequences and that measure specific language, cognitive, and reading skills with high precision and validity. We recommend that those interested in quantifying the benefits of Fast ForWord products in California schools pre- and post-test students with one of these high-resolution assessments in addition to the CSTs.

A table containing these recommendations can be found in Appendix A.

Recommendation 2: Randomization Tests for Proficiency Levels

Our preferred method for analyzing proficiency score changes is a Monte Carlo implementation of a Non-Parametric Randomization Test (McNPR test). Despite its fancy name, this test is really rather intuitive. The test operates on a group of students with two years of CST proficiency level data, and divides those students into three groups:

- Those who ***increased*** their proficiency level on the second test (e.g, a student who was at Level 3 on the first test and Level 4 on the second test).
- Those who ***decreased*** their proficiency level on the second test (e.g, a student who was at Level 5 on the first test and Level 3 on the second test).
- Those who ***maintained*** the same proficiency level from the first test on the second test.

The McNPR test evaluates the likelihood of the "null hypothesis", or the hypothesis that the educational intervention has no impact on student CST performance. If the null hypothesis is true, that means that any changes in student proficiency levels are due to other random factors – probably testing variability. Our assumed model of testing variability suggests that, on average, students are as likely to over-perform as under-perform[1], so we would expect to see roughly similar numbers of students increase their proficiency level as decrease their proficiency level. Figure 6 shows a possible distribution of 1,000 students' change groups if the null hypothesis were true.

[1] One could argue that this makes our test semi-parametric as opposed to non-parametric, although our assumption is that test-performance is from a *class* of distributions (i.e. distributions symmetrical around the student's true ability) rather than from a *specific* distribution.

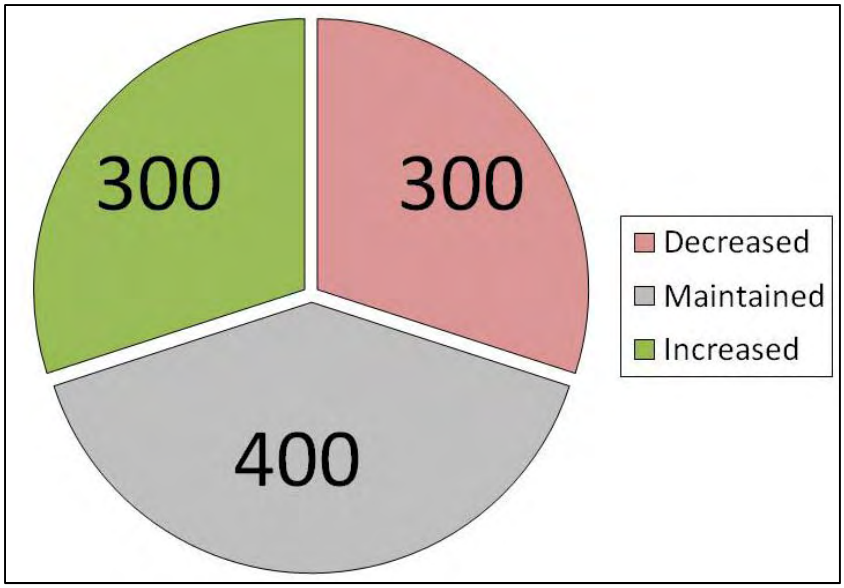

Figure 6. Sample distribution of students under the null hypothesis

If, on the other hand, the null hypothesis was clearly false and the intervention was pushing students towards higher proficiency levels, the distribution might look like Figure 7.

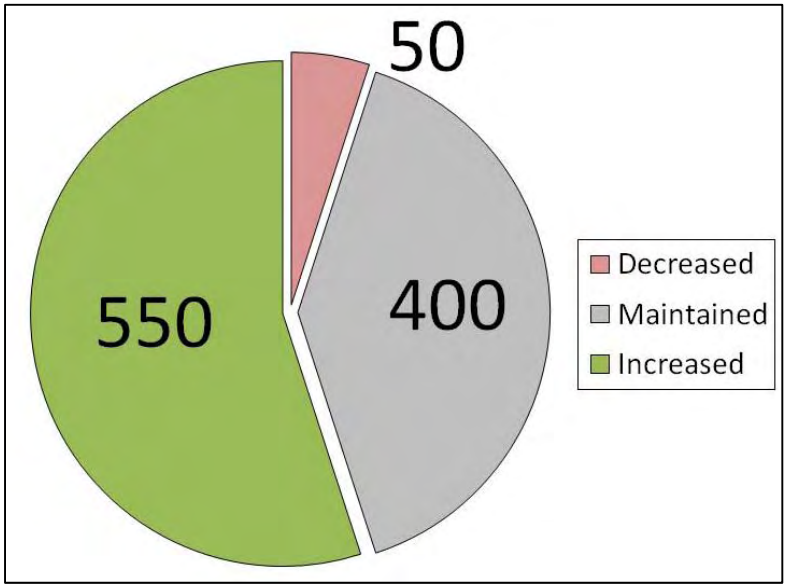

Figure 7. Sample distribution of students under the alternative hypothesis

We expect to see some distribution in between these two extremes. The McNPR test tries to decide if the observed distribution of proficiency level changes is more like the former example (random fluctuation) or the latter example (intervention has an effect). Details of the McNPR test are available in Appendix B.

McNPR Test Example

Assume we have 1,000 students with two years of CST test scores, with Fast ForWord used in between the two tests. The parameters for this example are:

$n_{increased}$ = 231 students
$n_{decreased}$ = 186 students
$n_{maintained}$ = 583 students
n_{total} = 1,000 students

The distribution of students is represented in the following chart (Figure 8).

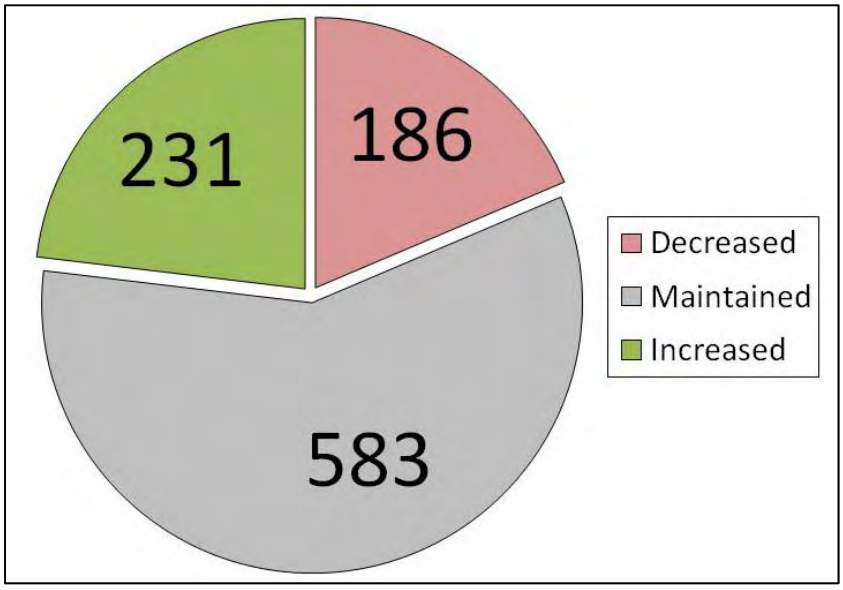

Figure 8. Sample distribution of 1,000 students

The number of students increasing their proficiency level is large relative to the number of students decreasing. The McNPR test indicated that a spread this large is very unlikely under the null hypothesis, which leads us to conclude that there is significant momentum towards improving CST scores for these students. We infer that this upward trend extends not just to those students who crossed a proficiency threshold, but to the majority of students who maintained their proficiency level as well – we expect that the low resolution of the CST proficiency level scores has obscured their growth. The McNPR test calculations for this example can be found in Appendix C.

Limitations of Other Analysis Approaches

There are several alternative statistical approaches one might consider that are, on reflection, not adequate for CST proficiency level datasets.

One might be tempted to use a test of two binomial proportions to see if a significantly larger proportion of students test at proficiency (i.e., proficiency level 3 or higher) on

the second test than the first test. This approach is low resolution (it effectively reduces the number of categories from five to two), it ignores gains that do not cross the proficiency threshold, and it doesn't consider the paired nature of the year-over-year CST scores.

Similarly, a chi-squared analysis does not consider the paired nature of the data, and proficiency level data may violate the test's requirement for at least five observations in each cell. Additionally, the chi-squared test makes parametric assumptions about the underlying distribution of CST scores which may be violated by some datasets.

Conclusion

The California Standards Tests (CSTs) are not designed to evaluate year-to-year changes in students' learning trajectories. Cross-grade comparisons are not permitted with CST scale scores. While such comparisons are permitted with proficiency level scores, these scores are limited by ordinality, low resolution, and ceiling and floor effects. Optimally, studies investigating how interventions impact students' learning trajectories will utilize alternative assessments that are better suited to serve as outcomes measures. When CST proficiency levels are the only available information for analyzing student gains, the strongest statistical analysis can be done using a Monte Carlo implementation of a Non-Parametric Randomization Test (McNPR test).

Appendix A: List of Recommended Alternative Assessments

When monitoring student progress, or evaluating the outcome of an intervention, it is important to select assessments suited to the product(s) being evaluated, the skills being monitored, and the testing format (individual- or group-administration). The following assessments are recommended for students using Scientific Learning products.

Fast ForWord Language/Literacy				
Test	**Administration**	**Publisher**	**Age/Grade**	**Skills**
Clinical Evaluation of Language Fundamentals (CELF)	Individual	Pearson	Ages: 5.0-Adult	Language (Receptive, Expressive), Cognitive (Memory, Sequencing)
Comprehensive Test of Phonological Processing (CTOPP)	Individual	Pro-Ed	Ages: 5.0 – 24.11	Cognitive (Phonological Awareness, Memory, Rapid Naming)
Oral and Written Language Scales (OWLS)	Individual	Pro-Ed	Ages: 5.0 – 21.0	Language (Receptive, Expressive)
Phonological Awareness Test (PAT)	Individual	LinguiSystems	Ages: 5.0 – 9.11 / Grades: K – 4th	Cognitive (Phonological Awareness)
Reading Progress Indicator (RPI)	Computer	Scientific Learning	Grades: K - Adult	Early Reading Skills
Test of Auditory Comprehension of Language (TACL)	Individual	Pearson	Ages: 3.0 – 9.0	Language (Receptive)
Test of Language Development (TOLD)	Individual	Pearson	Ages: 4.0 – 17.11	Language (Receptive, Expressive)
Test of Phonological Awareness (TOPA)	Group	Pro-Ed	Ages: 5.0 – 8.11 / Grades: K – 3rd	Cognitive (Phonological Awareness)
Fast ForWord Language to Reading/Literacy Advanced				
Test	**Administration**	**Publisher**	**Age/Grade**	**Skills**
Clinical Evaluation of Language Fundamentals (CELF)	Individual	Pearson	Ages: 5.0-Adult	Language (Receptive, Expressive), Cognitive (Memory, Sequencing)
Comprehensive Test of Phonological Processing (CTOPP)	Individual	Pro-Ed	Ages: 5.0 – 24.11	Cognitive (Phonological Awareness, Memory, Rapid Naming)

Oral and Written Language Scales (OWLS)	Individual	Pro-Ed	Ages: 5.0 – 21.0	Language (Receptive, Expressive)
Phonological Awareness Test (PAT)	Individual	LinguiSystems	Ages: 5.0 – 9.11 / Grades: K – 4th	Cognitive (Phonological Awareness)
Reading Progress Indicator (RPI)	Computer	Scientific Learning	Grades: K - Adult	Early Reading Skills
Test of Auditory Comprehension of Language (TACL)	Individual	Pearson	Ages: 3.0 – 9.0	Language (Receptive)
Test of Language Development (TOLD)	Individual	Pearson	Ages: 4.0 – 17.11	Language (Receptive, Expressive)
Test of Phonological Awareness (TOPA)	Group	Pro-Ed	Ages: 5.0 – 8.11 / Grades: K – 3rd	Cognitive (Phonological Awareness)
Woodcock Reading Mastery Test (WRMT)	Individual	Pearson	5.0 – 75+	Reading
Fast ForWord Reading				
Test	**Administration**	**Publisher**	**Age/Grade**	**Skills**
Gates-MacGinitie Reading Tests	Group	Riverside Publishing	Grades: K - Adult	Reading (Vocabulary, Comprehension)
Reading Progress Indicator (RPI)	Computer	Scientific Learning	Grades: K - Adult	Early Reading Skills
TerraNova	Group	CTB/McGraw-Hill	Grades: K – 12th	Reading
Reading Assistant				
Test	**Administration**	**Publisher**	**Age/Grade**	**Skills**
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)	Individual	University of Oregon Center on Teaching and Learning	Grades: K – 3rd	Reading (Fluency)
Gates-MacGinitie Reading Tests	Group	Riverside Publishing	Grades: K - Adult	Reading (Vocabulary, Comprehension)
Gray Oral Reading Test (GORT)	Individual	Pearson	Ages: 6.0 – 18.11	Reading (Fluency)
Test of Word Reading Efficiency (TOWRE)	Individual	Pearson	Ages: 6.0 – 24.11	Reading (Fluency)

Appendix B: Monte Carlo Non-Parametric Randomization Test

The McNPR test has the following parameters:

$n_{increased}$ = Number of students who increased one or more proficiency levels
$n_{decreased}$ = Number of students who decreased one or more proficiency levels
$n_{maintained}$ = Number of students who maintained the same proficiency level
$n_{total} = n_{increased} + n_{decreased} + n_{maintained}$
m = Number of simulations (typically 10,000 to 100,000)
α = the significance threshold for the statistical test (typically 0.05)

The McNPR test empirical determines the probability that a particular observation moved. The test assumes that the null hypothesis is true (until proven otherwise). Under this preliminary assumption, it is equally likely that an observation will move up as will move down, so the observed probability of movement is:

Eq. (1)

$$\hat{p} = \frac{n_{increased} + n_{decreased}}{n_{total}}$$

Under a normal approximation to the binomial distribution[2], the standard error of $\hat{p}$ is:

Eq. (2)

$$s_{\hat{p}} = \sqrt{\hat{p} \times (1 - \hat{p})}$$

The test statistic for the McNPR test is the observed difference between the number of observations that increased and the number of observations that decreased:

Eq. (3)

$$\omega_{observed} = n_{increased} - n_{decreased}$$

The McNPR test then runs the following simulation to empirically determine how extreme the observed results are:

[2] One could argue that this makes our test semi-parametric as opposed to non-parametric, although our assumption is that test-performance is from a *class* of distributions (i.e. distributions symmetrical around the student's true ability) rather than from a *specific* distribution.

1. Generate n_{total} identical student records.
2. Determine the probability of movement for this simulation iteration using a normal approximation the binomial distribution. Randomly select $\hat{p}_{sim}$ from a normal distribution with mean $\hat{p}$ and standard deviation $s_{\hat{p}}$.[3]
3. For each student, flip a coin to randomly determine whether they move (heads) or don't move (tails). The probability of the coin coming up heads should be the randomly selected $\hat{p}_{sim}$from the previous step.
4. Now that we have separated the students into a group of movers and non-movers, flip a coin to randomly determine whether the movers increased (heads) or decreased (tails). The probability of the coin coming up heads should be 0.5 – consistent with the null hypothesis that increase and decreases are due to equally random variation in testing.
5. Calculate the difference between the number of students who improved and students who declined. Call this value ω_i, where *i* is the simulation iteration number.
6. Repeat simulation steps 1 through 5 a total of *m* times.
7. Determine the percentile of $\omega_{observed}$ in the distribution of the *m* simulated ω_i's. If the percentile is in the most extreme α of the distribution (the one-sided α for a one-tailed test; either the upper or lower $\alpha/2$ for a two-tailed), reject the null hypothesis. Otherwise, fail to reject the null hypothesis.

The R code for the McNPR test is included in Appendix D.

[3] The normal approximation to the binomial distribution is generally quite good. However, it is less precise when the total number of observations is small (particularly when $\hat{p}$ is also very close to 0 or 1). Thus, using the McNPR test on small samples is not recommended; even though other corrections may be suitable (e.g., a Wilson score), conclusions from small samples are hard to generalize.

Appendix C: Calculations for the McNPR Test Example

For the McNPR example presented in the text, the parameters are:

$n_{increased}$ = 231 students
$n_{decreased}$ = 186 students
$n_{maintained}$ = 583 students
n_{total} = 1,000 students
m = 10,000 simulations
α = 0.05 significance level

For this example:

$$\hat{p} = \frac{n_{increased} + n_{decreased}}{n_{total}} = \frac{231 + 186}{1{,}000} = 0.417$$

$$s_{\hat{p}} = \sqrt{\hat{p} \times (1 - \hat{p})} = \sqrt{0.417 \times (1 - 0.417)} = 0.493$$

$$\omega_{observed} = n_{increased} - n_{decreased} = 231 - 186 = 45$$

The McNPR test determined that the empirical p-value for $\omega_{observed}$ was 0.015, which means that 45 is the 98.5th percentile of the distribution of ω under simulation. The distribution of ω and the placement of $\omega_{observed}$ is shown below in Figure 9.

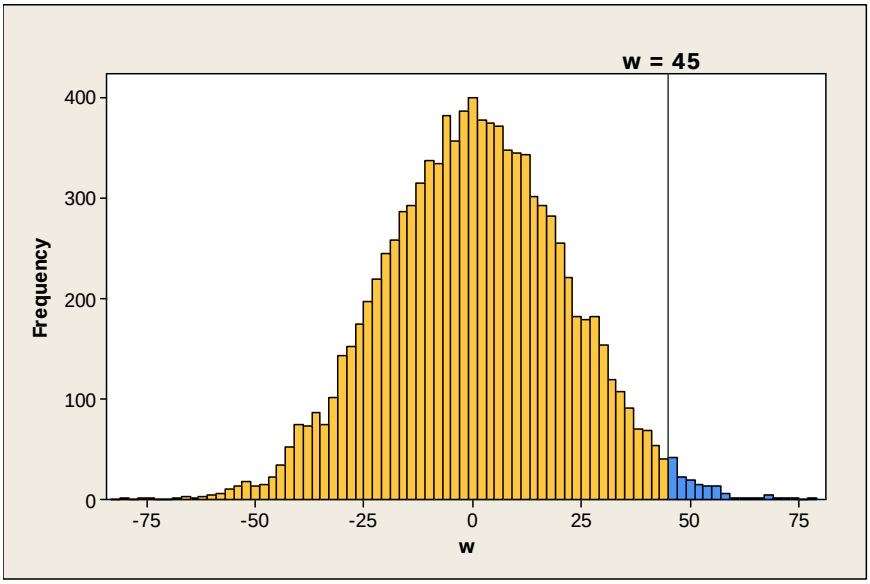

Figure 9. Distribution of simulated ω values and the location of $\omega_{observed}$

This simulation indicates that $\omega_{observed}$ is an extreme result under the null hypothesis (either one-sided or two-sided). Consequently, we reject the null hypothesis and conclude that the Fast ForWord intervention had a statistically significant positive impact on the CST performance of this group of students.

Appendix D:
R Code for the Monte Carlo Non-Parametric Randomization Test

The following code will run a Monte Carlo Non-Parametric Randomization Test in the free, open-source statistics package R (available for download at www.r-project.org).

```
# Code Start

# parameters

n.dec <- 186
n.inc <- 231
n.maint <- 583
n <- sum(n.dec, n.inc, n.maint)
w.obs <- n.inc - n.dec
m <- 10000

# determine p.hat distribution
p.hat <- (n.dec + n.inc) / n
p.se <- sqrt((p.hat*(1-p.hat))/n)

MyResults <- data.frame(N.maint = numeric(m),
                        N.dec = numeric(m),
                        N.inc = numeric(m))
# Simulation Loop
for (i in 1:m) {

        # determine p.hat for this simulation
        p.move <- rnorm(1, p.hat, p.se)
        p.maint <- 1-p.move

        # drop obs into move bins (0=maintain, 1=move)
        BinNum <- sample(0:1, n, replace = TRUE, prob = c(p.maint, p.move))

        # assign them to gains or losses
        Side <- rbinom(n, 1, .5)
        Side[Side == 0] <- -1

        # calculate final bins
        MyBin <- BinNum * Side

        # populate restuls
        MyResults$N.maint[i] <- length(MyBin[MyBin == 0])
        MyResults$N.dec[i] <- length(MyBin[MyBin < 0])
        MyResults$N.inc[i] <- length(MyBin[MyBin > 0])
}

# results
MyResults$w <- MyResults$N.inc - MyResults$N.dec

obs.percentile <- length(MyResults$w[MyResults$w >= w.obs])/m
obs.percentile

# write out results
write.table(MyResults, file = "Out.csv", sep = ",", row.names = FALSE, col.names = TRUE)

# Code End
```


Version 2010.04.08