
Assessing L2 proficiency growth: Considering regression to the mean and the standard error of difference

Rie Koizumi¹, Yo In'nami², Junichi Azuma³, Keiko Asano¹, Toshie Agawa¹, and Derek Eberl¹
rkoizumi@juntendo.ac.jp

1. *Juntendo University*

2. *Chuo University*

3. *Kobe Gakuin University*

Abstract

Regression to the mean (RTM) and the standard error of difference (SED) are two artifacts commonly observed in pretest–posttest designs, but they are rarely addressed in practice. We examined whether second language (L2) learners' change in scores reflected change in their L2 proficiency, by investigating whether their actual scores exceeded those that considered RTM and SED; we did so by using pretest–posttest data of the Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP) at a Japanese university across three years. We found moderate degrees of RTM, but also found that more than one-third (33.33–46.03%) of students increased their scores beyond RTM and the SED. We discuss the importance of considering RTM and SED in analyzing pretest–posttest data.

Keywords: considering errors in practice, pretest–posttest data, TOEFL ITP, Japanese university students

Change in pretest and posttest scores is often investigated using descriptive statistics such as means, or statistical significance tests such as paired *t*-tests. However, such change is subject to many factors other than change in true ability. Examples of other factors are maturation, history, and practice effects (Campbell & Stanley, 1963; Cook & Campbell, 1979; Shadish, Cook, & Campbell, 2002). Thus, analysis in change requires careful examination beyond descriptive statistics and/or statistical significance tests, before we can confidently conclude that a change in ability was indeed observed.

Regression to the mean (RTM) and the standard error of measurement (SEM) are two factors that affect pretest–posttest score changes and are commonly observed in language testing. RTM refers to a situation where pretest scores farther from the mean are probabilistically likely to cluster around the posttest mean. Thus, students who scored much lower than the pretest mean tend to increase their scores in the posttest more than those who scored a little lower than the pretest mean. Alternatively, students who scored much higher than the pretest mean tend to lower their scores more than those who scored a little higher than the mean.

Another consideration in pretest–posttest designs is that every test score includes measurement error, often known in practice in the form of the standard error of measurement (SEM). SEM refers to “the standard deviation of an individual's observed scores from repeated administrations of a test (or parallel forms of a test) under identical conditions”; it is “usually estimated from group data” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, pp. 223–224). For example, when a test taker obtains a score of 500, due to the SEM, his or her true score could be 495 or 510. This suggests that these three scores are all within the margin of error and that seemingly different scores are mere artifacts owing to measurement error. SEM is often used to interpret a single score or to compare two scores from the same test. When we compare two scores from the same test or parallel tests administered in different occasions to the same person, we can instead use a special version of the SEM that is tailored for repeated measurements. This is called the standard error of difference, and it refers to the measurement of error for scores obtained

from two test administrations (Harvill, 1991).¹ The standard error of difference is typically abbreviated as SE_{diff} or SED—the latter of which we use in this paper.

Although RTM and SEM (and SED) are frequently observed in language testing and widely known in the measurement literature, they seem to have been rarely considered in practice when interpreting score change. We will report on how to consider RTM and SED; we then illustrate this, using second language (L2) pretest–posttest data from the Test of English as a Foreign Language Institutional Testing Program (TOEFL ITP). For this purpose, we will first explain RTM and SED.

Regression to the Mean (RTM)

According to Campbell and Kenny (1999), many phenomena unrelated to language tests can also be explained by RTM. For example, tall parents tend to have children who are not as tall as their parents (i.e., children’s height moves nearer to the children’s mean); sports rookies who fared exceptionally well in their first year cannot be as successful in the following year (i.e., their performances regressed to those of the average players); and a sequel to a hit movie is less popular than the earlier one (i.e., the sequel performances became nearer to the mean of movies).

RTM is prevalent in language test data involving pretest–posttest or multi-time-point scores. Thus, an increase in pretest–posttest scores cannot be automatically attributed to the beneficial effect of treatment or to the actual growth of ability. Given that gain-score studies are widely used and yet their value could be undermined by RTM, many studies have proposed how to address RTM in psychology, education, medicine, and other fields (Campbell & Kenny, 1999).

There are several ways to deal with RTM, and they are categorized as those applied before data are collected and those after data are collected (Campbell & Kenny, 1999). The former method is recommended, as it allows one to design a study with RTM in mind.

There are three pre-data-collection methods. First, researchers should examine research designs carefully. For example, this includes (a) assigning the same number of participants to each group randomly in experimental/control groups designs, (b) avoiding pretest–posttest designs that have only experimental groups and no control groups, and (c) always conducting one or more pretests. A second approach is to consider the plausible effects of a third variable that affects the study findings and to model them (for example, as a covariate) when designing a study. For example, if experimental and control groups widely differ in terms of English-learning motivation, the pretest–posttest difference could be due to the treatment, the difference in their motivation, or both. Modeling a third factor—motivation, in this case—could reduce both the SEM and RTM. A third way is to use measures and tests that feature high reliability. The higher the reliability of a test is, the more consistently it measures knowledge and ability. This leads to a decrease in SEMs (e.g., Bonate, 2000). Concurrently, researchers should use tests that are shown to be parallel in content and difficulty and highly correlated when they are administered at the same time or on close occasions. This is because RTM occurs when two tests do not correlate perfectly. The more highly two tests correlate, the less RTM is observed; in contrast, a weaker intercorrelation leads to greater RTM.

In addition to these pre-data-collection methods for addressing RTM, there are several ways of dealing with this issue, even after data are collected—at the data analysis stage. We will present them in the Method section, below. All these pre- and post-data-collection methods are extensively discussed in Campbell and Kenny (1999, Chapter 10). While Marsh and Hau (2002) report that RTM cannot be completely controlled for, even with the latest analytical method of multilevel analysis, the pre- and post-data-collection methods we describe in this paper are still useful in even partially addressing the regression effect.

Despite the prevalent possibility of RTM effects in pretest–posttest designs, they have not been adequately addressed in L2 growth-assessment studies. One exception is Swinton (1983), who aimed to offer guidelines for establishing a language–growth benchmark at local institutions. His method was based on the recommendations of Cronbach and Furby (1970), and is as follows: administer three tests on the same examinees—namely, Pretest A at the beginning of a semester, Reliability Test B one week after Pretest A, and Posttest C at the end of the semester; develop a regression equation relating Reliability Test B to Pretest A, in order to estimate score change due to nonintervention (e.g., measurement error and practice effects); and examine whether a Posttest C score is higher than that expected from the regression equation. Suppose our regression equation is: $\text{Test B} = 200 + .7 * (\text{Pretest A})$, and a student scores 100 on Pretest A and 150 on Test B. Using the regression, the expected score on the Posttest C (when RTM occurs and there is no actual gain) is 270 ($= 200 + .7 * [100]$). If the actual score on the Posttest C is 280—that is to say, greater than 270—it indicates growth in ability. The gain is 10 ($= 280 - 270$), not 180 ($= 280 - 100$).

Perhaps the most comprehensive recent study on RTM is that of Marsden and Torgerson (2012). Using the database *Educational Resources Information Centre*, they conducted a methodological review of single-group, pretest–posttest empirical studies published in 13 educational research journals in 2009. Of 490 studies published in that year, 64 (13%) were found to have evaluated innovative interventions and used experimental, quasi-experimental, or pre-experimental designs. After excluding 48 studies that did not meet their inclusion criteria (e.g., studies that did not “have at least one quantified measure,” p. 588), Marsden and Torgerson had 16 studies left; they report that none of the study authors described the potential effect of RTM, although other potential factors (e.g., maturation and time) were mentioned. This indicates that RTM is not widely recognized among scholars, and that a better understanding of RTM could lead to improvements in the interpretation of study findings.

Standard Error of Difference (SED)

Test scores include errors caused by variations in the content and format of test items and whole tests, and by inconsistencies in test administration and scoring (see, for example, Fulcher, 2010; Hughes, 2003). These causes could lead to larger errors and lower test reliability, whereas tests that have high reliability have smaller test-score errors.

Equation 1 below (Harvill, 1991, p. 186, Formula 10) is used to calculate SEDs—that is, the measurement of errors for scores obtained from two administrations (Harvill, 1991). The value shows the degree to which a test score changes at 68% probability because of errors under two test administrations. In other words, it shows the 68% probability of score variation. Equation 1 indicates that tests with larger standard deviations (*SDs*) have a larger SED. Further, larger SED values indicate larger error, resulting in test scores featuring greater fluctuation (see, for example, Carr, 2011, on how to calculate *SD* and reliability). Greater confidence in score variation can be obtained by calculating the 95% probability of score variations, using Equation 2 (Harvill, 1991, p. 186); however, we use only Equation 1 because “score bands which are 68 percent confidence intervals ... are most commonly used in practice” (p. 184).

$$\begin{aligned} &\text{SED (for 68\% probability in comparing two scores from the same test taker}^2) \\ &= (\text{SD of the pretest}) * (\sqrt{2 - (\text{Reliability of the pretest}) - (\text{Reliability of the posttest})}) \end{aligned} \quad (1)$$

$$\text{SED (for 95\% probability)} = 1.96 * \text{SED (for 68\% probability)} \quad (2)$$

Suppose that the SED (for 68% probability) of the TOEFL ITP is 15. This would mean that under two test administrations, this test score can fluctuate by 15 at the probability of 68%. When a pretest score is 480 and a posttest score is 520, the error range of the pretest score is between 465 and 495; 520 is not included in this range. Thus, the pretest and posttest scores are highly likely to differ, and we can assert

this with confidence. However, if the posttest score is 485, it would be difficult to argue that the two test scores differ.

As with RTM, SEDs are not always reported or considered during score interpretation. Good practices are seen in the *TOEIC User Guide* (Educational Testing Service, 2007) and the *TOEIC Examinee Handbook* (Educational Testing Service, 2008), both of which explain that the 68% probability SED for each of the TOEIC listening and reading sections is 35. If a student's listening score improves from 300 to 340, this indicates real growth in that student's listening proficiency, as the score of 340 lies outside the 265–335 SED range.

The Current Study

We address three research questions (RQs) to examine whether RTM is observed (RQ1), and to what degree RTM and the SED affect findings (RQ2 and RQ3) in using the TOEFL ITP. The three RQs are as follows.

1. Is there any evidence of the regression to the mean (RTM) in pretest–posttest data?
2. What percentage of students increased or decreased their scores beyond RTM?
3. What percentage of students increased or decreased their scores beyond the SED?

The results could contribute to our understanding of how to separate students' real growth in proficiency from RTM and SED. This would, in turn, strengthen arguments regarding whether or not students had increased in ability.

Method

Participants and Instrument

We used data from first-year students at the tertiary level who took the TOEFL ITP (Level 1) twice—in April and December—at a private university in Chiba, across three years ($n = 120$ in 2012; $n = 125$ in 2013; $n = 126$ in 2014). The TOEFL ITP was conducted to assess growth in students' L2 English proficiency, and to place students into five English classes in the subsequent year. Additionally, each student needed to obtain a TOEFL ITP of 475 or higher, or a TOEFL Internet-based test (iBT) of 53 or higher, to advance to the second year. Thus, students were generally motivated to study hard to meet the requirement. The test was also conducted to evaluate the effectiveness of the English program.

The TOEFL ITP is designed to assess the English proficiency of nonnative speakers. It has three sections, all in paper-and-pencil multiple-choice formats; it consists (in order of appearance) of a listening section (50 items, 35 minutes), grammar section (40 items, 25 minutes), and reading section (50 items, 55 minutes). A total score ranges from 310 to 677, with an SEM of 13 (Educational Testing Service, n.d.). Each student receives score reports that show each of the three section scores and the total score.

Analyses

The TOEFL ITP total scores from the April and December administrations over the three-year period were used. Of the various methods available for analyzing the degree of RTM, we utilized two methods that we consider the most accessible. First, to address RQ1, we correlate the change scores (posttest minus pretest) and the pretest scores. If there is an RTM effect, we see a substantial and negative correlation. The higher a negative correlation is, the higher the degree of RTM will be (e.g., Marsden & Torgerson, 2012; Roberts, 1980; Rocconi & Ethington, 2009; Rogosa, 1995). The rationale here is that negative correlations are derived from lower pretest scorers who are likely gaining a higher posttest score and from higher pretest scorers who are likely gaining a lower posttest score.

While the first method is group-based and produces a single value that shows the overall extent of RTM, the second method is individual-based: It calculates expected individual posttest scores while assuming RTM, and compares them to the actual scores. Using Equation 3 below (Campbell & Kenny, 1999, p. 26), we calculated an expected posttest score per person and compared it to his or her actual posttest score, while bearing in mind RQ2. We calculated the percentage of students with actual posttest scores that were higher than their expected posttest scores—that is, the percentage of those whose growth exceeded RTM. Table 1 shows how our three-year data were applied to the Equation.

$$\text{Expected posttest score} = M_y + r_{xy}(SD_y/SD_x)(X - M_x) \quad (3)$$

Table 1
Expected Posttest Scores and Actual Posttest Scores

x	2012	2013	2014
M_y = posttest score mean	508.87	522.82	539.76
r_{xy} = correlation between pretest and posttest scores	.79	.83	.80
SD_y = standard deviation of posttest scores	34.17	42.58	36.16
SD_x = standard deviation of pretest scores	40.50	47.34	44.97
M_x = pretest score mean	507.13	508.01	510.51
X = Actual pretest score of a student (example)	500	517	557
Expected posttest score (example)	504	530	570
Actual posttest score (example)	510	547	560
	(gain)	(gain)	(no gain)

For example, for a student in 2013 with a pretest score of 517, his expected posttest score was 530 ($522.82 + .83[42.58/47.34][517 - 508.01]$); the actual posttest score was 547. This student's actual posttest score was larger than the posttest score forecast by RTM; this suggests that the score gain reflects improvement in his ability, rather than RTM. It should be noted that the standard deviations of the posttest scores were all smaller than those of the pretest scores; this indicates a narrower distribution of posttest scores and may serve as one piece of evidence of RTM.

To examine RQ3, Equation 1 was used; SEDs for 68% probability were calculated as follows.

$$\begin{aligned} \text{SED (for 68\% probability in comparing two scores from the same test taker)} \\ = (\text{Standard deviation of the pretest}) * (\sqrt{2 - (\text{Reliability of the pretest}) -} \\ (\text{Reliability of the posttest})) \end{aligned} \quad (1)$$

$$\text{SED in 2012} = (40.50) * (\sqrt{2 - (.96) - (.96)}) = (40.50) * (0.28) = 11.46$$

$$\text{SED in 2013} = (47.34) * (\sqrt{2 - (.96) - (.96)}) = (47.34) * (0.28) = 13.39$$

$$\text{SED in 2014} = (44.97) * (\sqrt{2 - (.96) - (.96)}) = (44.97) * (0.28) = 12.72$$

We used the reliability of .96 for the TOEFL ITP, as reported by Educational Testing Service (n.d.), as the reliability index for both the pretest and posttest.³ We calculated the percentage of students who had actual posttest scores higher than the SED—that is, the percentage of those whose growth exceeded the SED. All analyses were conducted using the *Comparing paired samples* (which includes paired *t*-tests) and *Correlation* pages in the langtest.jp Web App (Mizumoto, n.d.). The app runs on several well-known R packages and produces various useful figures based on data pasted into its website's designated space.

In summary, we will (a) examine correlations between change scores and pretest scores (for RQ1), (b) calculate the predicted posttest scores and the percentages of students whose posttest scores exceeded the predicted ones (for RQ2), (c) consider the SED and calculate the percentages of students whose posttest scores exceeded the SED (for RQ3), and (d) synthesize the findings from (b) and (c).

Results and Discussion

Data distribution

Means, standard deviations, and correlations between the pretest and posttest scores in the 2012–2014 data are presented in Table 1 above. Boxplots of pretest–posttest scores and changes in individual scores in 2012, meanwhile, are presented in the left-hand panel of Figure 1, while the right-hand panel shows clearly that pretest scores—especially extreme ones—tend to converge to the posttest mean. This is consistent with the smaller standard deviation for the posttest (34.17, compared to 40.50 for the pretest [see Table 1]). These results suggest a certain degree of RTM in the 2012 data.

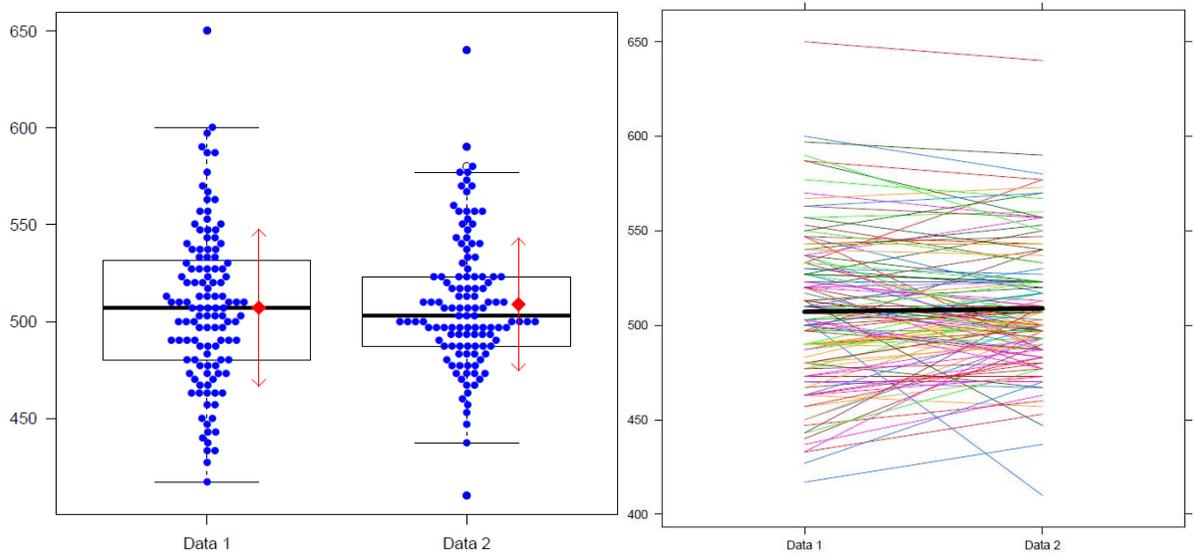


Figure 1. Left: Boxplots of pretest–posttest data in 2012. Data 1 = Pretest scores; Data 2 = Posttest scores. ± 1 standard deviations are represented by arrows. See Field (2009) for interpretation of the boxplots. This note also applies to the left-hand panel of Figures 2 and 3. Right: Changes in individual scores between pretest and posttest data in 2012. Thick line indicates the mean difference. This note also applies to the right-hand panel of Figures 2 and 3.

The left-hand panel of Figure 2 shows pretest–posttest scores from the 2013 data, and the right-hand panel shows overall that extreme pretest scores were likely to converge toward the posttest mean. The posttest standard deviation was smaller than the pretest standard deviation (42.58 and 47.34, respectively). These results, again, provide some evidence of RTM in the 2013 data.

Figure 3 shows pretest–posttest scores from the 2014 data, with the right-hand panel showing overall that extreme pretest scores were likely to converge toward the posttest mean. The posttest standard deviation was smaller than the pretest standard deviation (36.16 and 44.97, respectively). Again, these results provide some evidence of RTM in the 2014 data.

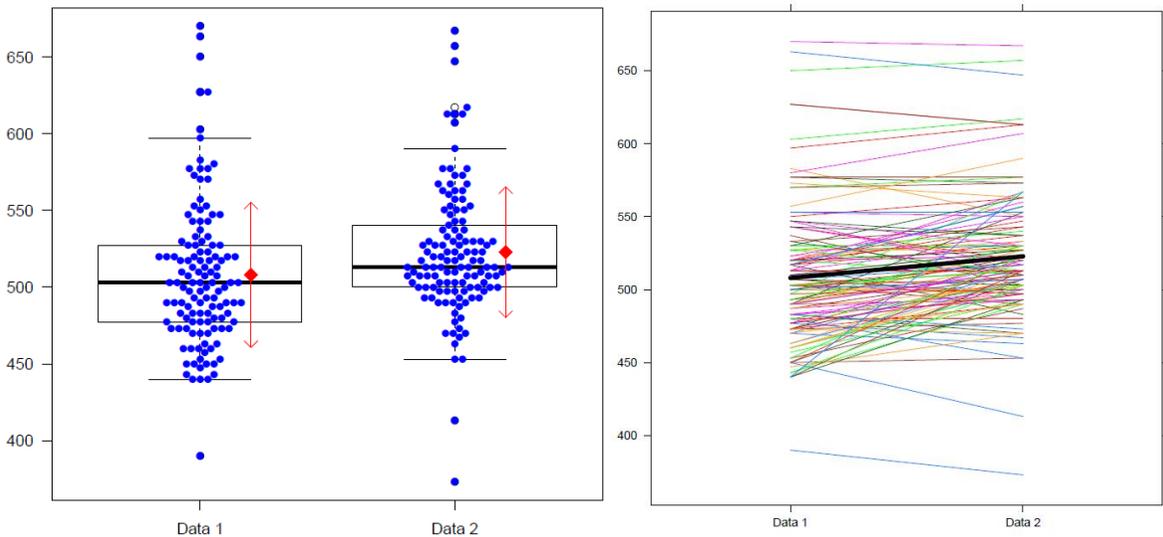


Figure 2. Left: Boxplots of pretest–posttest data in 2013. Right: Changes in individual scores between pretest and posttest data in 2013.

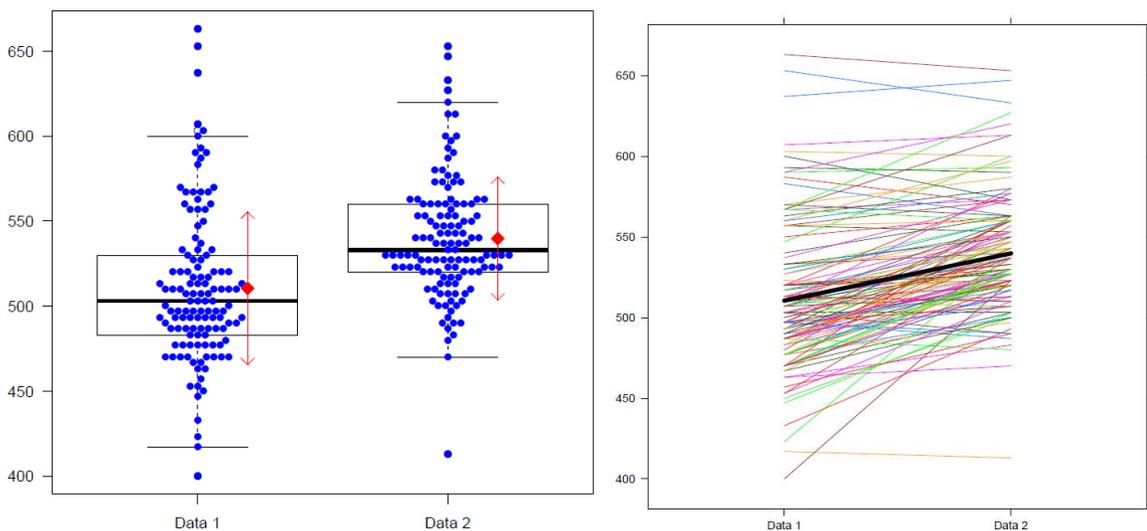


Figure 3. Left: Boxplots of pretest–posttest data in 2014. Right: Changes in individual scores between pretest and posttest data in 2014.

For reference purposes, we used paired t -tests to compare the pretest–posttest scores. Table 2 shows that in 2012, there was no significant difference between the two scores, with a negligibly small effect size according to Plonsky and Oswald’s (2014) guideline for interpreting effect sizes (i.e., within-group contrast $d = 0.6$ for small, 1.0 for medium, and 1.4 for large). In 2013, there was a significant difference with a negligibly small effect size, whereas in 2014, there was a significant difference between the two scores with a small effect size.

Table 2
Results of Paired T-tests and Effect Sizes

	Paired t-test	Effect size
2012	$t = -0.76, df = 119, p = .45$	$d [95\% CI] = 0.05 [-0.07, 0.16], g = 0.04 [-0.07, 0.16], \delta = 0.04$
2013	$t = -6.24, df = 124, p < .01$	$d [95\% CI] = 0.32 [0.22, 0.43], g = 0.32 [0.22, 0.43], \delta = 0.31$
2014	$t = -12.27, df = 125, p < .01$	$d [95\% CI] = 0.69 [0.56, 0.81], g = 0.68 [0.56, 0.80], \delta = 0.65$

Note. d = Cohen's d ; g = Hedges' g ; δ = Glass's delta.

RQ1: Is there any evidence of the regression to the mean (RTM) in pretest–posttest data?

Figures 4–6 show scatterplots and correlations between change scores and pretest scores; the relationships therein are consistently negative and moderate ($r = -.54, -.45, \text{ and } -.59$, respectively). Negative correlations indicate RTM, and the higher they are, the greater the degree of RTM (e.g., Roberts, 1980; Rocconi & Ethington, 2009; Rogosa, 1995). This is because negative correlations suggest a greater change in scores when pretest scores are lower, as well as less-positive-change scores or more-negative-change scores when pretest scores are higher. Negative and moderate correlations suggest moderate degrees of RTM. As evidenced by negative and larger correlations, the 2014 data show more serious RTM than the 2012 and 2013 data, and the 2012 data show more serious RTM than the 2013 data.

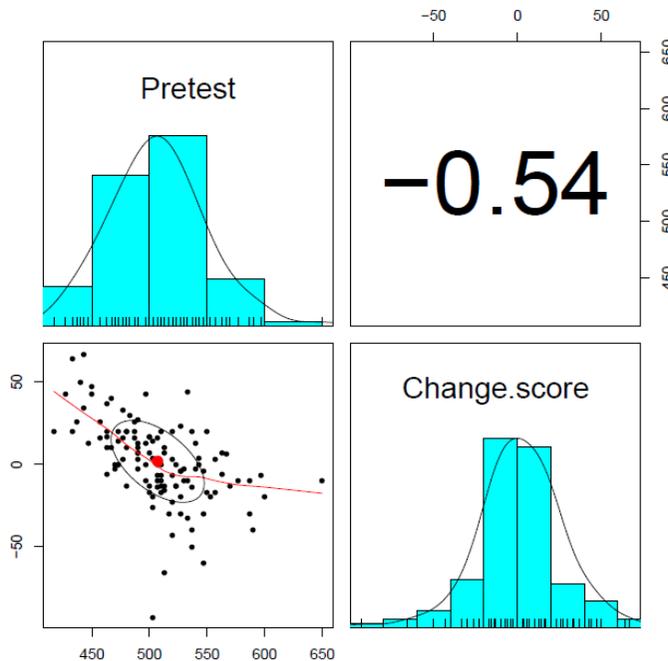


Figure 4. Scatterplot and correlations in 2012. The left bottom scatterplot has pretest scores on the X axis and pre–post change scores on the Y axis. The red curve shows loess smooths, the large black circle shows correlation ellipses, and the large red dot indicates the means of the X and Y axes (see Ogasawara, 2014; Revelle, 2014).

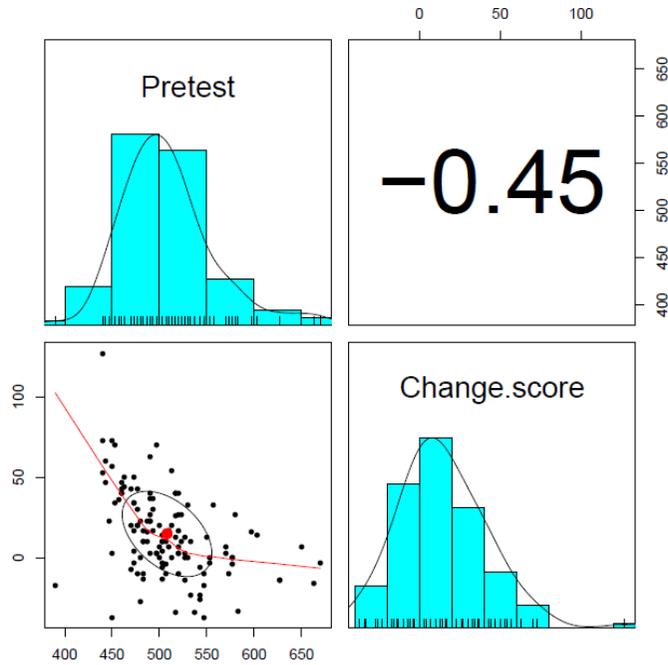


Figure 5. Scatterplot and correlations in 2013.

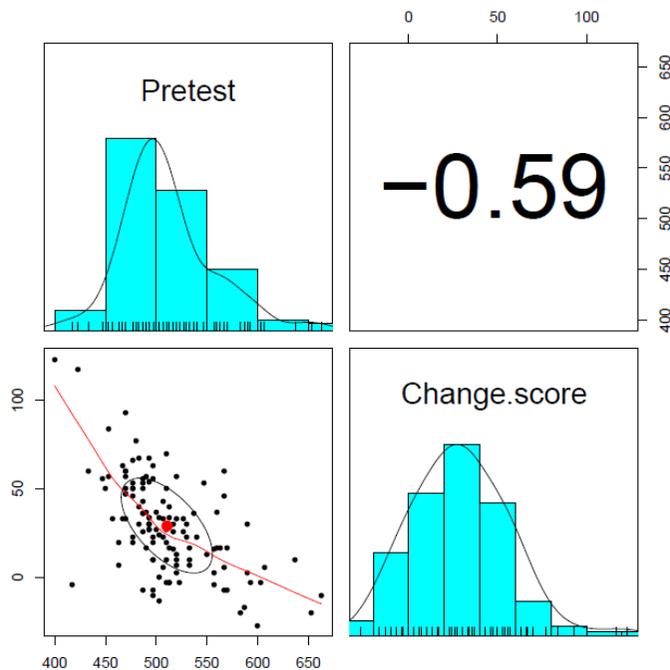


Figure 6. Scatterplot and correlations in 2014.

RQ2: What percentage of students increased or decreased their scores beyond RTM?

Table 3 compares the expected and actual posttest scores. The rightmost column presents favorable results, showing the number of students whose posttest scores exceeded the predicted ones on the basis of RTM. Across three years, approximately half of the students (51.67%, 49.60%, and 50.00%, respectively) showed a posttest score gain that exceeded that forecast by RTM.

Table 3
Comparisons of Expected and Actual Posttest Test Scores

	Expected > Actual	Expected = Actual	Expected < Actual
2012 (n = 120)	56 (46.67%)	2 (1.67%)	62 (51.67%)
2013 (n = 125)	62 (49.60%)	1 (0.80%)	62 (49.60%)
2014 (n = 126)	61 (48.41%)	2 (1.59%)	63 (50.00%)

Note. Expected = Expected posttest score; Actual = Actual posttest score. This note also applies to Table 4.

RQ3: What percentage of students increased or decreased their scores beyond the SED?

We used the SEDs of the 68% probability of score variability, and the values differed across years: 11.46 in 2012, 13.39 in 2013, and 12.72 in 2014. Table 4 shows that in 2012, 26.67% of the first-year students had lower actual posttest scores than their pretest score minus SED, 38.33% had actual posttest scores between their pretest score minus SED and their pretest score plus SED, and 35.00% had higher actual posttest scores than their pretest score plus SED. Thus, we can see in the fourth column that 35.00% of the first-year students in 2012, 47.20% in 2013, and 72.22% in 2014 had higher scores than the 68% probability SED. As mentioned by the reviewer, we are statistically supposed to have 16% of the students above the 68% probability SED (as 32% should be outside the 68% probability range and 16% should be above this range). The percentages 35.00–72.22% were all beyond 16%, which can be interpreted as a favorable result in examining growth.

Table 4
Comparisons of SED Range and Actual Posttest Test Scores

	Actual < [IPS - SED]	[IPS - SED] ≤ Actual ≤ [IPS + SED]	[IPS + SED] < Actual
2012	32 (26.67%) ^a	46 (38.33%)	42 (35.00%)
2013	15 (12.00%)	51 (40.80%)	59 (47.20%)
2014	5 (3.97%)	30 (23.81%)	91 (72.22%)

Note. IPS = Individual pretest score. ^a32/120*100.

Further, we combined the two results based on RTM and the SED and investigated what percentage of students increased their scores beyond both RTM and the SED. Although not reported in the tables, we found that in the case of 68% probability, 33.33% (n = 40) of the first-year students in 2012, 40.00% (n = 50) in 2013, and 46.03% (n = 58) in 2014 had higher scores than those forecast by RTM and the SED. Therefore, from the viewpoint of RTM and SED, more than one-third of the students earned scores that exceeded both RTM and the SED. Thus, we can reasonably claim that such students indeed increased their ability.

Conclusion

The current study aimed to examine students' genuine growth in English-language proficiency, while considering regression to the mean (RTM) and the standard error of difference (SED). It examined RTM by (a) using correlations between change scores and pretest scores and (b) calculating the percentages of students whose posttest scores exceeded those predicted by RTM. It also examined the SED by (c) calculating the percentages of students whose posttest scores exceeded the SED range. We found moderate degrees of RTM, but also found that more than one-third (33.33–46.03%) of students increased their scores beyond RTM and the SED. Additionally, we discussed the importance of considering RTM and the SED in analyzing pretest–posttest data.

In response to the relative dearth of studies that address these two artifacts, this study has shown how they can be examined, by using real-world data. The equations we used are simple, and the necessary values can be estimated by using simple statistics; the results were useful in providing stronger evidence of claims of growth. In practice, this in turn allows teachers and researchers to offer feedback to students with greater confidence. This is particularly true of the methods we used to investigate RQ2 and RQ3, in which each student's change in scores was examined against the change expected from RTM and SED.

Our study has three limitations. First, to use Equations 1 and 2 to calculate SED, we used the reliability of the TOEFL ITP, as reported by Educational Testing Service (n.d.). That report does not provide details of the examinees from whom the reliability had been calculated (e.g., the number of examinees or nationalities); however, we assume that the reliability of .96 was higher than that we would have obtained had we had access to the raw data, because reliability estimates publicly reported are usually based on representative samples of the populations of test takers. If we had used a reliability lower than .96, the SEDs would have increased. Thus, our current result using the smaller SEDs may have made more students seem to have improved more than they actually did (see also Notes 3 and 4). Second, although we had had only pretest–posttest scores of the TOEFL ITP, access to other instruments that relate to RTM could have allowed us to better examine the impact of RTM on change in pretest–posttest scores. For example, if there were covariates that were assessed during the period when pretest scores were assessed, we could have used additional statistical analyses that can take RTM effects into account. For example, analysis of covariance (ANCOVA)—a combination of analysis of variance (ANOVA) and regression analysis—allows researchers to adjust posttest scores that are affected by RTM by utilizing as covariates data other than pretest–posttest scores (e.g., Bonate, 2000; Chuang-Stein & Tong, 1997). Further, structural equation modeling (SEM) allows one to control for measurement error when constructing a model, and this can help researchers interpret posttest scores more precisely (e.g., Kline, 2011). Third, we used a single-group, pre–post design with no control group; this design is not desirable, given a number of threats that undermine causal interpretation (Campbell & Kenny, 1999). Replication studies that feature a control group are needed, if we are to more rigorously examine proficiency gain.

We hope that our examination of two types of statistical artifact is helpful for teachers and researchers who are interested in examining proficiency growth.

Notes

¹ Instead of using SED, we could instead use two SEMs from two tests by determining whether the two confidence intervals of the two SEM ranges overlap. We did not use this method, as the SED seems to be more widely used. Readers interested in the SEM method should refer to note 2, below.

² The Equation differs when calculating the SED for 68% probability in comparing *different* test-takers' scores on the same test as follows:

$$\sqrt{2} * [SD \text{ of the test}] * \sqrt{(1 - [\text{Reliability of the test}])} = \sqrt{2} * SEM$$

Interested readers are directed to Harvill (1991) for details.

³ As pointed out by the reviewer, the SEDs used in this study (11.46 to 13.39) may be a lower estimation. Educational Testing Service (n.d.) states that the SEM of the TOEFL ITP is 13.0. The SED is usually larger than the SEM, because the SED considers errors (SEMs) of a pretest and posttest (typically shown in the Equation in Note 2; however, this is an equation for a different purpose). The lower SEDs values may have been caused by the use of the higher reliability of .96, which was likely derived from representative samples of the populations of test takers with a larger range of proficiency levels than ours that were based on students at one university. Although Educational Testing Service (n.d.) does not report the type of test takers used to compute this value, we assume that the reliability was higher as values publicly reported are usually based on representative samples. However, this is the only value we could obtain because we require raw data to calculate the reliability based on the students in this study. Because we used a possibly inflated value of reliability in the calculation, our SEDs may be lower. Thus, this is one of the limitations in our study.

⁴ It should be noted that the Equation in Note 2 cannot be used in our context because it should be used in comparing *different* test-takers' scores on the same test. On the other hand, Equations 1 and 2 should be used in comparing two scores from the *same* test taker (Harvill, 1991).

Acknowledgments

This work was partially supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant-in-Aid for Scientific Research (C), Grant Number 26370737. We are deeply indebted to the editor, the reviewer, and Atsushi Mizumoto for his invaluable comments.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Bonate, P. L. (2000). *Analysis of pretest–posttest designs*. Boca Raton, FL: Chapman & Hall/CRC.
- Campbell, D. T., & Kenny, D. A. (1999). *A primer on regression artifacts*. New York, NY: Guilford.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, UK: Oxford University Press.
- Chuang-Stein, C., & Tong, D. M. (1997). The impact and implication of regression to the mean on the design and analysis of medical investigations. *Statistical Methods in Medical Research*, 6, 115–128. doi:10.1177/096228029700600203
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”—OR should we? *Psychological Bulletin*, 74, 68–80. doi:http://dx.doi.org/10.1037/h0029382
- Educational Testing Service. (2007). *TOEIC user guide—Listening and reading*. Princeton, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/Test_of_English_for_International_Communication/TOEIC_User_Gd.pdf
- Educational Testing Service. (2008). *TOEIC examinee handbook—Listening and reading*. Ewing, NJ: Author. Retrieved from http://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf

- Educational Testing Service. (n.d.). TOEFL® ITP. Retrieved from https://www.ets.org/s/toefl_itp/pdf/toefl_itp_score.pdf
- Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London, UK: Sage.
- Fulcher, G. (2010). *Practical language testing*. London, UK: Hodder Education.
- Harvill, L. M. (1991). An NCME instructional module on standard error of measurement [Instructional topics in educational measurement]. *Educational Measurement: Issues and Practice*, 10(2), 181–189. Retrieved from <http://www.ncme.org/pubs/items/16.pdf>
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.
- Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.
- Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, 38, 583–616. doi:10.1080/03054985.2012.731208
- Marsh, H. W., & Hau, K.-T. (2002). Multilevel modeling of longitudinal growth and change: Substantive effects or regression toward the mean artifacts. *Multivariate Behavioral Research*, 37, 245–282. doi:10.1207/S15327906MBR3702_04
- Mizumoto, A. (n.d.). *Langtest*. Retrieved from <http://langtest.jp/#app>
- Ogasawara, O. (2014). *R graphical manual*. Retrieved from http://rgm3.lab.nig.ac.jp/RGM/R_rdfile?f=psych/man/pairs.panels.Rd&d=R_CC
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. doi:10.1111/lang.12079
- Revelle, W. (2014). An overview of the psych package. Retrieved from <http://cran.r-project.org/web/packages/psych/vignettes/overview.pdf>
- Roberts, A. O. H. (1980). Regression toward the mean and the regression-effect bias. In G. Echternacht (Guest Ed.), *New directions for testing and measurement* (Vol. 8, pp. 59–82). San Francisco, CA: Jossey-Bass Inc.
- Rocconi, L. M., & Ethington, C. A. (2009). Assessing longitudinal change: Adjustment for regression to the mean effects. *Research in Higher Education*, 50, 368–376. doi:10.1007/s11162-009-9119-x
- Rogosa, D. (1995). Myths and methods: “Myths about longitudinal research” plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Erlbaum.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Swinton, S. S. (1983). *A manual for assessing language growth in instructional settings*. TOEFL Research Report, 83–17. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/research/policy_research_reports/rr-83-17_toefl-rr-14