Using difference-in-differences to compare cohort-level TOEIC L&R scores

Jean-Pierre J. Richard
richard.jean-pierre@u-nagano.ac.jp
The University of Nagano

Abstract

The effects of two years of online classes, due to measures put in place to prevent the spread of COVID-19 among university students in Japan, remains largely unknown. This study investigated four cohorts of students (N = 854) at a prefectural university in regional Japan who completed the Test of English for International Communication, Listening and Reading (TOEIC L&R) at the start and end of their two-year required English language program. TOEIC gain scores were examined in relation to expected regression toward the means and to the standard error of difference. These data were also analyzed using a difference-in-differences quasi-experimental design. Key findings indicated that all four cohorts made large, significant gains. Members of the cohort that experienced two years of online learning, however, made significantly fewer gains compared with the other three cohorts on Listening, and significantly fewer gains compared with its successor cohort on Reading. Overall, the findings suggest that the two years online had a deleterious effect with regard to TOEIC L&R scores. However, several important limitations were addressed. Most importantly, the quality of the online TOEIC L&R when it was introduced in 2020 leaves some doubt about the conclusions drawn.

Keywords: TOEIC, online learning, COVID-19, regression to the mean, standard error of difference, difference-in-differences, pretest-posttest data, Japanese university students

The COVID-19 pandemic affected education systems worldwide. In Japan, primary and secondary schools were temporarily closed in the Spring of 2020 (Iwabuchi et al., 2022). In contrast to short-term school closures, many universities implemented online classes, and the length of time between transitioning in and out of online learning varied. Many universities, in fact, continued to have online classes for two years, April 2020 to March 2022. At Shōzan University, a pseudonym, where this study took place, the cohort of students who entered the university in April 2020 experienced their first two years of university education online. The aim of this study is to estimate whether this cohort experienced gains at similar rates as other cohorts on a standardized test, the Test of English for International Communication, Listening and Reading (TOEIC L&R[1]), which the university uses for quality assurance of its English program.

Literature Review

Media reports, as early as the spring of 2020, indicated that much learning was lost as classes shifted from in-person learning to online learning: "Learning loss from Covid-19 is a national catastrophe" (Harvard Crimson, 2022, Para. 2). Academic studies have mostly concurred with these media reports. In a meta-analysis of 42 studies from 15 countries, it was estimated that 35% of a typical year of learning was lost during the pandemic (Cohen's d = -0.14, 95% CI -0.17 to -0.10), and losses were greatest for children from lower socio-economic backgrounds (Betthäuser et al., 2023).[2] In the United States, math and reading scores fell between .20–.27 SDs and .09–.18 SDs (Kuhfeld, et al., 2022). In the Netherlands and Germany, primary school children were estimated to have lost up to one-fifth of the school year (0.08 SD) (Engzell et al., 2021; Schult, et al., 2022). In Brazil, a difference-in-differences analysis estimated that losses for public secondary school students were 0.32 SDs (Lichand et al., 2022). In Mexico, losses were 0.34–0.45 SD in reading and 0.62–0.82 in mathematics (Hevia et al., 2022). In Sweden, where schools did not close, no losses were observed (Hallin et al., 2022).

In Japan, primary and secondary school closures began in March 2020, near the end of the 2019–2020 academic year and, depending on the region, continued into the start of the next academic year, until the middle of May in most of Japan, and the end of May in greater Tokyo. Once opened, schools reduced class time, class size, and staggered hours to reduce contacts (Iwabuchi et al., 2022). "The school closure, despite being temporary, led to a huge disparity in student learning between schools" (Iwabuchi et al., 2022, p. 127). However, it was thought that by December 2020, most schools had caught up with the curriculum, and in January 2021, the national Common Test for University Admissions was held as scheduled (Iwabuchi et al., 2022). Online services for junior and senior high school students in Japan "mitigated the negative consequences" of school closures (Ikeda & Yamaguchi, 2021), p. 472); nonetheless, differences due to home resources and school quality were observed.

Most large-scale studies of pandemic-related effects on education have focused on primary and secondary education; however, effects were also experienced by students in higher education with lower enrollments, less study time, lower

graduation rates, and loss of jobs and job offers (Aucejo et al., 2020). American students across multiple universities, completing the same standard assessment, saw average losses of 0.2 SDs after online learning was introduced (Orlov et al., 2021). University students were overwhelmed with online assignments when closures began, and this had a detrimental effect on learning outcomes (Motz et al., 2021). Students were given "busywork" (Motz et al., p. 79), which students believed did not relate to improved learning, leading to demotivation, missed deadlines, lack of uptake in feedback, and academic failure; moreover, this busywork collided with daily lives, such as pandemic-affected full-time care for family members and employment struggles. The decrease in motivation and performance of university students was also attributed to a lack of infrastructure necessary to support learning and a loss of support from classmates and instructors (Tan, 2021). In addition, EFL learners initially expressed positive attitudes to online learning, but preferred face-to-face classrooms, and many lost interest and motivation, which was thought to be attributed to a lack of online classroom participation or engagement from peers (Sukman & Mhunkongdee, 2021. Grit and resilience were factors to help keep first-year university students focused on goals (Lytle & Shin, 2022).

Most universities in Japan postponed classes and/or began remote classes in April 2020, at the start of the academic calendar (Kang, 2021). Many problems related to moving classes online as a result of the COVID-19 pandemic were observed. For example, first-year university students experienced elevated levels of academic distress (Horita et al., 2021), as well as psychological and economic stresses (Sato et al., 2023). Obara (2022) found that students entering Japanese universities in 2020 had difficulties establishing positive relationships with peers and others. On the other hand, one possible positive outcome was the potential for the use of blended learning models, in which students have greater choice in terms of delivery mode, location, and time. However, this might have been most relevant for older, graduate students (Shindo et al., 2022). In a large-scale difference-in-differences study, student-level course evaluations showed improvements in 2020 from previous years, but when considering instructor-level teaching-quality, the improvements in overall course evaluations were weak or nonexistent, which suggested that gains in evaluations were attributable to greater choice in class-taking for students and not to improvements in teacher quality (Kashima & Yamamoto, 2021).

Nagata (2022) reported on the effects of an online learning system on TOEIC scores among two groups of university learners in a study that had begun before COVID-19 forced universities to go online. The online system Nagata described included both one-to-one online English conversation and other L2 English-language training software. On average, students increased their TOEIC Listening[3] and Reading scores by 0.52 and 0.35 points per hour of study, gains that were approximately twice as large as gains in face-to-face classes. However, there was much variation in usage of the online system; for example, on average students spent 80-100 hours using the one-to-one conversation component, but standard deviations were over 50 hours. In addition, the percentage of students whose scores increased beyond the standard error of difference (SE_{diff}) was small, 23% for the online learners but 29% for the classroom-based learners.

The inclusion by Nagata (2022) of a discussion of SE_{diff} was useful; however, he did not address regression toward the mean (RTM). Koizumi et al. (2015) argued that it was important to consider both RTM and SE_{diff} in pretest-posttest research designs in order to address the issue of actual gains, or lack thereof, in ability. The aim of the present study is to estimate whether the group of students who entered university in April 2020 experienced growth in posttest TOEIC scores from their pretest scores at comparable rates to other cohorts. To meet this aim, two research questions will be addressed.

Research Questions:

1. Is there evidence of RTM in Time 1–Time 3 data? If yes, what percentage of students increased their scores beyond RTM, and what percentage of students increased their scores beyond the SE_{diff}?
2. By how much did each cohort grow in comparison with other cohorts?

Methods

Context and Participants

The study took place at a small public regional university in central Japan which opened in 2018, and has two faculties of non-English majors. One goal at the university is for all Year 2 students to participate in a short-term overseas program (OP) of two-to-four weeks for academic study in their major and English learning. Due to COVID-19, only the students from the first cohort went abroad. The second and third cohorts had online OP; the fourth and fifth cohorts will go abroad in 2023. To prepare for the OP, students complete the required two-year English Program for Global Mobility (EPGM). The EPGM consists of four 100-minute lessons per week in four 7-week academic quarters in Year 1, and two-to-four 100-minute lessons per week, depending on faculty, in three 7-week quarters in Year 2 (approximately 325 hours of class time

in the EPGM). Half of the lessons focus on English accuracy, primarily through the teaching of listening and reading, and half of the lessons focus on English fluency, primarily through the teaching of speaking and discussion. Because the university is new, only four cohorts have completed the EPGM. Courses and course objectives have remained consistent since its opening, in accordance with Ministry of Education guidelines for new academic institutions.

The university uses two standardized English-language tests: the Computerized Assessment System for English Communication (CASEC) for class placement, and TOEIC is used for quality assurance. Students complete CASEC in mid-to-late March before entering university, and TOEIC three times: the beginning (Time 1) and end of Year 1 (Time 2), and the end of Year 2 (Time 3). Students in Cohort 1, however, did not take the test at Time 2. Table 1 displays basic characteristics for Cohorts 1 to 4, including combined TOEIC Listening and Reading scores at Times 1 and 3. Incoming cohorts had similar mean scores on CASEC (see the note in the table). Publicly available national cohort data show that TOEIC scores rose dramatically in 2020 when the online test was introduced, then fell somewhat in 2021 but were still much higher than previous scores (IIBC, 2022). The TOEIC mean scores reported in Table 1 appear to mirror these national cohort results.

Table 1
Cohort characteristics and mean scores for CASEC and TOEIC L&R

	Cohort 1 (n = 211)	Cohort 2 (n = 217)	Cohort 3 (n = 216)	Cohort 4 (n = 210)
Year of entry	2018	2019	2020	2021
CASEC (M)[1]	564	574	577	571
Year 1 classes	face-to-face	face-to-face	online	online
Year 2 classes	face-to-face	online	online	face-to-face
Overseas Program	overseas	online	online	June 2023
TOEIC 1 (M)	paper (413)	paper (417)	online (507)	online (465)
TOEIC 3 (M)	paper (556)	paper (547)	online (620)	online (627)

Note. [1] A one-way ANOVA tested for cohort-level differences on CASEC scores. The Levene's test showed that there was equal variance for all four groups: F(3,850) = 1.81, p =.144. The ANOVA showed no significant effect of cohort on CASEC scores; F(3, 850) = 1.44, p = .231 (with trivial effect sizes: η^2 = .005, η^2_p = .002).

As reported above, the students are non-English majors in two faculties, Economics and Human Sciences, both pseudonyms, with students in the former comprising approximately 70% of the students at the university. Approximately 50% of the students are from the prefecture where the university is located, 1% are students from other Asian countries (e.g., Taiwan, Malaysia), and the remainder are from other prefectures in Japan. In all, 903 students in Cohorts 1-4 had complete data sets. However, there were 49 outliers (approximately 5%), that were spread across the four cohorts, of which 27 were low-scorers and 22 were high-scorers, including all of the international students: [Cohort 1: 12 students (5 lower, 7 upper); Cohort 2: 11 (6, 5); Cohort 3: 14 (9, 5); Cohort 4: 12 (7, 5)]. (A 4 x2 χ^2 test for independence was carried out to assess the distribution of outliers across cohorts. The test was not significant: χ^2 (3) = 1.40, p = .70, indicating independence of variables.) After removing outliers, the data set consisted of 854 students.

Data Analyses

Similar to Koizumi et al. (2015), means, standard deviations, correlations, and paired-sample t-tests are first reported. Note that all analyses were carried out using JASP, Version 0.17.1 (JASP Team, 2023).

Regression toward the Mean (RTM). RTM was investigated at the group level and individual level. At the group level, differences in scores between the posttest and pretest were calculated for each student. Then, these differences were used in correlation analyses with the pretest scores to investigate RTM. If present, large negative correlations will be observed. At the individual level, expected scores at Time 3 were calculated for each student, and these were compared with actual outcomes. Equation (1) was used.

$$\text{Expected posttest score} = M_y + r_{xy}(SD_y/SD_x)(X - M_x) \tag{1}$$

where M_y is the mean score at Time 3, r_{xy} is the correlation between scores at Time 1 and Time 3, SD_y and SD_x are the standard deviations at Time 3 and Time 1 respectively, M_x is the mean score at Time 1, and X is the score for each individual at Time 1. See Appendix A for expected posttest score calculations. In addition, χ^2 tests for independence were carried out

to investigate whether the number of students gaining or not gaining were similar in each cohort.

Standard Error of Difference (SE_{diff}). The percentage of students with score gains beyond the SE_{diff} were estimated, using the probability score of 68%, with equation (2).

$$SE_{diff} = (Time\ 1\ SD) * (\sqrt{[2 - (Reliability\ at\ Time\ 1) - (Reliability\ at\ Time\ 2)]}) \qquad (2)$$

The TOEIC score user guide from Educational Testing Service (ETS) reported a reliability coefficient of $\alpha = .90$ for both Listening and Reading (ETS, 2022, p. 19); however, ETS researchers reported a reliability coefficient of $\alpha = .94$ (Cid et al., 2017, p. 11) for a large-scale study of test-takers in Japan and Korea specifically. The reliability coefficient from Cid et al. was used.[4] For calculations, see Appendix B. A second calculation using the known SE_{diff} (ETS, 2022) was undertaken. The percentage of students with score gains beyond this SE_{diff} (i.e., +35 for each of Listening and Reading) were calculated. In addition, χ^2 tests for independence were carried out to investigate whether the number of students gaining or not gaining were similar in each cohort.

Difference-in-differences

Difference-in-differences is a quasi-experimental statistical procedure (World Bank, 2023) frequently used in economics, management studies, and other social sciences (see, for example, Fredriksson & Magalhães de Oliveira, 2019) to calculate differences in outcomes between groups. The procedure controls for starting points and assumes similarities across groups, including parallel slopes, and estimates average treatment effects (for the treatment group). In Figure 1, the treatment group (T_1) has an assumed similar slope (dotted line) to its expected outcome ($T_{expected}$) as the control group (C). However, the actual outcome of the treatment group ($T_{observed}$) is different and is attributed to the treatment effect. The estimated true treatment effect is thus not the difference between $T_{observed}$ and C_2, but rather the difference between $T_{observed}$ and $T_{expected}$,

Figure 1
Representation of a basic difference-in-differences model
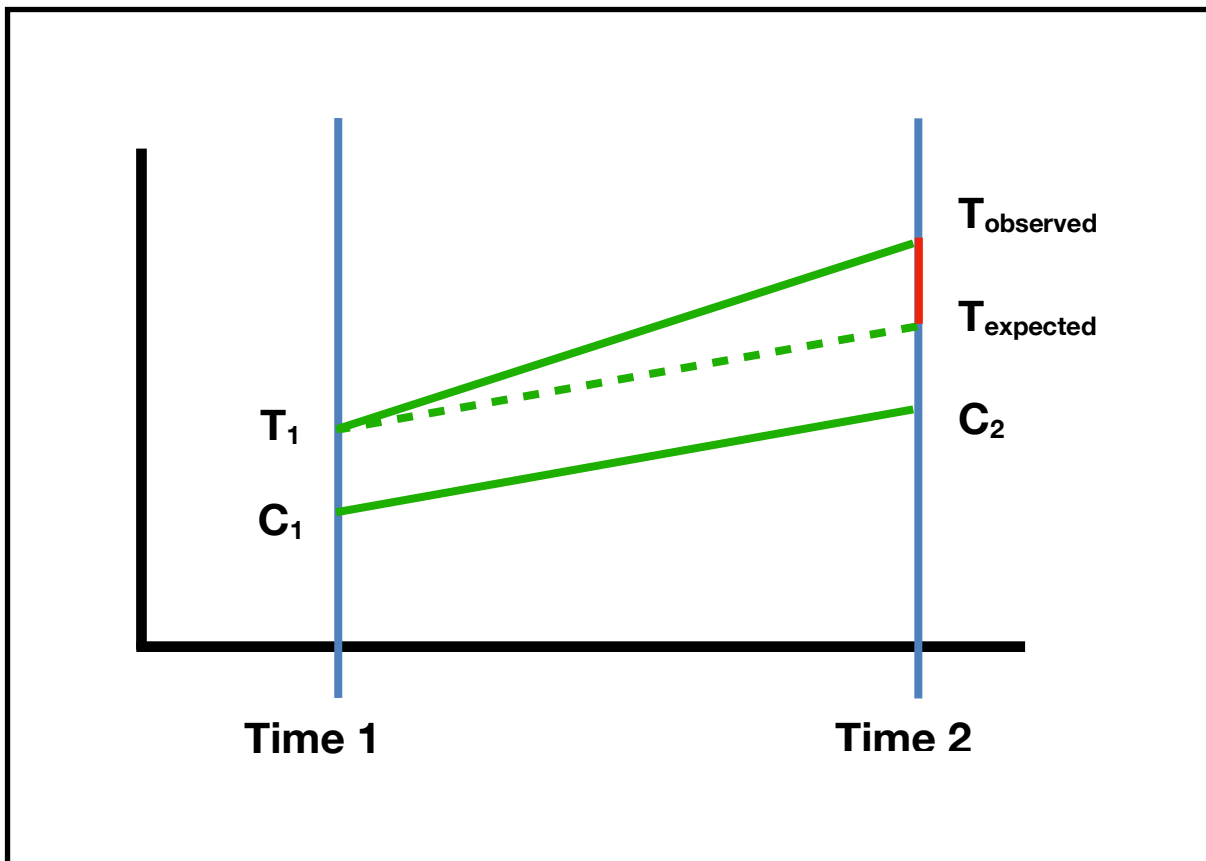
represented by the small red line.

Note that in the current study, the four cohorts have similar population characteristics: they began the two-year EPGM with similar CASEC scores, completed TOEIC at Times 1 and 3, and experienced similar English courses. Thus, any differences in final outcomes between the cohorts are assumed to be related to lesson type: face-to-face or online.

Difference-in-differences can be calculated using equation (3).

$$\delta_{DD} = \bar{Y}_{Treat,2} - [\bar{Y}_{Treat,1} + (\bar{Y}_{Control,2} - \bar{Y}_{Control,1})] \tag{3}$$

where δ_{DD} represents the difference in differences; $\bar{Y}_{Treat,1}$ and $\bar{Y}_{Treat,2}$ represent the average of the treatment group at Times 1 and 2, and $\bar{Y}_{Control,1}$ and $\bar{Y}_{Control,2}$ represent the average of the control group at Times 1 and 2. The difference-in-differences analysis can be demonstrated with the sample mean scores for the Treatment and Control groups at the Pretest and the Posttest that are displayed in Table 2, with the counterfactual showing the starting point of the Treatment group but the growth rate of the Control group if the growth rates were similar. In other words, the counterfactual shows the expected growth for the treatment group if the treatment had not occurred. This counterfactual addresses the assumption that growth over time for both groups would be parallel. Using the sample data and equation (3), the difference in differences is: 330 - [281 + (342 - 268)] = -25.

Table 2
Example for calculating difference in differences

Test	Treatment	Control	Counterfactual
Pre (Time 1)	281 ($\bar{Y}_{Treat,1}$)	268 ($\bar{Y}_{Control,1}$)	281 ($\bar{Y}_{Treat,1}$)
Post (Time 2)	330 ($\bar{Y}_{Treat,2}$)	342 ($\bar{Y}_{Control,2}$)	355 [$\bar{Y}_{Treat,1} + (\bar{Y}_{Control,2} - \bar{Y}_{Control,1})$]

Researchers can also calculate difference in differences using regression analysis, with dummy codes, for example, D_{TREAT}, (0 = no treatment; 1 = treatment), and D_{POST}, (0 = not post-treatment, 1 = post-treatment), as shown in equation (4).

$$Y_i = b_o + \delta_T D_{TREATi} + \delta_T D_{POSTi} + \delta_{DD}(D_{TREATi})(D_{POSTi}) \tag{4}$$

where Y_i represents the predicted outcome variable. The ordinary least squares estimate of δ_{DD} provides the average treatment effect. The output from a sample regression model is shown in Table 3, based on the same data that were used to calculate the means in Table 2. Difference in differences, -25.157 (p = 0.005), is the bolded unstandardized coefficient in the bottom row, and it is the same (rounded) as that shown with equation (3). The advantage of using the regression model is that it includes a significance value, confidence intervals, and allows for a causal inference.

Table 3
Example of using regression to calculate difference in differences

Model (H₁)	Unstandardized Coefficients	SE	p*	95% CI* Lower	Upper
(Intercept)	268.01	4.02	< .001	260.09	275.75
D_{TREATi}	12.54	5.51	0.021	1.76	23.42
D_{POSTi}	74.34	5.88	< .001	62.52	85.47
$D_{TREATi}*D_{POSTi}$	-25.16	8.21	0.005	-40.32	-8.19

In this current study, four dummy codes were used: three dummy codes for the four cohorts, and one for treatment (i.e., pre or post treatment). Note that when dummy coding for groups, there needs to be one less dummy code than groups (k -1) because the reference group is represented by the intercept, b_o (Jeon, 2015). If the number of dummy codes is equal to the number of groups, singularity (i.e., perfect correlation) due to a redundant variable will be present in the data (Tabachnick & Fidell, 2007). See Appendix C for dummy codes used in this study.

Results and Discussion

Table 4 displays the means and standard deviations for Cohorts 1–4 for TOEIC Listening and Reading. At Time 1, Cohort 3 had higher mean scores than other cohorts for both Listening and Reading. At Time 3, Cohort 4 had higher mean scores for Listening, and near similar mean scores as Cohort 3 for Reading. Between Times 1 and 3, each of the four cohorts saw large average gains in TOEIC scale scores for both Listening and Reading: Cohort 1: 75 points and 68 points; Cohort 2: 73

points and 58 points; Cohort 3: 55 points and 58 points; and Cohort 4: 75 points and 88 points. Cohorts 1, 2, and 4 experienced similarly large average gains for Listening, while Cohort 3 experienced smaller gains. Cohort 4 experienced the largest gains for Reading, and Cohorts 2 and 3 experienced the smallest average gains for Reading.

Table 4
Means (SDs) for cohorts 1–4 at times 1 and 3, for TOEIC listening and reading

	Listening M (SD)		Reading M (SD)	
Cohort	Time 1	Time 3	Time 1	Time 3
1 (n = 211)	235.7 (53.7)	310.7 (70.3)	177.4 (52.0)	245.4 (68.3)
2 (n = 217)	231.5 (50.2)	304.4 (60.3)	185.0 (47.2)	242.8 (64.3)
3 (n = 216)	279.5 (53.3)	334.0 (62.2)	227.7 (48.3)	285.8 (66.5)
4 (n = 210)	267.7 (52.0)	342.5 (59.9)	197.7 (60.3)	284.4 (74.6)

Intra-test correlations for both Listening-Listening, and Reading-Reading between Times 1 and 3, were quite similar for Cohorts 1, 2, and 4. Cohort 3, however, had much smaller correlations: Cohort 1: r = .55, r = .62 (r^2 = .30, r^2 = .38); Cohort 2: r = .53, r = .55 (r^2 = .28, r^2 = .30); Cohort 3: r = .36, r = .30 (r^2 = .13, r^2 = .09); and Cohort 4: r = .61, r = .59 (r^2 = .37, r^2 = .35). The strength of the correlations for Cohorts 1, 2 and 4 were two-to-four times that of Cohort 3. In short, the students in Cohorts 1, 2 and 4 were more consistent across the two years compared with Cohort 3.

Paired sample t-tests, with a Bonferroni adjusted alpha level of .006 per test (.05/8), compared the scores from Time 1 and Time 3. Table 5 shows that for each cohort, scores at Time 3 for both Listening and Reading were significantly higher than at Time 1. Cohorts 1, 2, and 4 had medium-to-large effect sizes for Listening and Reading, while effect sizes were small-to-medium for Cohort 3. (For effect sizes, see Plonsky & Oswald, 2014.)

Table 5
Cohort-level paired-sample T-tests with effect sizes (95% CIs) for TOEIC listening and reading, between Times 1–3

Cohort	Listening	Reading
1 (n = 211)	t(210) = -17.96, p < .001	t(210) = -18.13, p < .001
	-1.24 (-1.42, -1.06)	-1.25 (-1.43, -1.07)
2 (n = 217)	t(216) = -19.83, p < .001	t(216) = -15.52, p < .001
	-1.35 (-1.53, -1.16)	-1.05 (-1.22, -0.89)
3 (n = 216)	t(215) = -12.17, p < .001	t(215) = -12.28, p < .001
	-0.83 (-0.98, -0.67)	-0.84 (-0.99, -0.68)
4 (n = 210)	t(209) = -21.70, p < .001	t(209) = -20.04, p < .001
	-1.50 (-1.69, -1.30)	-1.38 (-1.57, -1.197)

Research Question 1

There was evidence of RTM at the group level for all four cohorts, for both Listening and Reading, respectively. Small negative correlations were observed between scores at Time 1 and change in scores at Time 3 for Cohorts 1 (-.26, -.18), 2 (-.33, -.21) and 4 (-.30, -.26). Moderate negative correlations were observed for Cohort 3 (-.47, -.41). In all, the group-level RTM effects were found to be small-to-moderate. In addition to group-level analysis, RTM was also investigated for individuals. Results are reported in Table 6. For Cohorts 1–4, approximately 50–53% had scores at Time 3 that were greater than their RTM expected (estimated) scores for both Listening and Reading. The number of students per cohort with scores greater than their RTM expected scores for both skills was 31–35%. No large differences were observed between skills or between cohorts. In three chi square tests, rows were represented by the number of students making gains greater than the expected RTM (True) or not (False); columns were the cohorts. The chi square statistics were small and had non-significant p-values.

There was also evidence of score gains beyond the SE$_{diff}$. As shown in Table 7, a large majority of students in all four cohorts made gains greater that the estimated SE$_{diff}$, for both Listening and Reading. For Listening, Cohorts 1, 2, and 4 experienced similar gains. Approximately 85% of the students in these cohorts had gains that were greater than the SE$_{diff}$. In contrast, only 73% of students in Cohort 3 did so. For Reading, the percentage of students who saw gains ranged from 76-81% across all four cohorts. The percentage of students who experienced gains on both skills was similar for Cohorts 1, 2 and 4, but much less for Cohort 3. The three chi square tests resulted in significant chi square statistics, largest for

Listening, and smallest for the combined score.

Regarding gains greater than the ETS SE_{diff}, more than three-quarters of the students in Cohorts 1, 2 and 4 saw gains in Listening. Cohort 4 also saw similar gains for Reading, whereas the percentage of students in Cohorts 1–3 who experienced gains in Reading ranged from 63-68%. The percentage of students who gained on both skills was greatest for Cohort 4, and smallest for Cohort 3. In Cohort 3, fewer than half of the students gained on both skills. Note that for all four cohorts, for both skills, the ETS SE_{diff} (i.e., 35 scale points) was greater than the estimated SE_{diff} (i.e., 16–21 scale points, depending on skill and cohort). The three chi square tests resulted in significant chi square statistics, which was largest for Reading. Based on the results above, students in Cohort 4 made the largest gains, followed by students in Cohorts 1 and 2, and lastly Cohort 3.

Research Question 2

Two multiple linear regression models using forward data entry, one each for Listening and Reading, were run to estimate the difference in differences between cohorts at Time 1 and Time 3. For Listening the model explained 31.4% of the variance in Listening scores, $F(7, 1702) = 111.04$, $p < .001$, $R^2 = .314$, and for Reading the model explained 29.4% of the variance, $F(7, 1702) = 101.23$, $p < .001$, $R^2 = .294$. Table 8, the table of coefficients, with Cohort 1 as the reference (i.e., baseline category), shows the unstandardized coefficients (i.e., the scores in TOEIC scale scores), including difference in differences (bolded and highlighted). The interpretations of the coefficients are shown below the table. The means at the ends of lines 1–8 and 9–16 are the means for Listening and Reading for each cohort at Times 1 and 3 that were shown previously in Table 4 (with small differences due to rounding).

     Listening:
1. Intercept: Cohort 1 m at Time 1 = 235.7
2. D1: Cohort 2 m at Time 1: 235.7 - 4.2 = 231.5
3. D2: Cohort 3 m at Time 1: 235.7 + 43.8 = 279.5
4. D3: Cohort 4 m at Time 1: 235.7 + 32.0 = 267.7
5. Post: Cohort 1 m at Time 3: 235.7 + 75.0 = 310.7
6. D1*Post: Cohort 2 m at Time 3: 235.7 - 4.2 + 75.0 - 2.2 = 304.4
7. D2*Post: Cohort 3 m at Time 3: 235.7 + 43.8 + 75.0 - 20.6 = 334.0
8. D3*Post: Cohort 4 m at Time 3: 235.7 + 32.0 + 75.0 -0.17 = 342.5

     Reading:
9. Intercept: Cohort 1 m at Time 1: 177.4
10. D1: Cohort 2 m at Time 1: 177.4 + 7.6 = 185.0
11. D2: Cohort 3 m at Time 1: 177.4 + 50.4 = 227.7
12. D3: Cohort 4 m at Time 1: 177.4 + 20.4 = 197.7
13. Post: Cohort 1 m at Time 3: 177.4 + 68.0 = 245.4
14. D1*Post: Cohort 2 m at Time 3: 177.4 + 7.6 + 68.0 - 10.2 = 242.8
15. D2*Post: Cohort 3 m at Time 3: 177.4 + 50.4 + 68.0 - 9.9 = 285.8
16. D3*Post: Cohort 4 m at Time 3: 177.4 + 20.4 + 68.0 + 18.6 = 284.4

In short, all four cohorts made large gains on Listening, but Cohort 3 made fewer gains. Cohorts 1, 2 and 4 had similar outcomes for Listening, gaining on average approximately 75 TOEIC scale points. The difference in differences in TOEIC scale points between Cohorts 1 and 2 (75 - 2.17 = 2 points) and Cohorts 1 and 4 (75 - 0.17 = 0 points) were small and similar, and these differences were not significant ($p = .785$, $p = .983$). Also, the difference in differences between Cohorts 4 and 2 was also 2 scale points (74.83 - 72.83). Note that this p-value was not calculated for this comparison but is estimated to be not significant. Cohort 3 gained, on average, 55 points. The difference in differences between Cohort 1 and 3 was approximately -21 TOEIC scale points (75 - 20.56), and this was significant ($p < .010$). The difference in differences between Cohorts 2 and 3, and 3 and 4 were similarly large, 18 scale points (72.83 - 54.44) and 20 scale points (74.83 - 54.44). Note that the p-values were not calculated for these latter two comparison, but they are estimated to be significant.

Also, while all four cohorts made large gains on Reading, Cohort 4 showed the greatest gains at approximately 87 TOEIC scale points. The difference in differences between Cohort 1 and 2 was -10 (-10.20) scale points ($p = .221$), between Cohort 1 and 3 was -10 (-9.93) scale points ($p = .233$), and between Cohort 1 and 4 was 19 (18.60) scale points ($p = .027$). The difference in differences between Cohorts 4 and 2, and 4 and 3 were similarly large, at 29 scale points (86.61 - 57.81) and 29 scale points (86.61 - 58.08). Note that the p-values were not calculated for these two comparisons, but they are also

estimated to be significant.

Comparing gains for Listening, average gains for Cohorts 1, 2, and 4 were approximately 137% greater than the average gains made by Cohort 3. Comparing gains for Reading, average gains for Cohort 4 were approximately 150% greater than the average gains made by Cohorts 2 and 3, and 128% greater than Cohort 1.

Although combined TOEIC scores were not analyzed in detail above, a multiple linear regression, using forward data entry, was also run to estimate the difference in differences between cohorts at Time 1 and Time 3 in combined TOEIC scores. The model explained 35.4% of the variance in combined TOEIC scores, $F(7, 1702) = 133.33$, $p < .001$, $R^2 = .354$. On average Cohort 1 gained 143.0 TOEIC scale points, Cohort 2 gained 12.4 fewer points ($p = .392$), Cohort 3 gained 30.5 fewer points ($p = .035$), and Cohort 4 gained 18.4 more points ($p = .205$).

Based on the approximate number of class hours over two years (325), these gains per skill, per cohort, represent approximate hourly gains of 0.17–0.23 scale points for Listening, 0.17–0.27 scale points for Reading, and 0.35–0.50 scale points for combined TOEIC.[5] The results are summarized in Table 9. Cohorts 1, 2, and 4 made similar gains per class hour for Listening. For Reading, Cohort 4 made the most hourly gains. In all, the largest gains were made by Cohort 4, followed by Cohort 1. The fewest gains were made by Cohort 3, which might imply that two years of online classes had a deleterious effect on TOEIC scores for this cohort.

Limitations

There are several important limitations to the study. First, known differences among the students within cohorts were ignored for this study. For example, students in the two faculties perform differently on CASEC and TOEIC tests. On average, the scores from the Faculty of Economics are significantly higher than those of the Faculty of Human Sciences, and anecdotally, their motivations to learn English differ because their career paths differ greatly. In addition, there are also known differences between students who enter on general examinations (GT) and those who enter on selected or recommended examinations (RS) (i.e., GT > RS). However, these differences were ignored for this study because the percentage of students within each faculty, and percentage of students based on entrance type are similar among the four cohorts. In short, there are differences within cohorts, but the cohorts are parallel. However, it is plausible that the two years of online learning affected negatively one of these groups more than another. That is a question for future research.

A second limitation is that TOEIC data from Time 2 were not analyzed. Adding data from Time 2 might have helped to clarify how cohorts changed after Year 1 and Year 2 in the EPGM. Importantly, while Cohorts 2 and 4 both experienced one year of face-to-face classes and one year of online classes, they did so differently, Cohort 2 had face-to-face classes in their first year, whereas Cohort 4 had face-to-face classes in their second year. Cohort 3 had online classes in both years of the EPGM. Adding data from Time 2 might have helped to clarify how these differences in classes affected TOEIC scores. Unfortunately, Cohort 1, the cohort with two years of face-to-face classes, did not take TOEIC at Time 2, which adds further complications. In addition, the university where this study took place is a small university that only opened in 2018 and only four cohorts of students have completed the EPGM. Although CASEC scores have showed that incoming first-year students are similar on average, the university does not have a long history to compare cohorts from before COVID-19.

The final and perhaps most important limitation relates to the reliability of the online TOEIC, especially from 2020 when it was first used by a large number of institutions in Japan. To the best of my knowledge, ETS and IIBC have yet to release detailed research reports. Reports released to date have been limited to annual mean scores of the TOEIC L&R, which show that in 2020 when the online test was introduced, average TOEIC scores grew nationwide, but have since fallen (IIBC, 2022). As noted, this rising-falling pattern was mirrored in the present data. To be sure, the most recent TOEIC Score User Guide indicates that the reliability for the TOEIC L&R is unchanged, $\alpha = .90$ (ETS, 2022). However, a reasonable question is whether the results reported here for Cohort 3 are an artifact of a lower quality test in the spring of 2020 when Cohort 3 first took this test.

Conclusion

The current study aimed to investigate whether Cohort 3, a group which experienced only online classes during 2020–2022, experienced gains at similar rates as other cohorts on TOEIC at a small regional university in central Japan. To this end, gains, in relation to RTM and SE_{diff}, were examined. At the group level, Cohort 3 appeared to show more RTM than other cohorts, but at the individual level, a similar percentage of students, approximately 50% in each cohort, experienced gains greater than their expected RTM scores. In addition, approximately 75–85% of students in each cohort, on both skills, had

gains that were greater than the estimated SE_{diff}, and approximately 58–79% had gains that were greater than the ETS SE_{diff}. It was noted, however, that fewer students in Cohort 3 experienced gains greater than the ETS SE_{diff}. The size of the estimated SE_{diff} used in this study were approximately 60% the size of the ETS SE_{diff}. As Cohort 3 made gains similar to the other cohorts on the former but not the latter implies that gains made by individual students in Cohort 3 were, on average, smaller compared with students in other cohorts, and this we know from other analyses reported above.

This study is valuable in that an analysis tool that is not commonly used in applied linguistics was adopted. The quasi-experimental difference-in-differences technique was used to estimate the amount of gains made by each cohort, relative to each other, while accounting for their different starting points. The results showed that for Listening, Cohorts 1, 2, and 4 made gains, and that Cohort 3 made significantly fewer gains. For Reading, Cohort 4 made the greatest gains, significantly different from the other cohorts. A number of limitations were noted. In particular, the quality of the online TOEIC at Time 1 for Cohort 3 causes concern. However, without evidence from ETS or IIBC on the test quality, the tentative conclusion is that the smaller percentage of gains made by members of Cohort 3 was most likely due to their EPGM classroom learning experiences. That is, compared with other cohorts, Cohort 3 had two years of online classes, compared with zero (Cohort 1) or one year (Cohorts 2 and 4), and this negatively impacted learning for Cohort 3 as measured by one standardized language test, the TOEIC L&R.

Notes
[1] The TOEIC testing program includes Listening and Reading (TOEIC L&R) and Speaking and Writing (TOEIC S&W). In this paper, unless otherwise indicated, TOEIC refers to the TOEIC L&R.
[2] In most studies of COVID-19 affected learning loss, losses were typically greatest for children from disadvantaged or lower socioeconomic groups, but these differences, while important, are not a focus of this paper.
[3] In this study, initial-letter capitalised Listening or Reading refer to the TOEIC skill-based tests. Without an initial capital letter, listening or reading refer to the skills in general.
[4] Cid et al. (2017) reported on a study involving Japanese (n = 2045) and Korean (n = 1628) test-takers, most likely representing a greater range of abilities than are found at one Japanese university. As one reviewer pointed out, using the reliability coefficient from Cid et al. (α =. 94) to estimate SE_{diff} in the current study is problematic because it likely results in a deflated SE_{diff}. Unfortunately, as Koizumi et al. (2015) noted, it is not possible to estimate the reliability of the TOEIC L&R for one's institution. Using the ETS-reported α = .9 reliability coefficient instead results in SE_{diff} (ranging from 21.1 to 27.0) that are 129% larger than the SE_{diff} estimates shown in Appendix B. However, using this second reliability coefficient is also problematic because it is based on results from millions of test-takers worldwide. Also, ETS does not report SDs. Therefore, a workaround is to include the reliability reported by Cid et al. (2017) and the SDs from Shōzan University. Due to the problematic nature of estimating an SE_{diff} without knowing the reliability of the test for one institution, I also included a second calculation based on the much larger SE_{diff} of 35 (ETS, 2022). The inaccuracy of the SE_{diff} for the students at Shōzan University is an obvious limitation of this study.
[5] The same reviewer also noted that gains per hour are likely not uniform for students within different score bands. Indeed, Saegusa (1985) reported on this. For example, he estimated that an employee (not a student) would need approximately 450 hours to increase their combined TOEIC L&R score of 450 to 650, but an employee would need 500 hours to increase their score from 650 to 850.

Acknowledgements

References

Aucejo, E. M., French, J., Tymms, P., Ulgade Araya, M. P., & Zafar, B. (2020). The impact of COVID-19 on student experiences and expectations: Evidence from a survey. *Journal of Public Economics, 191,* 104271. https://doi.org/10.1016/j.jpubeco.2020.104271

Betthäuser, B.A., Bach-Mortensen, A.M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behaviour*, *7*, 375–385. https://doi.org/10.1038/s41562-022-01506-4

Cid, J., Wei, Y., Kim, S., & Hauck, C. (2017). *Statistical Analyses for the Updated TOEIC® Listening and Reading Test. Research Memorandum: ETS RM-17-05*. ETS. https://www.ets.org/Media/Research/pdf/RM-17-05.pdf

Engzell, P., Frey, A., & Verhagen, M. D. (2021). Learning loss due to school closures during the COVID-19 Pandemic. *PNAS: Proceedings of the National Academy of Sciences of the United States of America*, *118*(17), e2022376118. https://doi.org/10.1073/pnas.2022376118

ETS. (2022). *TOEIC Score User Guide, TOEIC Listening and Reading Test.* https://www.ets.org/content/dam/ets-org/pdfs/toeic/toeic-listening-reading-score-user-guide.pdf.

Fredriksson, A., & Magalhães de Oliveira, G. (2019). Impact evaluation using difference-in-differences. *RAUSP Management Journa*l, *54*(3), 519–532. https://doi.org/10.1108/RAUSP-05-2019-0112

Hallin, A. E., Danielsson, H., Nordström, T., & Fälth, L. (2022). No learning loss in Sweden during the pandemic: Evidence from primary school reading assessments. *International Journal of Educational Research*, *114*, 102011. https://doi.org/10.1016/j.ijer.2022.102011

The Crimson Editorial Board. (2022, November 10). Years after outbreak, Covid's Educational Toll Looms Large [Editorial]. *Harvard Crimson.* https://www.thecrimson.com/article/2022/11/10/editorial-covid-educational-loss/

Hevia, F. J., Vergara-Lope, S., Velásquez-Durán, A., & Calderón, D. (2022). Estimation of the fundamental learning loss and learning poverty related to COVID-19 pandemic in Mexico. *International Journal of Educational Development*, *88*, 102515. https://doi.org/10.1016/j.ijedudev.2021.102515

Horita, R., Nishio, A., & Yamamoto, M. (2021). The effect of remote learning on the mental health of first year university students in Japan. *Psychiatry Research*, *295*, 113561. https://doi.org/10.1016/j.psychres.2020.113561

IIBC. (2022). *2021 nendo jukenshasuu to heikin sukoa* [2021 Test-takers and average scores]. https://www.iibc-global.org/library/default/toeic/official_data/pdf/DAA.pdf

Ikeda, M., & Yamaguchi, S. (2021). Online learning during school closure due to COVID-19. *Japanese Economic Review, 72*, 471–507. https://doi.org/10.1007/s42973-021-00079-7

Iwabuchi, K., Hodama, K., Onishi, Y., Miyazaki, S., Nakae, S., & Suzuki, K. H. (2022). Covid-19 and education on the front lines in Japan: What caused learning disparities and how did the government and schools take initiative? In F. M. Reimers (Ed.), *Primary and Secondary Education During Covid-19* (pp. 125–151). Springer. https://doi.org/10.1007/978-3-030-81500-4_5

JASP Team (2023). *JASP* (Version 0.17.1) [Computer software].

Jeon, E. H. (2015). Multiple regression. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 131-158). Routledge.

Kang, B. (2021). How the COVID-19 pandemic is reshaping the education service. In J. Lee & S. H. Han (Eds.), *The Future of Service Post-COVID-19 Pandemic, Volume 1, The ICT and Evolution of Work* (pp. 15–36). https://doi.org/10.1007/978-981-33-4126-5_2

Kashima, R., & Yamamoto, Y. (2021). The effects of large-scale online classes on students' course evaluations: Evidence from a Japanese university under the COVID-19 pandemic. *Working Paper Series WP2021-03*, *Mori Arinori Institute for Higher Education and Global Mobility*. https://arinori.hit-u.ac.jp/wp2022/wp-content/uploads/2021/03/0103603a410aea7b4304337746c8786f.pdf

Koizumi, R., In'nami, Y., Azuma, J., Asano, K., Agawa, T., & Eberl, D. (2015). Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference. *Shiken*, *19*(1), 3–15. https://hosted.jalt.org/teval/node/22

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test score patterns across three COVID-19-impacted school years. *EdWorkingPaper: 22-521. Annenberg Institute at Brown University.* https://doi.org/10.26300/ga82-6v47

Lichand, G., Dória, C. A., Leal-Neto, O., & Cossi, J. (2022). The impacts of remote learning in secondary education: Evidence from Brazil during the pandemic. *Nature Human Behaviour*, *6*, 1079-1086. https://doi.org/10.1038/s41562-022-01350-6

Lytle, A., & Shin, J. L. (2022). Resilience and grit predict fewer academic and career concerns among first-year undergraduate students during COVID-19. *Social Psychology of Education, 26*, 227–240. https://doi.org/10.1007/s11218-022-09741-3

Motz, B. A., Quick, J. D., Wernert, J. A., & Miles, T. A. (2021). A pandemic of busywork: increased online coursework following the transition to remote instruction is associated with reduced academic achievement. O*nline Learning Journal*, *25*(1), 70–85. http://dx.doi.org/10.24059/olj.v25i1.2475

Nagata, M. (2022). Mantsu-man gata onrain eikawai gakushu ga tōikku sukoa ni ataeru kōka no teiryō teki bunseki

[Quantitative analysis of the effects of one-on-one online English conversation learning on TOEIC scores]. *Reitaku Journal of Economic Studies*, *29*(1), 12–23.

Obara, Y. (2022). The practice of online class under COVID-19 situation and its awareness survey: Through the questionnaires in English class. *Bulletin of Policy Management, Shobi University*, *38*, 17–41. http://id.nii.ac.jp/1506/00000742/

Orlov, G., McKee, D., Berry, J., Boyle, A., DiCiccio, T., Ransom, T., Rees-Jones, A., & Stoye, J. (2021). Learning during the COVID-19 pandemic: It is not who you teach, but how you teach. *Economics Letters*, *202*, 109812. https://doi.org/10.1016/j.econlet.2021.109812

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Saegusa, Y. (1985). Prediction of English proficiency progress. *Musashino English and American Literature*, *18*, 165–185.

Sato, Y., Yatsuya, H., Saijo, Y., Yoshioka, E., & Tabuchi, T. (2023). Psychological distress during the coronavirus disease 2019 pandemic and associated factors among undergraduate students in Japan. *Disaster Medicine and Public Health Preparedness, 17*, E294. https://doi.org/10.1017/dmp.2022.245

Schult, J., Mahler, N., Fauth, B., & Lindner, M. A. (2022). Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. S*chool Effectiveness and School Improvement*, *33*(4), 544–563. https://doi.org/10.1080/09243453.2022.2061014

Shindo, Y., Hitomi, E., & Iwano, M. (2021). The possibility of blended e-learning education in graduate program: The analysis follow up survey results on distant classes for prevention of the Covid-19 infection. *Yamaguchi Prefectural University Gakujitsu Jōhō*, 14, 57–75. http://ypir.lib.yamaguchi-u.ac.jp/yp/1664

Sukman, K., & Mhunkongdee, T. (2021). Thai EFL learners' voices on learning English online during the COVID- 19 pandemic. *International Journal of English Language Teaching*, *9*(2), 1–9. http://dx.doi.org/10.2139/ssrn.3824069

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Allyn & Bacon/Pearson Education.

Tan, C. (2021). The impact of COVID-19 on student motivation, community of inquiry and learning performance. *Asian Education and Development Studies*, *10*(2), 308–321. https://doi.org/10.1108/AEDS-05-2020-0084

World Bank. (2023). Difference-in-differences. https://dimewiki.worldbank.org/Difference-in-differences

Appendix A

Calculations for Evidence of RTM

Expected posttest score $= M_y + r_{xy}(SD_y/SD_x)(X - M_x)$

Cohort 1
Listening $= 310.7 + (.55)(70.3/53.8)(X - 235.7)$
$= 310.7 + (.55)(1.31)(X - 235.7)$
$= 310.7 + (0.721)(X - 235.7)$

Reading $= 245.4 + (.62)(68.3/52.0)(X - 177.4)$
$= 245.4 + (.62)(1.31)(X - 177.4)$
$= 245.4 + (0.814)(X - 177.4)$

Cohort 2
Listening $= 304.4 + (.53)(60.3/50.2)(X - 231.5)$
$= 304.4 + (.53)(1.20)(X - 231.5)$
$= 304.4 + (0.637)(X - 231.5)$

Reading $= 242.8 + (.55)(64.3/47.2)(X - 185)$
$= 242.8 + (.55)(1.36)(X - 185)$
$= 242.8 + (0.749)(X - 185)$

Cohort 3
Listening $= 334 + (.36)(62.2/53.3)(X - 279.5)$
$= 334 + (.36)(1.17)(X - 0.32)$
$= 334 + (0.420)(X - 0.32)$

Reading $= 285.8 + (.30)(66.5/48.3)(X - 227.7)$
$= 285.8 + (.30)(1.38)(X - 227.7)$
$= 285.8 + (0.413)(X - 227.7)$

Cohort 4
Listening $= 342.5 + (.61)(59.9/52.0)(X - 267.7)$
$= 342.5 + (.61)(1.15)(X - 267.7)$
$= 342.5 + (0.702)(X - 267.7)$

Reading $= 284.4 + (.59)(74.6/60.3)(X - 197.7)$
$= 284.4 + (.59)(1.24)(X - 197.7)$
$= 284.4 + (0.730)(X - 197.7)$

Appendix B

Calculations for Evidence of Growth beyond SE$_{diff}$

Cohort 1
Listening = (53.754) * ($\sqrt{2 - (.94) - (.94)}$]
               = (53.754) * (.346)
               = 18.599

Reading = (51.976) * ($\sqrt{2 - (.94) - (.94)}$]
               = (51.976) * (.346)
                = 17.984

Cohort 2
Listening = (50.231) * ($\sqrt{2 - (.94) - (.94)}$]
               = (50.231) * (.346)
               = 17.380

Reading = (47.232) * ($\sqrt{2 - (.94) - (.94)}$]
               = (47.232) * (.346)
                = 16.342

Cohort 3
Listening = (53.514) * ($\sqrt{2 - (.94) - (.94)}$]
               = (53.514) * (.346)
               = 18.516

Reading = (48.286) * ($\sqrt{2 - (.94) - (.94)}$]
               = (48.286) * (.346)
                = 16.707

Cohort 4
Listening = (51.992) * ($\sqrt{2 - (.94) - (.94)}$]
               = (51.992) * (.346)
                = 17.989

Reading = (60.335) * ($\sqrt{2 - (.94) - (.94)}$]
               = (60.335) * (.346)
                = 20.876

Appendix C

Dummy Codes Used in the Difference-in-Differences Analyses

(Cohort:) (D1) (D2) (D3) (Post, i.e., not post treatment)
1_x: 0000
2_x: 1000
3_x: 0100
4_x: 0010
(Cohort:) (D1) (D2) (D3) (Post, i.e., not post treatment)
1_y: 0001
2_y: 1001
3_y: 0101
4_y: 0011