SHIKEN

A Journal of Language Testing and Evaluation in Japan

Volume 27 • Number 1 • June 2023

https://doi.org/10.37546/JALTSIG.TEVAL27.1

Contents

Testing and Evaluation SIG

ISSN 1881-5537

Shiken: A Journal of Language Testing and Evaluation in Japan

Volume 27 No. 1
June 2023

https://doi.org/10.37546/JALTSIG.TEVAL27.1

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

https://doi.org/10.37546/JALTSIG.TEVAL27.1-1

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 Shozan 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	Cohort 2	Cohort 3	Cohort 4
	(n = 211)	(n = 217)	(n = 216)	(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:* $1^2 = .005$, $1^2_p = .002$*).*

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 x^2 test for independence was carried out to assess the distribution of outliers across cohorts. The test was not significant: x^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.

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

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, x^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).

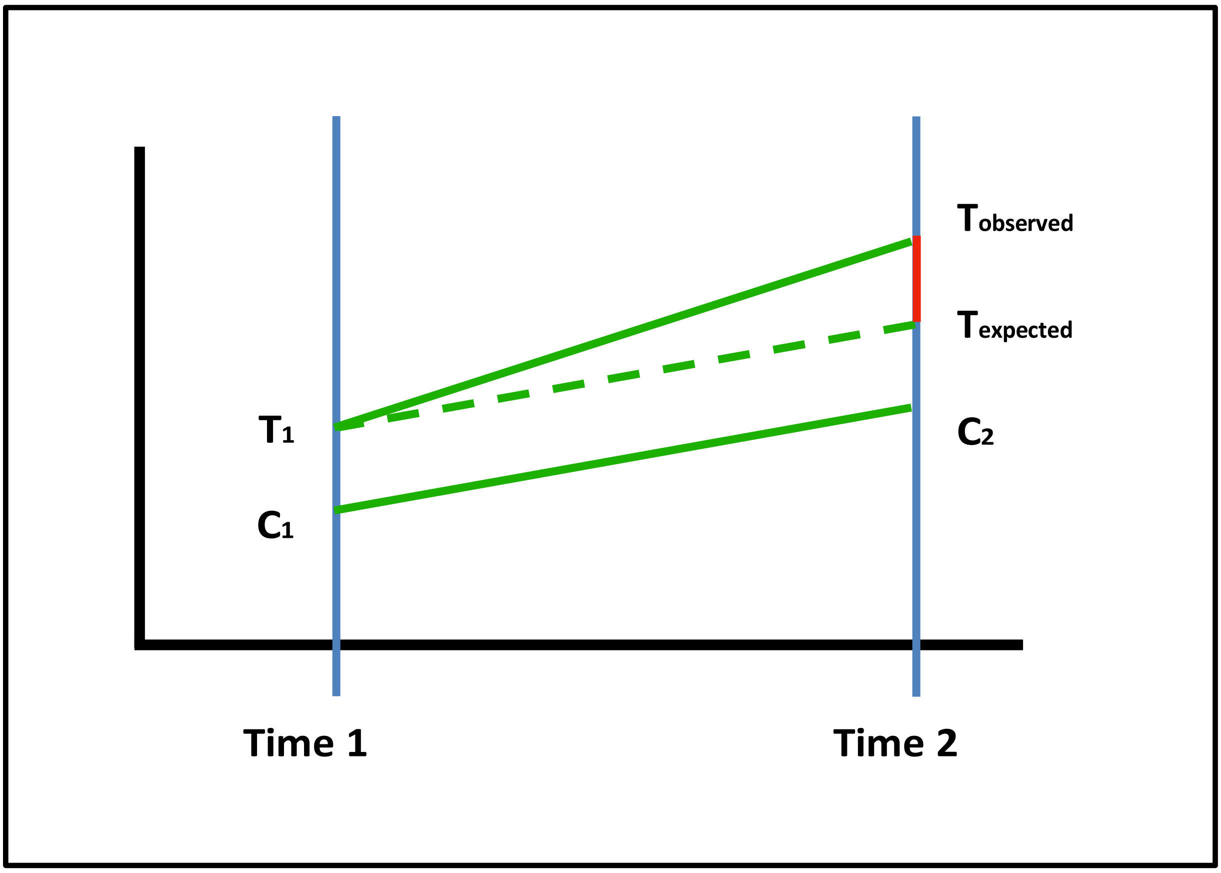
$$SE_{diff} = (\text{Time 1 SD}) * (\surd[2 - (\text{Reliability at Time 1}) - (\text{Reliability at Time 2})]) \qquad (2)$$

The TOEIC score user guide from Educational Testing Service (ETS) reported a reliability coefficient of a =.90 for both Listening and Reading (ETS, 2022, p. 19); however, ETS researchers reported a reliability coefficient of a =.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, x^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 $T_{expected}$, represented by the small red line.

Figure 1
Representation of a basic difference-in-differences model



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).

$$8_{DD} = Y_{Treat,2} - [Y_{Treat,1} + (Y_{Control,2} - Y_{Control,1})] \tag{3}$$

where 8_{DD} represents the difference in differences; $Y_{Treat,1}$ and $Y_{Treat,2}$ represent the average of the treatment group at Times 1 and 2, and $Y_{Control,1}$ and $Y_{Control,2}$ 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 ($Y_{Treat,1}$)	268 ($Y_{Control,1}$)	281 ($Y_{Treat,1}$)
Post (Time 2)	330 ($Y_{Treat,2}$)	342 ($Y_{Control,2}$)	355 [$Y_{Treat,1} + (Y_{Control,2} - Y_{Control,1})$]

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).

$$Y_i = b_o + 8_T D_{TREATi} + 8_T D_{POSTi} + 8_{DD}(D_{TREATi})(D_{POSTi}) \tag{4}$$

where Y_i represents the predicted outcome variable. The ordinary least squares estimate of 8_{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

	Unstandardized			95% CI*	
Model (H)	Coefficients	SE	p*	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 -1.24 (-1.42, -1.06)	t(210) = -18.13, p < .001 -1.25 (-1.43, -1.07)
2 (n = 217)	t(216) = -19.83, p < .001 -1.35 (-1.53, -1.16)	t(216) = -15.52, p < .001 -1.05 (-1.22, -0.89)
3 (n = 216)	t(215) = -12.17, p < .001 -0.83 (-0.98, -0.67)	t(215) = -12.28, p < .001 -0.84 (-0.99, -0.68)
4 (n = 210)	t(209) = -21.70, p < .001 -1.50 (-1.69, -1.30)	t(209) = -20.04, p < .001 -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, a =.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. (a =. 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 a = .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 Shozan 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 Shozan 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

I am grateful to the two expert readers who commented upon earlier drafts of this article. Any errors or omissions remain my own.

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 toikku sukoa ni ataeru koka no teiryo 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 Joho*, 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$_{\text{diff}}$

Cohort 1
Listening = (53.754) * (✓[2 - (.94) - (.94)]
= (53.754) * (.346)
= 18.599

Reading = (51.976) * (✓[2 - (.94) - (.94)]
= (51.976) * (.346)
= 17.984

Cohort 2
Listening = (50.231) * (✓[2 - (.94) - (.94)]
= (50.231) * (.346)
= 17.380

Reading = (47.232) * (✓[2 - (.94) - (.94)]
= (47.232) * (.346)
= 16.342

Cohort 3
Listening = (53.514) * (✓[2 - (.94) - (.94)]
= (53.514) * (.346)
= 18.516

Reading = (48.286) * (✓[2 - (.94) - (.94)]
= (48.286) * (.346)
= 16.707

Cohort 4
Listening = (51.992) * (✓[2 - (.94) - (.94)]
= (51.992) * (.346)
= 17.989

Reading = (60.335) * (✓[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

jMetrik Guide

Aaron Olaf Batty
abatty@keio.jp
Keio University

Introduction

This guide will walk you through getting your data into jMetrik and running an item analysis. jMetrik does many more things than just classical test theory (CTT) item analysis, though. You may want to explore more as you learn. Rasch/IRT analysis, DIF analysis, multiple models, etc., are all available! But for now, let's just look at some items with CTT analyses.

Keywords: software, how-to, item analysis

Necessary software

You need the following software:

1. A text editor

—*Not* a word processor—a program that is just for working with text files. I recommend Notepad++ for Windows or BBEdit for Mac (the free version is fine). If you already have a text editor you like (e.g., Sublime, Atom); you don't need me to tell you what a text editor is.

2. jMetrik
—Of course!
3. Microsoft Excel
—Or an Excel-alike such as LibreOffice Calc. Please don't try to use Apple Numbers for *anything*.

Overall process

The process has the following steps:

1. Installing jMetrik
2. Doing the analyses in jMetrik
3. Exporting your analyses to text and/or CSV
4. *Extra:* Getting your data into jMetrik

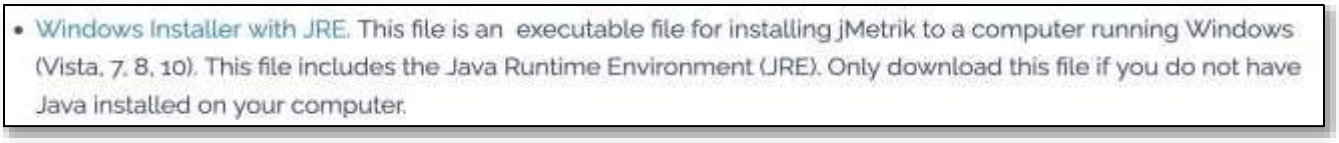
Let's begin!

1. Installing jMetrik

The installation files can be downloaded from the jMetrik download page here. Please refer to the platform-specific instructions below.

Windows

Unless you know that you have Java installed (you probably don't), please download the "Windows Installer with JRE" installer from the jMetrik download page as shown below.



- Windows Installer with JRE. This file is an executable file for installing jMetrik to a computer running Windows (Vista, 7, 8, 10). This file includes the Java Runtime Environment (JRE). Only download this file if you do not have Java installed on your computer.

After downloading, simply install the software. By default, the software will start after installation, but there's nothing to do with it yet, so please just close it.

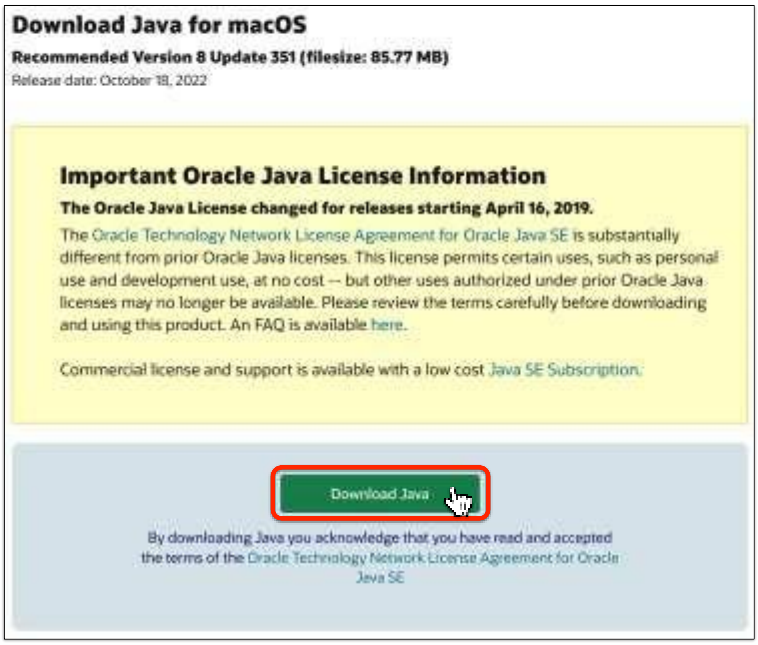
Macintosh

macOS 12 (Monterey) and above does not like the Java installer used by jMetrik, and claims you need to log in as root (the highest-level user of your computer, which is disabled by default) to install. Doing this is possible[1], but should be avoided by novice users. A workaround for more novice users is to install Java separately, then run jMetrik directly. Novice instructions follow, and a video of the process may be viewed here.

Step 1. Install Java

You will need to install Java manually and then run jMetrik as a file. Don't be alarmed. Java is just software that other software can run on. That's why the program can run on Windows, Mac, Linux… *whatever*, and still look the same. It's actually running in a Java virtual machine. There's nothing you need to do but install it.

1. Download Java from here.



Download Java for macOS

Recommended Version 8 Update 351 (filesize: 85.77 MB)

Release date: October 18, 2022

Important Oracle Java License Information

The Oracle Java License changed for releases starting April 16, 2019.

The Oracle Technology Network License Agreement for Oracle Java SE is substantially different from prior Oracle Java licenses. This license permits certain uses, such as personal use and development use, at no cost -- but other uses authorized under prior Oracle Java licenses may no longer be available. Please review the terms carefully before downloading and using this product. An FAQ is available here.

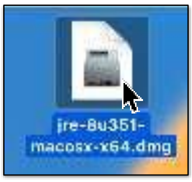
Commercial license and support is available with a low cost Java SE Subscription.

Download Java

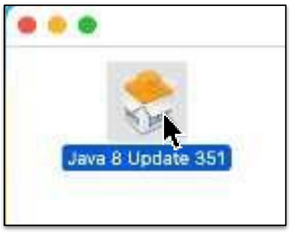
By downloading Java you acknowledge that you have read and accepted the terms of the Oracle Technology Network License Agreement for Oracle Java SE

2. Double-click the Java installer disk image.

NOTE: The name will be slightly different as new versions come out.

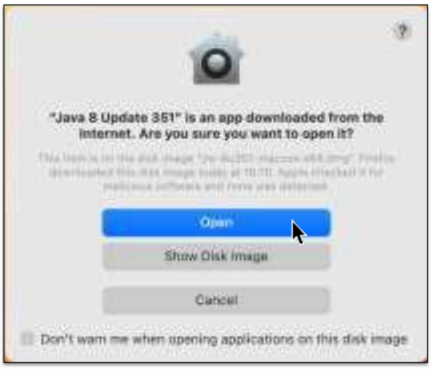


3. A window will open with the installer icon. Double-click the installer.



[1] Refer to the Apple support page on enabling/disabling the root user (https://support.apple.com/en-us/HT204012). Once enabled, log in as root and install the jMetrik package that includes the Java Runtime Environment (JRE). Then, log out as root and log in with your normal administrator account. Finally, *disable the root user* again. jMetrik will run normally and not request root access again.

4. Apple will then ask you if you really want to run the software you have just double-clicked. Tell them that yes, in fact, you *do* want to do the thing you have just done. Click “Open."

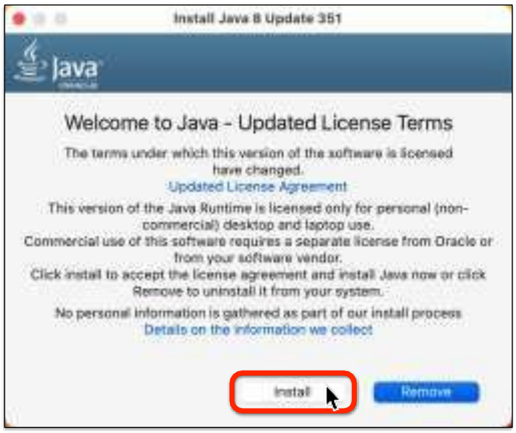


"Java 8 Update 351" is an app downloaded from the Internet. Are you sure you want to open it?

This item is on the disk image "jre-8u351-macosx-x64.dmg". [illegible] downloaded this disk image today at [illegible]. Apple checked it for malicious software and none was detected.

Open
Show Disk Image
Cancel
□ Don't warn me when opening applications on this disk image

5. The installer opens. Click “Install," even though “Remove," bizarrely, is the default.



Install Java 8 Update 351

Welcome to Java - Updated License Terms

The terms under which this version of the software is licensed have changed.

Updated License Agreement

This version of the Java Runtime is licensed only for personal (non-commercial) desktop and laptop use.

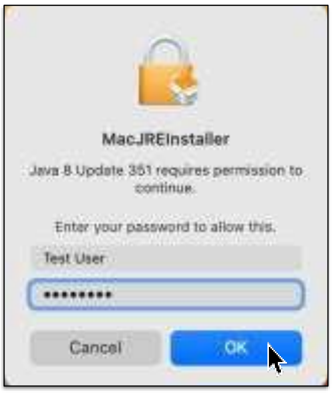
Commercial use of this software requires a separate license from Oracle or from your software vendor.

Click install to accept the license agreement and install Java now or click Remove to uninstall it from your system.

No personal information is gathered as part of our install process

Details on the information we collect

6. You will need to enter your Mac password to continue.



MacJREInstaller

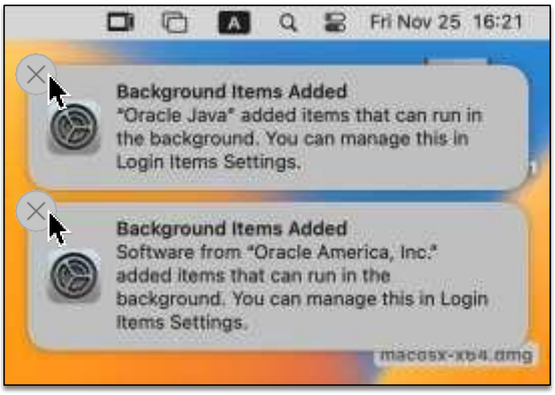
Java 8 Update 351 requires permission to continue.

Enter your password to allow this.

Test User

•••••••••

7. Apple will then helpfully tell you that you have installed things that can run. This is very helpful information indeed. *Click the Xes on these and never think about them again.*

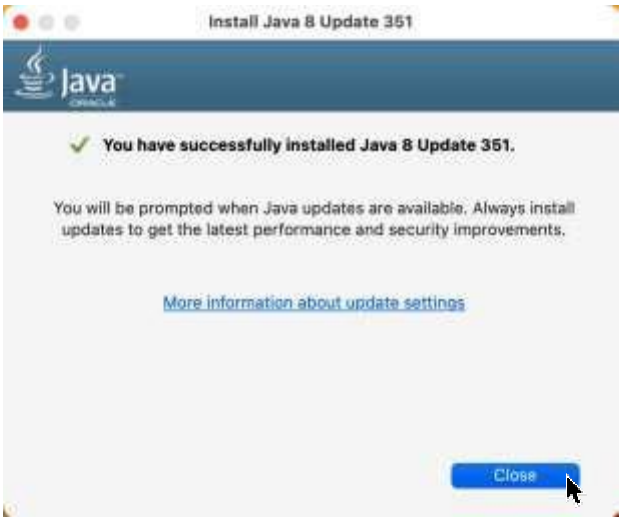


Fri Nov 25 16:21

Background Items Added
"Oracle Java" added items that can run in the background. You can manage this in Login Items Settings.

Background Items Added
Software from "Oracle America, Inc." added items that can run in the background. You can manage this in Login Items Settings.

8. Finally, the installation will finish. Click “Close."



Install Java 8 Update 351

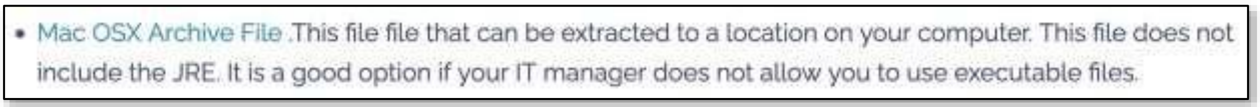
✓ You have successfully installed Java 8 Update 351.

You will be prompted when Java updates are available. Always install updates to get the latest performance and security improvements.

More information about update settings

Step 2. Download and decompress jMetrik

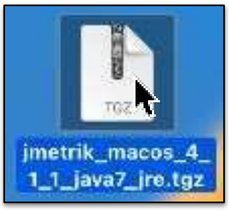
1. Download the **Mac OSX Archive File** installer from the jMetric download page as shown below.



- Mac OSX Archive File .This file file that can be extracted to a location on your computer. This file does not include the JRE. It is a good option if your IT manager does not allow you to use executable files.

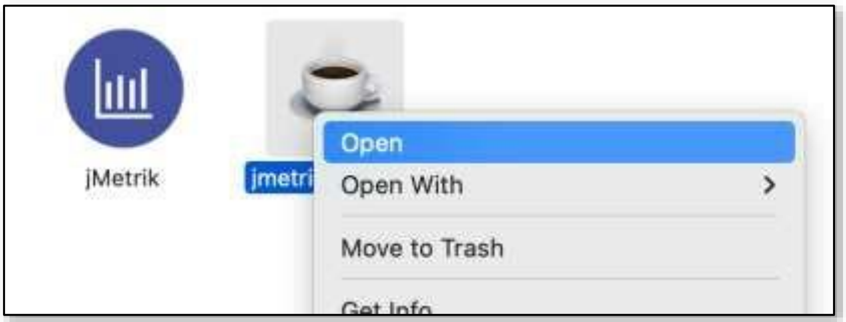
2. Move the downloaded file to someplace convenient. (The Downloads folder is not convenient.)

3. Double-click the archive file.



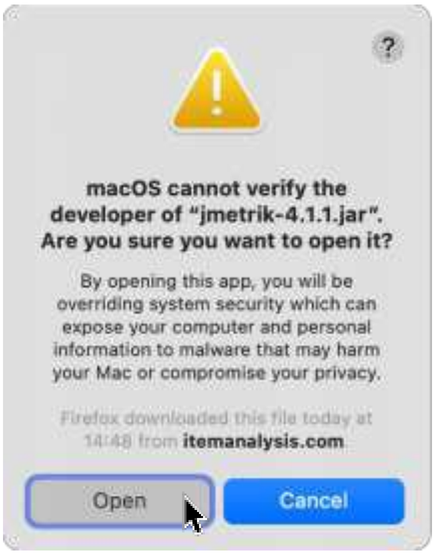
4. A folder called “jmetrik" will appear. Open it.

5. ***Right-click*** the “jmetrik-4.X.X.jar" file (*NOT* the “jMetrik" file) and select “Open":



—If you double-click it, you will get a message refusing to open it because it is from an “unidentified developer." We are bypassing that.

6. You may get a message asking if you really want to open it. Click “Open."



macOS cannot verify the developer of “jmetrik-4.1.1.jar”. Are you sure you want to open it?

By opening this app, you will be overriding system security which can expose your computer and personal information to malware that may harm your Mac or compromise your privacy.

Firefox downloaded this file today at 14:48 from itemanalysis.com

7. The software will (finally) open.

2. Getting your data into jMetrik

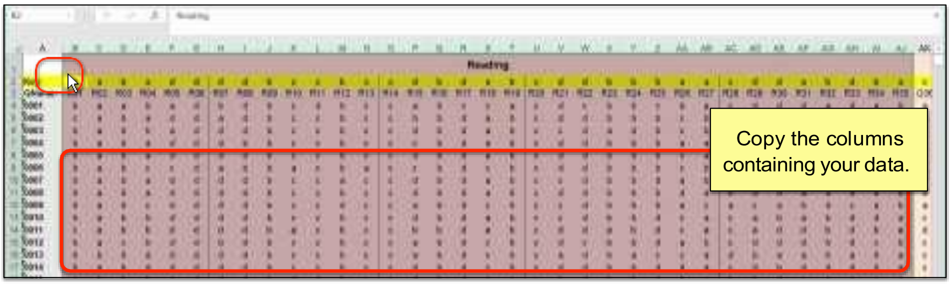
jMetrik reads plaintext files. There are myriad ways to make them, but I'm going to show you the way that I prefer. (*NOTE:* I am showing the process on Windows, but the process on the Mac is exactly the same.)

The overall process is as follows:

1. Make a data file for use in jMetrik
2. Make a score file for use in jMetrik

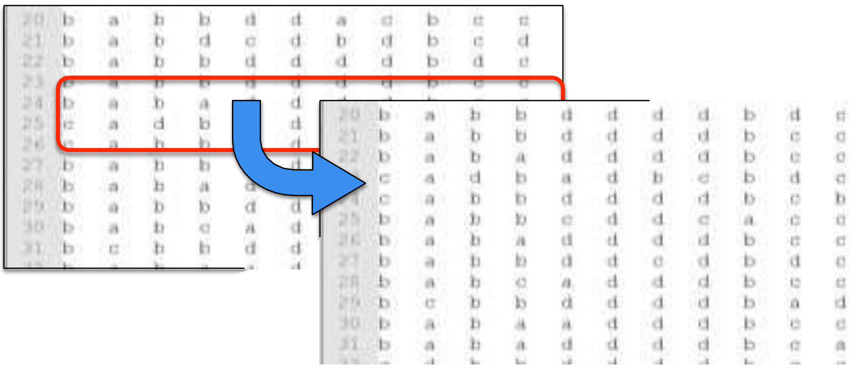
Step 1. Making the data file for jMetrik

1. Open your data file in Excel.
2. Select the columns (you can click the row names and drag) with the data for items you'd like to analyze and copy them (ctrl-C on Windows, ⌘-C on Mac, or click the “Copy" button on the Home tab of the ribbon, or right-click and select “Copy"—however you normally do it!). ***Only copy the columns with data in them.*** You do not want any columns with information like student numbers or names. Just the answers.



Copy the columns containing your data.

3. Make a new file in your text editor (by default, most editors make a new file when you open them).
4. Paste into the new text file (ctrl-V on Windows, ⌘-V on Mac, right-click…).
5. The top line needs to be the “names" of the questions. Simply highlight and delete anything above that.



6. Save the file with an easy-to-identify name (e.g., ReadingABCD.txt).

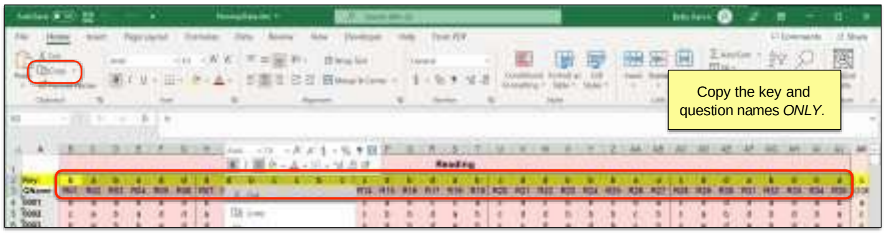
Step 2. Making the score file for jMetrik

jMetrik needs to be given information on how to score the data. This can be done in jMetrik directly, but it's time-consuming. I find it much faster and easier to simply give it a file with the code it needs. I have prepared an Excel file that creates these files:

https://hosted.jalt.org/teval/sites/jalt.org.teval/files/jMetrik%20Score%20File%20Maker.xls

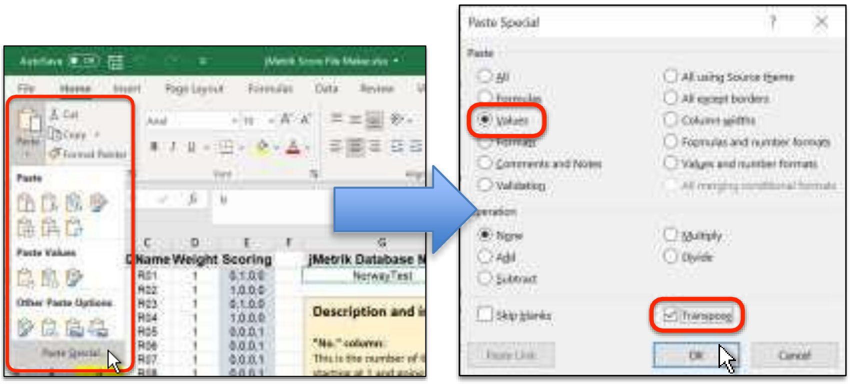
—The file includes text boxes that tell you how to use it. It can be adapted to be used with any data you have with minimal changes.

1. Open the “jMetrik Score File Maker.xls" file and make sure it is open to the “jMetrik Key" tab.
2. Return to your dataset in Excel and copy the key and names of your items. **Take care to *only* copy the key and the question names.**



Copy the key and question names *ONLY*.

3. Paste your item names and key into the jMetrik Score File Maker. Be sure to paste values only. You may neet to transpose (Click the first row in the “Key" column and then click the drop-down arrow next to “Copy" in the Home tab of the Ribbon. Select “Paste Special" and in the window that appears, select “Values" and—if necessary—“Transpose," then click OK.)



Paste Special

Paste

○ All ○ All using Source theme
○ Formulas ○ All except borders
● Values ○ Column widths
○ Formats ○ Formulas and number formats
○ Comments and Notes ○ Values and number formats
○ Validation ○ All merging conditional formats

Operation

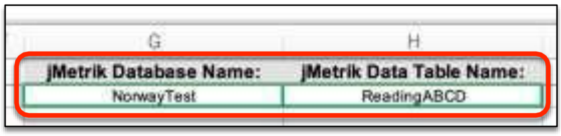
● None ○ Multiply
○ Add ○ Divide
○ Subtract

□ Skip blanks ☑ Transpose

Paste Link OK Cancel

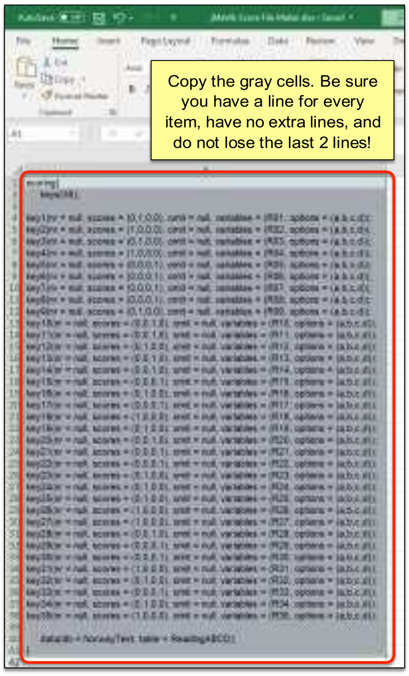
4. *(May be necessary)* If you need to add or remove lines, do so. Be sure that the first column of sequential numbers goes all the way down, and you copy the “Scoring" code down if you have more than 30 items.

5. Enter the jMetrik database and table names you plan to use. ***NOTE:* These must match what you make in jMetrik later so that jMetrik knows how to use the score file!**



G	H
jMetrik Database Name:	jMetrik Data Table Name:
NorwayTest	ReadingABCD

6. Click on the “jMetrik Score File" tab.
7. *(May be necessary)* If you need to add or remove lines, do so. Ensure that you do not lose the last two lines of the gray cells, as they tell jMetrik what database and table the scoring data pertains to.
8. Select and copy the gray cells.



Copy the gray cells. Be sure you have a line for every item, have no extra lines, and do not lose the last 2 lines!

9. Make a new file in your text editor.
10. Paste into the new text file.
11. Save the file with an easy-to-remember name (e.g., “Reading Score File.txt")

Now you can do the analysis in jMetrik!

3. Doing the analysis in jMetrik

jMetrik uses a database with tables of information in it to do the analyses. For this reason, there are a few steps we need to take before we can start working on a project. The basic process is as follows:

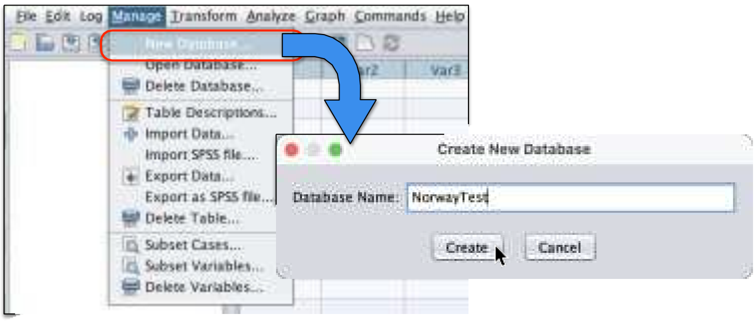
1. Make and open a new database
2. Import data
3. Import the score file and run it
4. Run the analyses

You can use the same database as much as you like, but I recommend making a new one for any new test. For example, it would make sense to use the same database for all of the sections of a particular test, but if you started working on a different test, it would make sense to make a new one and use that instead.

You will need to open a database every time you use the program. Don't be alarmed if you start the program and you don't see anything listed. You just need to tell it which database you want to use!

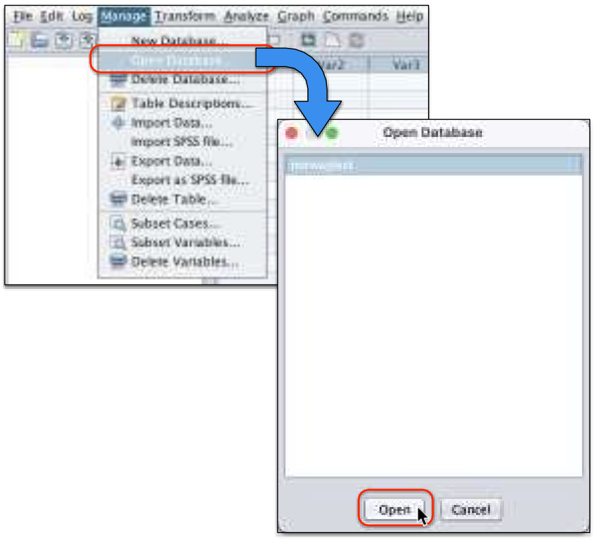
Making and opening a new database

1. Start jMetrik (if you are using a non-Intel Macintosh, it might take a little while to start the first time, as Rosetta 2 needs to make a translated binary to run on the Apple ARM CPUs—if you don't know what that means, don't worry; just wait. It will run normally after the first time.).
2. Go to the Manage menu and select “New Database…"
3. In the window that comes up, give your database a name (e.g., “ReadingTest") and click “Create." ***NOTE:* This name must match the database name you set in the "jMetrik Score File Maker" Excel file!**



Database Name: NorwayTest

4. Return to the Manage menu and select “Open Database…"
5. Select your database from the list and click “Open."



Importing data

(See the next page for a diagram.)

1. Go to the Manage menu and select “Import Data…"
2. Give a name to your table (e.g., “ReadingABCD").
3. NOTE: This name must match the table name you set in the “jMetrik Score File Maker" Excel file, if you're making your own files!
4. Click “Browse" and locate the score file you made.
5. Select “Tab" for the Delimiter and “In first row" for the Variable Names.
6. Click OK.
7. On the “Import" window, click “Import."

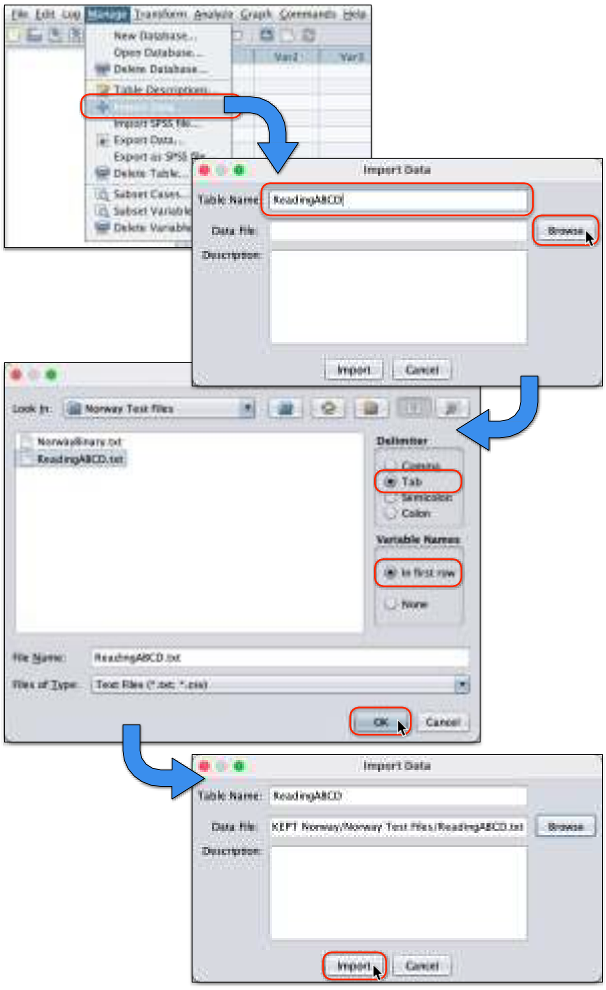


Table Name: ReadingABCD

Data File:

Description:

Delimiter: ● Tab

Variable Names: ● In first row

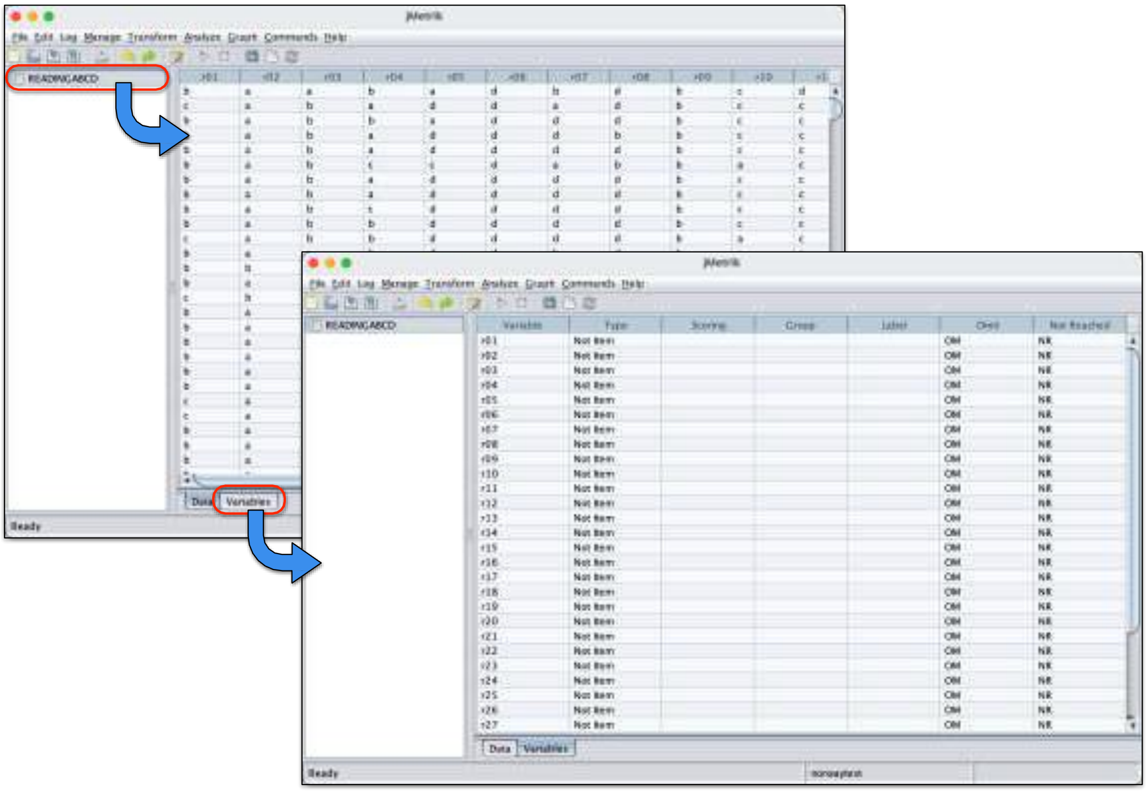
File Name: ReadingABCD.txt

Table Name: ReadingABCD

Data File: KEPT Norway/Norway Test Files/ReadingABCD.txt

Description:

8. You will see a new entry on the left side of the main jMetrik window. Click it to see your imported data.
9. Click the “Variables" tab at the bottom to see the scoring… Uh-oh… It doesn't know these are items. We have to give it the score file so it knows (next section).

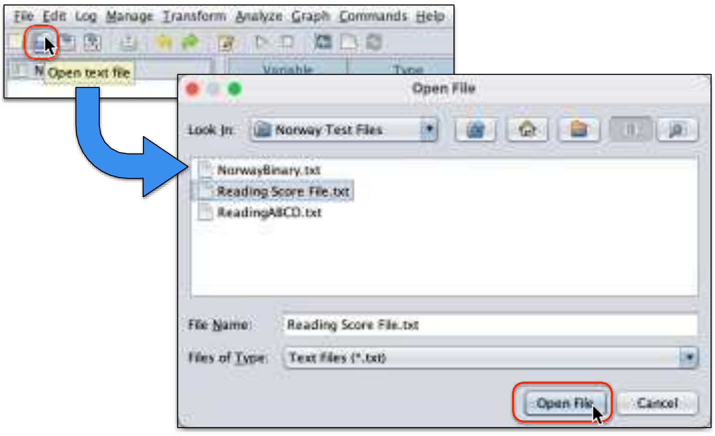

jMetrik
File Edit Log Manage Transform Analyze Graph Commands Help
READINGABCD
Data
Variables
Ready
jMetrik
File Edit Log Manage Transform Analyze Graph Commands Help
READINGABCD
Data
Variables
Ready
norwaytest

Variable	Type	Scoring	Group	Label	Omit	Not Reached
r01	Not Item				OM	NR
r02	Not Item				OM	NR
r03	Not Item				OM	NR
r04	Not Item				OM	NR
r05	Not Item				OM	NR
r06	Not Item				OM	NR
r07	Not Item				OM	NR
r08	Not Item				OM	NR
r09	Not Item				OM	NR
r10	Not Item				OM	NR
r11	Not Item				OM	NR
r12	Not Item				OM	NR
r13	Not Item				OM	NR
r14	Not Item				OM	NR
r15	Not Item				OM	NR
r16	Not Item				OM	NR
r17	Not Item				OM	NR
r18	Not Item				OM	NR
r19	Not Item				OM	NR
r20	Not Item				OM	NR
r21	Not Item				OM	NR
r22	Not Item				OM	NR
r23	Not Item				OM	NR
r24	Not Item				OM	NR
r25	Not Item				OM	NR
r26	Not Item				OM	NR
r27	Not Item				OM	NR

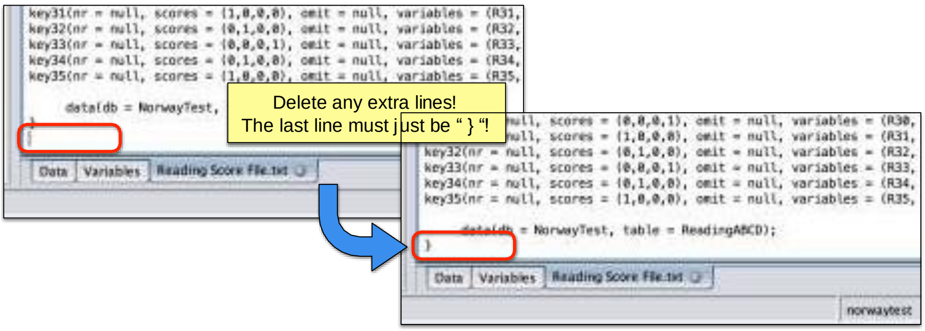
Importing and running the score file

This process tells jMetrik that the variables are items, and how to score them.

1. Click the “Open text file" button in the jMetrik toolbar (second button from the left).
2. Locate your score file and click “Open File." Your file will open in a new tab in jMetrik


File Edit Log Manage Transform Analyze Graph Commands Help
Open text file
Variable
Type
Open File
Look In: Norway Test Files
NorwayBinary.txt
Reading Score File.txt
ReadingABCD.txt
File Name: Reading Score File.txt
Files of Type: Text Files (*.txt)
Open File
Cancel

3. Your text editor and/or jMetrik might have snuck an extra line or two at the bottom of the file. You need to delete those for it to work. The last line needs to be just a single curly bracket (}).

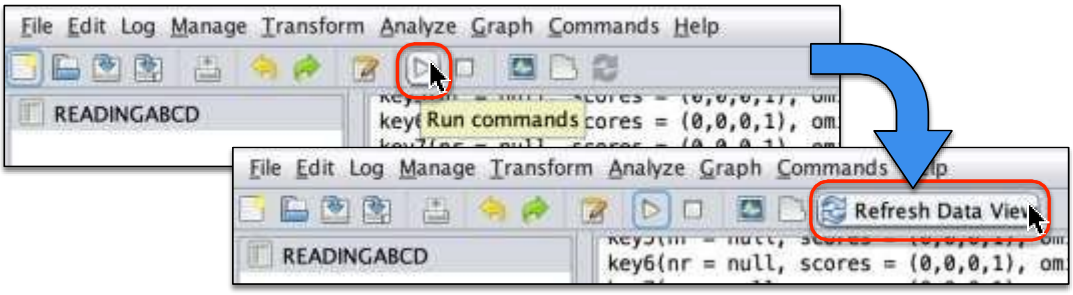


```
key31(nr = null, scores = (1,0,0,0), omit = null, variables = (R31,
key32(nr = null, scores = (0,1,0,0), omit = null, variables = (R32,
key33(nr = null, scores = (0,0,0,1), omit = null, variables = (R33,
key34(nr = null, scores = (0,1,0,0), omit = null, variables = (R34,
key35(nr = null, scores = (1,0,0,0), omit = null, variables = (R35,

    data(db = NorwayTest,
}
```

Delete any extra lines!
The last line must just be “ } “!

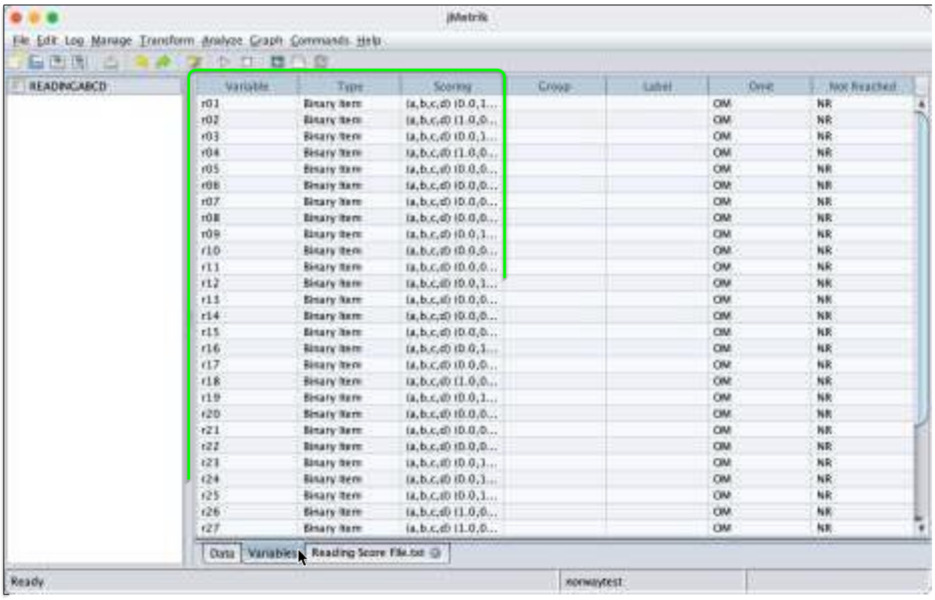
```
        null, scores = (0,0,0,1), omit = null, variables = (R30,
        null, scores = (1,0,0,0), omit = null, variables = (R31,
key32(nr = null, scores = (0,1,0,0), omit = null, variables = (R32,
key33(nr = null, scores = (0,0,0,1), omit = null, variables = (R33,
key34(nr = null, scores = (0,1,0,0), omit = null, variables = (R34,
key35(nr = null, scores = (1,0,0,0), omit = null, variables = (R35,

    data(db = NorwayTest, table = ReadingABCD);
}
```

4. Click the “Run commands" button in the toolbar (it looks like a “play" button—a triangle pointing right). This runs the code in the score file.
5. The “Refresh Data View" button becomes clickable. Click it!



```
key6(nr = null, scores = (0,0,0,1), om
```

6. Return to the “Variables" tab and check that your items have all been scored:



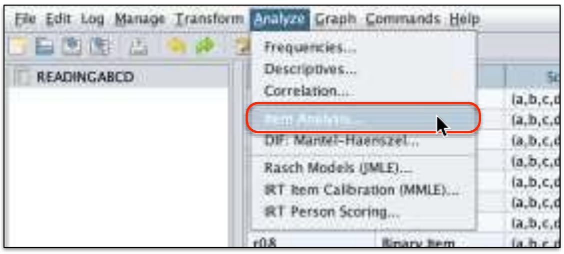
Variable	Type	Scoring	Group	Label	Omit	Not Reached
r01	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r02	Binary Item	{a,b,c,d} {1.0,0...			OM	NR
r03	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r04	Binary Item	{a,b,c,d} {1.0,0...			OM	NR
r05	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r06	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r07	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r08	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r09	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r10	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r11	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r12	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r13	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r14	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r15	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r16	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r17	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r18	Binary Item	{a,b,c,d} {1.0,0...			OM	NR
r19	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r20	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r21	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r22	Binary Item	{a,b,c,d} {0.0,0...			OM	NR
r23	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r24	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r25	Binary Item	{a,b,c,d} {0.0,1...			OM	NR
r26	Binary Item	{a,b,c,d} {1.0,0...			OM	NR
r27	Binary Item	{a,b,c,d} {1.0,0...			OM	NR

Now it's time to run some analyses!

Running the analyses

We're finally ready for the fun part.

1. Go to the Analyze menu and select “Item Analysis…" A window with settings appears.



2. Click the “ > > “ button to move all the items to the right side. This tells jMetrik that we want to look at all of them (You can also do individual items, etc., if you like.)
3. Click the “Item deleted reliability" checkbox.
4. *(Optional, but recommended.)* Click the “Save" button to save the results as a data table. This can be really convenient for exporting to Excel if you need to do that. Give the table an easy name (I always call item analyses “IFILE") and click OK.
5. Click the Run button (see the image on the next page).

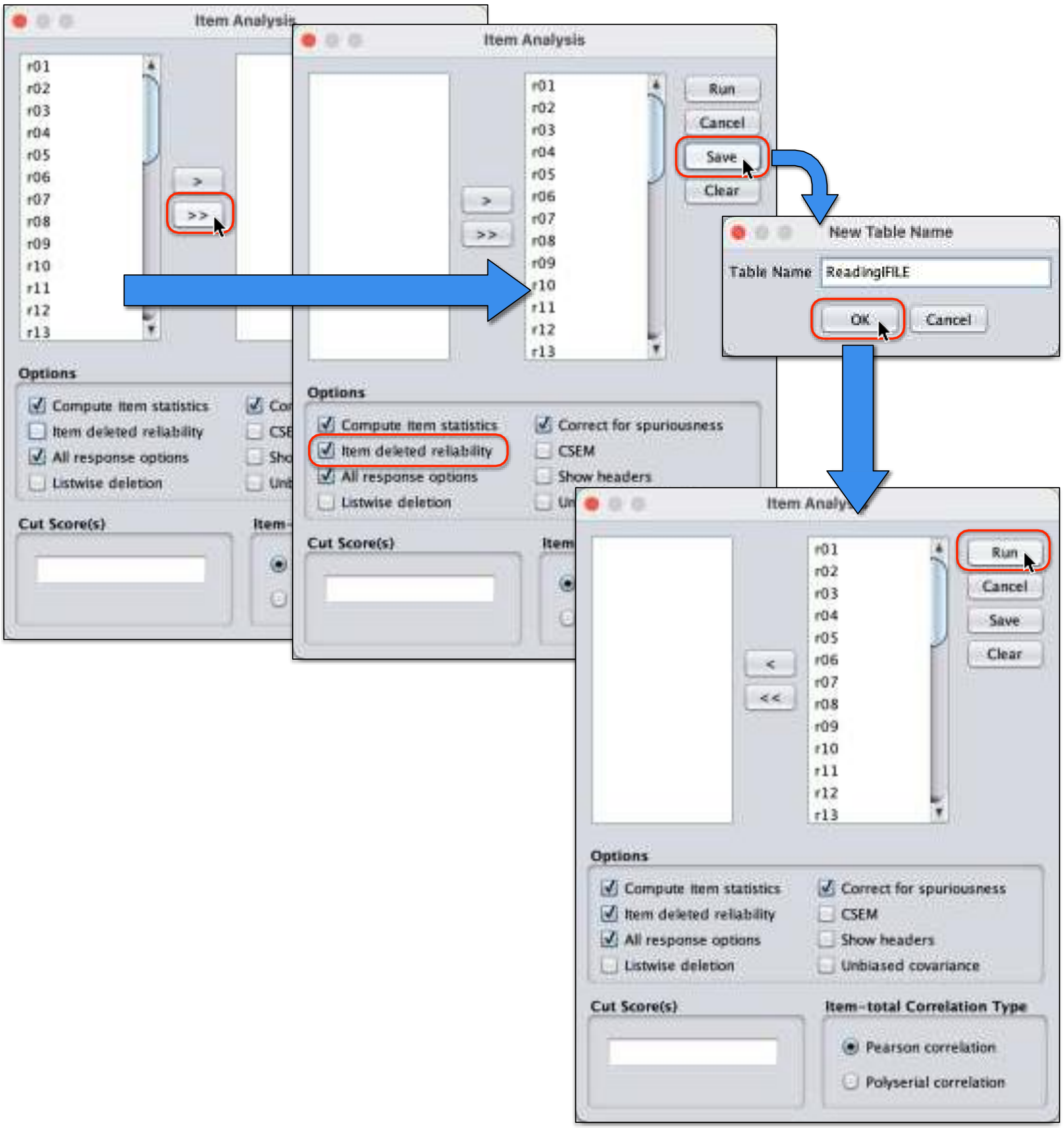


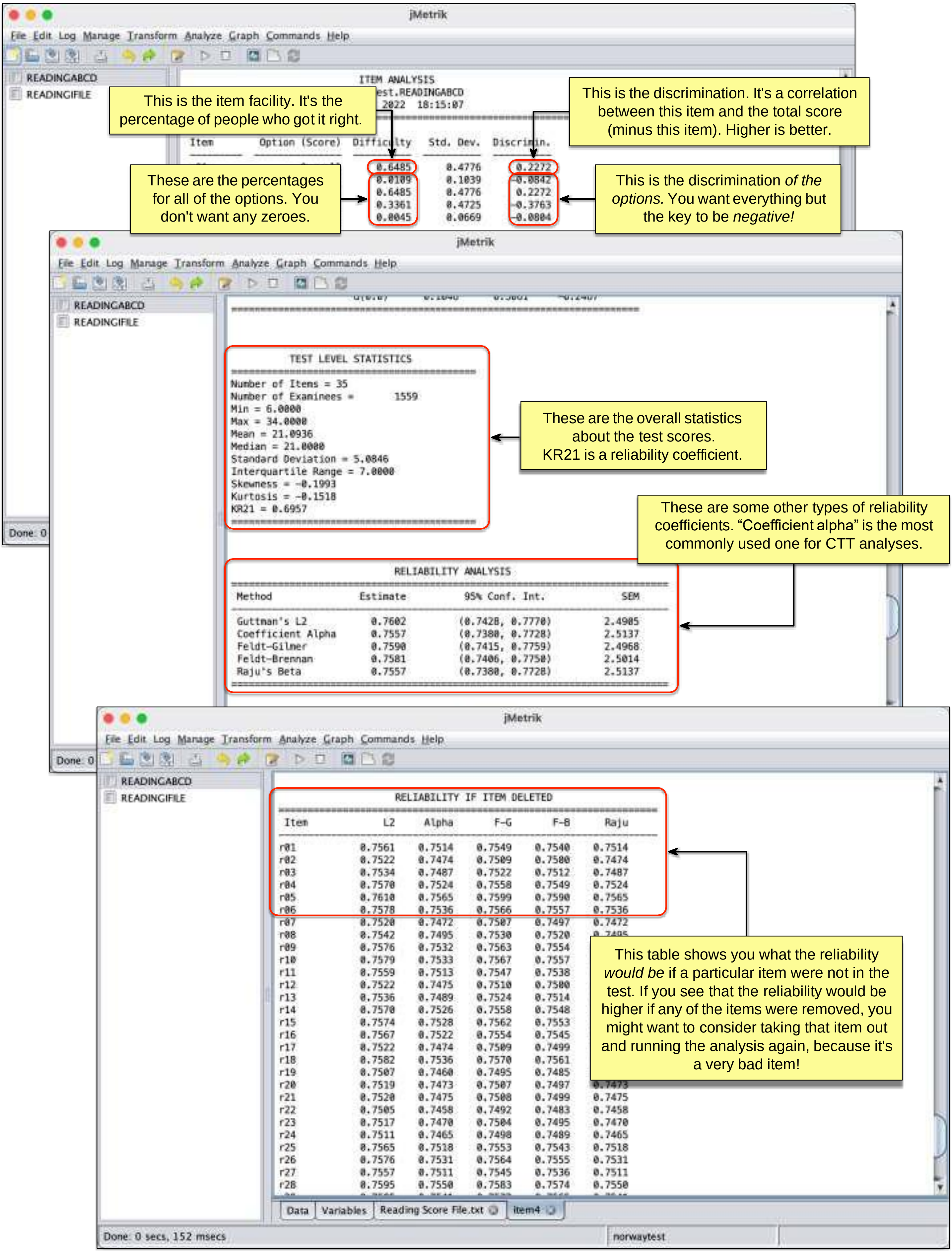
Table Name: ReadingIFILE

Options
☑ Compute item statistics
☑ Item deleted reliability
☑ All response options
□ Listwise deletion
☑ Correct for spuriousness
□ CSEM
□ Show headers
□ Unbiased covariance

Cut Score(s)

Item-total Correlation Type
● Pearson correlation
○ Polyserial correlation

6. The analysis runs and the results appear in a text window. See the next page for a breakdown.



```
                        ITEM ANALYSIS
                     est.READINGABCD
                      2022  18:15:07

Item        Option (Score)  Difficulty   Std. Dev.   Discrimin.
                              0.6485      0.4776      0.2272
                              0.0109      0.1039     -0.0842
                              0.6485      0.4776      0.2272
                              0.3361      0.4725     -0.3763
                              0.0045      0.0669     -0.0804
```

This is the item facility. It's the percentage of people who got it right.

This is the discrimination. It's a correlation between this item and the total score (minus this item). Higher is better.

These are the percentages for all of the options. You don't want any zeroes.

This is the discrimination *of the options*. You want everything but the key to be *negative!*

```
              TEST LEVEL STATISTICS
===================================================
Number of Items = 35
Number of Examinees =        1559
Min = 6.0000
Max = 34.0000
Mean = 21.0936
Median = 21.0000
Standard Deviation = 5.0846
Interquartile Range = 7.0000
Skewness = -0.1993
Kurtosis = -0.1518
KR21 = 0.6957
===================================================
```

These are the overall statistics about the test scores. KR21 is a reliability coefficient.

These are some other types of reliability coefficients. "Coefficient alpha" is the most commonly used one for CTT analyses.

```
                        RELIABILITY ANALYSIS
==========================================================================
 Method              Estimate       95% Conf. Int.               SEM
--------------------------------------------------------------------------
 Guttman's L2         0.7602      (0.7428, 0.7770)            2.4985
 Coefficient Alpha    0.7557      (0.7380, 0.7728)            2.5137
 Feldt-Gilmer         0.7590      (0.7415, 0.7759)            2.4968
 Feldt-Brennan        0.7581      (0.7406, 0.7750)            2.5014
 Raju's Beta          0.7557      (0.7380, 0.7728)            2.5137
==========================================================================
```

```
                     RELIABILITY IF ITEM DELETED
=====================================================================
 Item            L2      Alpha       F-G       F-B      Raju
---------------------------------------------------------------------
r01          0.7561    0.7514    0.7549    0.7540    0.7514
r02          0.7522    0.7474    0.7509    0.7500    0.7474
r03          0.7534    0.7487    0.7522    0.7512    0.7487
r04          0.7570    0.7524    0.7558    0.7549    0.7524
r05          0.7610    0.7565    0.7599    0.7590    0.7565
r06          0.7578    0.7536    0.7566    0.7557    0.7536
r07          0.7520    0.7472    0.7507    0.7497    0.7472
r08          0.7542    0.7495    0.7530    0.7520    0.7495
r09          0.7576    0.7532    0.7563    0.7554
r10          0.7579    0.7533    0.7567    0.7557
r11          0.7559    0.7513    0.7547    0.7538
r12          0.7522    0.7475    0.7510    0.7500
r13          0.7536    0.7489    0.7524    0.7514
r14          0.7570    0.7526    0.7558    0.7548
r15          0.7574    0.7528    0.7562    0.7553
r16          0.7567    0.7522    0.7554    0.7545
r17          0.7522    0.7474    0.7509    0.7499
r18          0.7582    0.7536    0.7570    0.7561
r19          0.7507    0.7460    0.7495    0.7485
r20          0.7519    0.7473    0.7507    0.7497    0.7473
r21          0.7520    0.7475    0.7508    0.7499    0.7475
r22          0.7505    0.7458    0.7492    0.7483    0.7458
r23          0.7517    0.7470    0.7504    0.7495    0.7470
r24          0.7511    0.7465    0.7498    0.7489    0.7465
r25          0.7565    0.7518    0.7553    0.7543    0.7518
r26          0.7576    0.7531    0.7564    0.7555    0.7531
r27          0.7557    0.7511    0.7545    0.7536    0.7511
r28          0.7595    0.7550    0.7583    0.7574    0.7550
```

This table shows you what the reliability *would be* if a particular item were not in the test. If you see that the reliability would be higher if any of the items were removed, you might want to consider taking that item out and running the analysis again, because it's a very bad item!

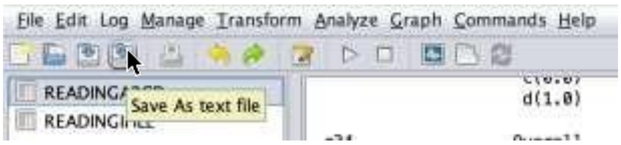
4. Exporting the results

We often want to use the results in other programs, or just save them so they are easy to send to others. This section shows you how to do that both as a text file and as a CSV that you can open in Excel.

Exporting as text

Exporting the output as text preserves all the headings, etc. of the main output window. If you don't need to do anything with the data but look at it, this is usually the best option.

1. Click the "Save as text file" button in the jMetrik toolbar (4th from the left).

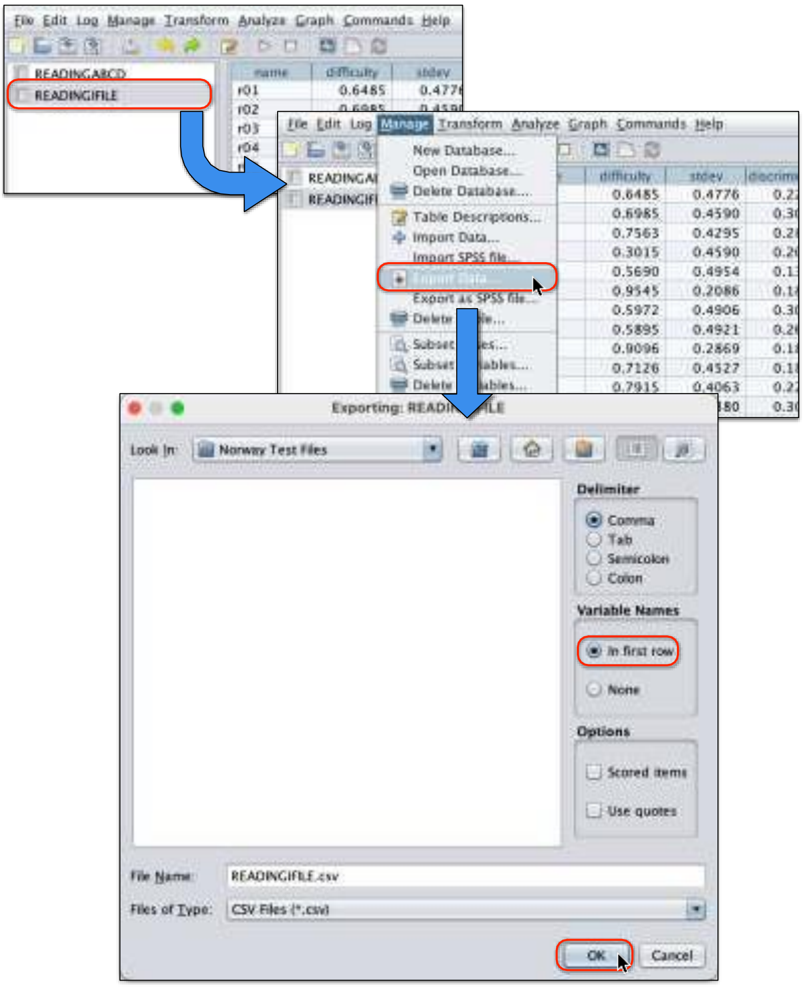


2. Choose a location, give it a name (I like to call these "IFILE"} and save like any other file.

Exporting as a CSV

“CSV" means “comma-separated values," and it's a kind of table that will open in just about anything. It will probably open in Excel by default on your computer. This is the best option if you want to do things like sort the list of items by various metrics. ***NOTE:*** **For you to be able to do this, you had to have clicked the "Save" button when you set up the item analysis. If you didn't, you can re-do the analysis and click the "Save" button. It won't hurt anything.**

1. Click on the name of the item analysis table in the list on the left side of the main jMetrik window.
2. Go to the Manage menu and select "Export Data."
3. Make sure that the "Variable Names" section has "In first row" selected. (Default}
4. Select where you'd like to save the file, give it a name, and click OK!



Delimiter: ● Comma ○ Tab ○ Semicolon ○ Colon

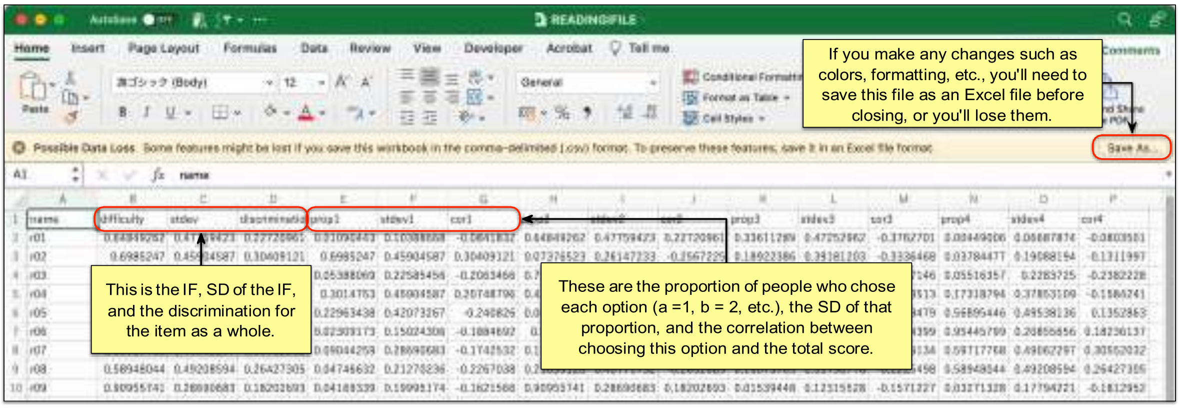
Variable Names: ● In first row ○ None

Options: □ Scored items □ Use quotes

File Name: READINGIFILE.csv

Files of Type: CSV Files (*.csv)

5. You can now open the file in Excel. ***NOTE:* You will need to save it as an Excel file (.xlsx) if you make any formatting changes, or they will be lost.** Here is how you read the file in Excel:


READINGIFILE
Home Insert Page Layout Formulas Data Review View Developer Acrobat Tell me
Save As...

Possible Data Loss Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 name

name difficulty stdev discrimination prop1 stdev1 cor1 prop2 stdev2 cor2 prop3 stdev3 cor3 prop4 stdev4 cor4

If you make any changes such as colors, formatting, etc., you'll need to save this file as an Excel file before closing, or you'll lose them.

This is the IF, SD of the IF, and the discrimination for the item as a whole.

These are the proportion of people who chose each option (a =1, b = 2, etc.), the SD of that proportion, and the correlation between choosing this option and the total score.

Conclusion

That's the basics. The program does much, much more, and it's astonishing that it's free. If you continue on the tester's path, I hope you find it useful. I have!

Call for Papers

Shiken: A Journal of Language Testing and Evaluation in Japan is seeking submissions for future issues on an ongoing basis. After checking the guidelines for publication below, please send your submission to the editor at tevalpublications@gmail.com.

Overview

Shiken aims to publish articles concerning language assessment issues relevant to assessment professionals, researchers, classroom practitioners and language program administrators. *Shiken* is dedicated to publishing articles on assessment in various educational contexts in and around Japan.

Article formats include research papers, replication studies, and review articles, all of which should typically not exceed 7000 words; as well as informed opinion pieces, technical advice articles, and interviews, all of which should typically not exceed 3000 words. Please contact the editor if you wish to submit an article that differs from these formats.

Novice researchers are encouraged to submit, but should aim for papers that focus on a single main issue. Please review recent issues of *Shiken* to understand the level of detail, depth of discussion and provision of empirical evidence that is required for acceptance.

Format

To facilitate double-blind peer review, the name(s) of the author(s) should not be mentioned in the manuscript. All citations and references to the author or co-author's work should read (Author, Year) e.g. "(Author, 2017)."

Articles should be formatted according to the guidelines of the *Publication Manual of the American Psychological Association (APA), 7th Edition*. Submissions should be formatted in Microsoft Word (.docx format) using 12-point Times New Roman font. The page size should be set to A4, with a 2.5 cm margin. The body text should be block justified, and single-spaced. Paragraphs should be separated by a blank line space. Each new section of the paper should have a section heading. Please review the most recent issues of *Shiken* for examples of section headings.

Research articles must be preceded by an abstract that succinctly summarizes the article content in less than 200 words. The reference section should begin on a new page immediately following the body text. Authors are responsible for comprehensive and accurate referencing including adding DOI or URL information wherever possible. Tables and figures should be numbered and titled. Separate sections for tables and figures should follow the references. Any appendices should be numbered and should follow the tables and figures.

Evaluation

All papers are double-blind peer-reviewed by two expert reviewers. Initial evaluation is usually completed within four weeks. The whole process from submission to publication is normally completed within six months.