International Journal of Educational Methodology
Volume 11, Issue 1, 17 - 25
ISSN: 2469-9632
https://www.ijem.com/

The Puzzle of Regression to the Mean

Don A. Affognon[*]
University of Nevada, USA

Received: November 28, 2024 ▪ Revised: December 22, 2024 ▪ Accepted: January 25, 2025

Abstract: Although regression to the mean is pervasive in data analysis, educational researchers often misconstrue it as evidence of genuine change and mistakenly attribute random changes to treatment effects. A statistical phenomenon where extreme values naturally move closer to the average after repeated treatment, regression to the mean is especially susceptible to misinterpretations in educational studies with pretest-posttest or longitudinal designs. In such studies, observed changes are frequently assumed to be the effects of treatment, even in cases where the changes are statistical artifacts. Using a hypothetical case and two real-world studies, this paper investigates the technical challenges that regression to the mean poses and introduces a hybrid Bayesian model that mitigates its effects more effectively than conventional approaches, such as multiple baseline adjustments and formulaic corrections. In particular, the hybrid Bayesian model relies on multiple baseline measurements to minimize distortions associated with regression to the mean during the pretest phase and leverages prior knowledge—such as standard deviations and population means—to refine post-test data adjustments. It follows that the model provides educational researchers with an innovative tool for accurately evaluating interventions and enhancing the effectiveness of various research-driven educational policies and practices.

Keywords: *Bayesian regression, causal inferences, pretest–posttest designs, regression to the mean.*

To cite this article: Affognon, D. A. (2025). The puzzle of regression to the mean. *International Journal of Educational Methodology*, *11*(1), 17-25. https://doi.org/10.12973/ijem.11.1.17

Introduction

Regression to the mean (RTM) is a purely statistical phenomenon that exerts a pervasive effect on datasets, yet many of its implications are widely misunderstood (Pinker, 2021). Whenever extreme values in a dataset move toward the average after repeated treatment, researchers often misinterpret such movements as evidence of causal effects rather than recognizing them as mere statistical artifacts (Isaac & Michael, 1995; Linden, 2013; Smith & Smith, 2005). In educational research, this tendency to misattribute random changes to treatment effects is problematic not just for the assessment of teacher effectiveness but also for the evaluation of teaching methods and curricula. In other words, the failure to recognize RTM can lead to inaccurate claims about the effectiveness of certain interventions, instructional methods, and curriculum changes. This problem is particularly pronounced in academic studies that rely upon pretest-posttest or longitudinal designs (Asbury, 1974; see also Kahneman & Tversky, 1973). In educational research, the problem is compounded by the heavy reliance on standardized test scores and other such outcome measures that are inherently variable, often leading researchers to overlook RTM or treat random fluctuations as real performance changes.

For example, Smith and Smith (2005) reported cases in which researchers attributed improvements in test scores to instructional interventions, even when many of the changes were due to statistical regression to the mean. Thorndike (1942) provided another example of this interpretation problem in a study where college students who had high standardized test scores but low grade-point averages (GPA) were considered “underachievers,” while their peers who had high GPAs but low-test scores were all considered “overachievers.” As with other cases, these interpretations did not reflect the possibility that RTM could explain such discrepancies. Furthermore, the psychologists Daniel Kahneman and Amos Tversky have emphasized that not only does RTM distort data interpretation, but it also leads to misguided policy decisions arising from a failure to account for statistical nuances. In a widely referenced paper, Kahneman and Tversky (1973) described how a group of flight school instructors enacted a training policy of positive reinforcement. Following the advice of school psychologists, the instructors consistently praised students for successfully executing complex maneuvers. However, they later observed that outstanding performances were often followed by a measurable decline

[*] **Correspondence:**
Don Affognon, 4505 S. Maryland Parkway, Las Vegas, NV 89154, USA. ✉ affognon@unlv.nevada.edu

© 2025 The author(s); licensee IJEM by RAHPSODE LTD, UK. **Open Access** - This article is distributed under the terms and conditions of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/).

in subsequent attempts. Misinterpreting this pattern as evidence against the effectiveness of positive reinforcement, they abandoned the policy altogether. What they failed to recognize was that these changes were not causal effects but rather predictable instances of RTM (Kahneman & Tversky, 1973).

This paper serves two purposes. First, it discusses how RTM insinuates itself in educational studies involving groups of students and how it is routinely addressed during study design or data analysis. Second, it relies on a hypothetical dataset and two empirical cases to contextualize RTM and introduce a hybrid Bayesian model that more effectively mitigates its distorting effects. The underlying argument is that despite its prevalence in educational research, RTM is not intractable.

Theoretical Framework

RTM is a counterintuitive statistical phenomenon that was first identified in two major experiments carried out by the polymath Francis Galton (1822–1911). In the first experiment, Galton (1886) studied RTM by planting sweet pea seeds of different sizes, grouping them into seven categories based on diameter. He then documented the offspring seeds that each plant yielded and calculated the average diameter of 100 seeds from each. The results revealed a distinct pattern: Whereas the smallest seeds tended to produce slightly larger offspring, the largest seeds produced slightly smaller ones, causing offspring seed sizes to regress toward the average diameter of the 100 seeds. Galton called this pattern *regression to the mean*, a term that would later give rise to the concept of regression in modern statistics. In the second experiment, Galton expanded his study of the same phenomenon to human height by reviewing the family records of 205 pairs of parents and their 928 adult children. Because the average man is 8% taller than the average woman, Galton scaled female heights by multiplying them by a factor of 1.08 before comparing them to male heights. He then calculated an average mid-parent height by averaging the heights of each mother and father. He then divided all the mid-parent heights into nine categories and calculated the median height of the children in each category.

Once more, Galton's (1886) findings demonstrated RTM: Parents who were taller or shorter than average had children whose heights gravitated toward the population average. Specifically, for every inch a parent's height deviated from the mean, the child's height was more likely to deviate by only 0.69 inches in the same direction. The consistency in the results of these seminal studies and the many others that followed make RTM a real statistical problem. However, as Pinker (2021) has suggested, RTM applies not only to the heights and IQs of parents and their children but also, more generally, to any two variables that are not perfectly correlated. An extreme value in one variable will tend to be associated with a less extreme value in the other. This, of course, does not mean that, eventually, all children will be of average height. Nor does it mean that the population will converge to an IQ of 100. What it does mean is that whenever an extreme value appears in a bell-shaped distribution, any other variable that is paired with it is unlikely to duplicate its outlier status. The puzzle is that in too many research studies, RTM is mistaken for evidence of change (Nesselroade et al., 1980; Pinker, 2021; see also Stigler, 1997). For example, in a study examining reading interventions for struggling students, researchers initially documented significant gains in reading scores following their treatments (Streiner, 2001). However, further analysis later showed that the gains were mainly attributable to RTM, as students selected on the basis of low initial scores were likely to improve mainly as a result of statistical probability, not as an effect of the treatments.

In another study that investigated behavioral interventions in elementary schools, researchers found that students initially identified for disruptive behaviors showed significant improvements over time (Marsden & Torgerson, 2012). However, as Streiner (2001) pointed out, most of these improvements were subsequently attributed to RTM—upon repeated measurements, those extreme or disruptive behaviors naturally regressed toward the average. Building on this understanding, Illenberger et al. (2019) explored how RTM can similarly bias findings in policy evaluations when using synthetic controls (a method that estimates treatment effects by creating a weighted combination of control units to approximate an untreated counterfactual). Through simulations in a Difference-in-Differences (DID) framework, the researchers compared synthetic controls to nearest-neighbor matching and found that synthetic controls exacerbate RTM bias. This, in turn, increases the likelihood of falsely detecting treatment effects when no actual effects exist. Indeed, the synthetic control method aggregates information from all control units, which reduces estimator variance but, at the same time, increases confidence in biased outcomes. As a result, Type I error rates for synthetic controls could be nearly twice as high as those of other methods under certain conditions.

To further analyze this effect, Illenberger et al. (2019) used permutation tests to evaluate the null hypothesis of no treatment effect. By systematically relabeling treated and control units in their dataset, the authors demonstrated how RTM bias can lead to spurious results. This analysis highlights the trade-off between improved precision and increased bias in synthetic controls and underscores the need for adjustments to ensure reliable inferences in the presence of RTM. Illenberger et al.'s (2019) findings align with those from behavioral intervention studies, highlighting the widespread impact of RTM across research contexts. These cases and many others illustrate how easily RTM can be mistaken for a genuine treatment effect. But as Yu and Chen (2015) have observed, this RTM problem is more prevalent in studies with pretest-posttest or longitudinal designs, perhaps because in such studies, researchers routinely select participants exhibiting extreme baseline scores to evaluate the impact of interventions so that for example, students with extremely low test scores could be assigned to a remedial program. The point is that without reliable controls, any improvements in repeated assessments may be attributed to interventions rather than RTM (Marsden & Torgerson, 2012). Another way of explaining this phenomenon is that in pretest-posttest or longitudinal research studies, repeated measurements are subject to internal variability and measurement errors. And given that extreme scores are influenced by such errors, subsequent measurements are likely to be less extreme as the internal variability and errors level off. The problem is that whenever RTM is unaccounted for, natural fluctuations can be treated as measures of true change (Yu & Chen, 2015).

Measuring the RTM Effect

To see how weighty RTM can be, consider a hypothetical longitudinal study involving a population of 100 students with these attributes: Mean test score is 75 (0–100 scale); standard deviation (how spread the scores are across all students) is 10; and the within-student standard deviation σ_w (the random fluctuations in scores due to test-taking conditions) is 6. Consider also that *between-students standard deviation* σ_b is derived from the total variance equation $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$, where σ_t^2 is the total variance, σ_w^2 is the within-student variance *(variability in test scores for the same student) over repeated tests, and* σ_b^2 is *the between-students variance or the true* differences in students' abilities and levels of preparation. Given that $\sigma_t = 10$ (the total standard deviation) and $\sigma_w = 6$ (or within-student standard deviation), $\sigma_b^2 = \sigma_t^2 - \sigma_w^2 = 10^2 - 6^2 = 100 - 36 = 64$, which makes the between-students standard deviation (σ_b) = $\sqrt{64}$ = 8. This value suggests that the true difference in ability or preparation among the students contributes *eight variability points* to the overall score distribution. Using the same student population, let us further assume that (a) all the students' test scores are normally distributed, (b) only students scoring below 60 are selected for some remedial programs, (c) only six students scored below 60, (d) the mean initial score of these six students is 55.05, and (e) their mean follow-up or post-intervention score is 60.30, making the mean change in test scores (follow-up scores minus initial scores) equal to +5.25. If this change suggests significant performance improvement, the question is whether it could also be an artifact of RTM.

The way to find out is to use the expected RTM formula: RTM = $\sigma_w \cdot (1 - \rho) \cdot C(z)$, where σ_w is the within-student standard deviation that reflects how much a student's test scores change as a result of random factors (test anxiety, etc.); ρ is the correlation between the initial and follow-up test scores and also reflects how reliably scores predict each other depending on the ratio of true student ability difference to the total variation in the population; and $C(z)$ is a scaling factor reflecting the extremeness of the selection cutoff point—that is, $C(z)$ increases as the cutoff (*z* score) becomes much more extreme or moves closer to the tails of the distribution. In the case of the hypothetical student population we are working with, $C(z)$ is determined using the statistical table to find the *z score of 1.5* (cutoff point of 60, mean of 75, standard deviation of 10), so that the value of $C(z)$ is *1.85.* We can calculate the expected RTM using these givens: $\sigma_w = 6$, $\rho = 0.64$ (σ_b^2/σ_t^2) = $8^2/10^2 = 0.64$, and $C(z) = 1.85$ (from the statistical tables), by plugging those givens in the RTM formula: $\sigma_w \cdot (1 - \rho) \cdot C(z) = 6 \times (1 - 0.64) \times 1.85 = 6 \times 0.36 \times 1.85 = 4.19$. Comparing the expected RTM effect of 4.19 points to the 5.25 mean change in test scores calculated earlier, we find that the expected RTM effect is 79.81% (4.19 divided by 5.25) of the improvement. Put differently, the expected RTM effect accounts for 79.81% (≈80%) of the mean change in this case.

Real-World Cases

In one of the cases referenced earlier, Marsden and Torgerson (2012) reviewed a dataset compiled during a previous pretest–posttest study. To see if RTM affected the change scores of the students involved, the authors plotted those scores against pretest scores. If RTM was present, a negative correlation would emerge because on average students with high pretest scores would make smaller gains than students with lower pretest scores. Figure 1 shows that there was "a strong negative correlation (−0.65, $p < .001$) between the pre-test and change scores" (Marsden & Torgerson, 2012, p. 586).

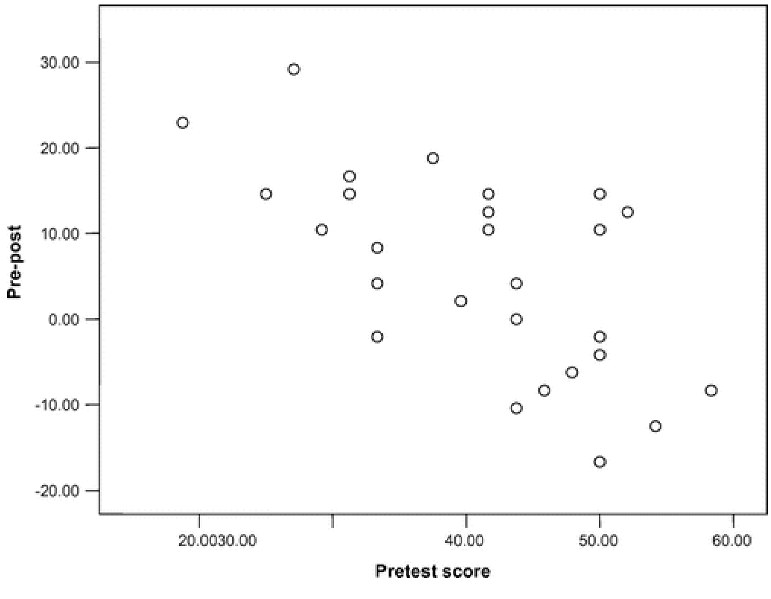

Figure 1. Pretest Scores Plotted Against Gains Between Pre- and Posttests
Note. Data analyzed were adapted from Marsden and Torgerson (2012)

While studying the effects of RTM, Marsden and Torgerson (2012) compared 16 pre- and posttest gains across lower and upper quartile groups were compared based on pretest scores in four specific skill areas: listening, speaking, reading, and writing. Of the 16 comparisons (i.e., pre-to-post and pre-to-delayed posttests in two groups with four outcome measures), six demonstrated statistically significant improvement for the lower group, and four were statistically marginal. But in nearly all other cases, the participants in the lower group demonstrated much greater improvement than their peers in the upper group. These results further show how the effects of RTM can create the impression of bigger gains for underperforming students and produce serious misinterpretations (Smith, 1997). The bottom line is that in virtually all educational assessments, a certain portion of the variation in scores arises from random error. But this error is more pronounced at the extremes of the distribution, where test scores are farther from the average, and so when students are subsequently retested, the scores at the distribution tails are more likely to move toward the mean than the scores closer to the center. And even though not every student's scores will regress to the mean, the majority will, causing the average scores at the extremes to converge toward the overall sample mean in subsequent tests (Marsden & Torgerson, 2012). Marsden and Torgerson's review provides empirical evidence that the effects of RTM are more problematic in studies involving students whose baseline or preintervention scores were either very low or very high.

In another pretest-posttest study, Nielsen et al. (2007) looked at variations in the learning styles of two cohorts of Danish college students at the start of the academic year (t_1) and again a year later (t_2). To document the students' initial learning style scores, the authors used 14 learning styles to place each student in a low or high starting category. Anticipating that the effect of RTM could distort their findings, Nielsen et al. made appropriate corrections before analyzing the changes in the data they collected. The degree to which the effects of RTM would have distorted the outcomes of this study are presented in Table 1, where the mean changes in each learning style from t_1 to t_2 are shown with and without correction for RTM effect. This table shows that among students whose initial levels of learning styles were low, the corrected results show minimal changes except in the "internal" style, for which a significant increase was noted. The corrected results show slight reductions in learning style scores over time. In contrast, the uncorrected mean changes for the same low-starting students incorrectly show that all learning styles increased, with more than 50% of the increases appearing significant. This contrast is yet another clue that RTM can lead researchers to mistake statistical artifacts for true changes.

Table 1. Mean Differences in Learning Style Scores (t_2 - t_1) Corrected and Uncorrected for RTM Effect

Learning style	Level of specific learning styles at t_1			
	Low		High	
	Corrected (real change)	Uncorrected (false change)	Corrected (real change)	Uncorrected (false change)
Legislative	−0.76	3.14^{*}	−0.18	-1.66^{*}
Executive	0.98	3.63^{**}	−0.30	-2.48^{**}
Judicial	−1.34	1.15	-2.32^{**}	-3.59^{**}
Monarchic	0.38	2.62^{**}	0.19	-1.76^{*}
Hierarchic	−1.16	1.44	−0.98	-1.61^{**}
Oligarchic	−0.87	1.05	0.64	2.10
Anarchic	−0.64	0.50	−0.04	-2.61^{*}
Democratic	−0.02	2.53^{**}	0.90	−0.61
Global	−0.07	2.56^{*}	−0.60	-3.69^{**}
Local	0.10	3.16^{**}	−0.75	-3.07^{**}
Internal	-1.54^{*a}	0.46	−0.74	-2.52^{**}
External	0.44	2.07^{**}	−0.31	-1.45^{**}
Liberal	−1.18	0.52	0.57	−2.05
Conservative	1.30	3.00^{**}	0.35	-2.28^{*}

Note. Adapted from Nielsen et al. (2007)

Controlling the RTM Effect

According to Nielsen et al. (2007) and Marsden and Torgerson (2012), the RTM effect can be accounted for during study design or data analysis. The authors suggested that educational researchers can improve their study design by randomly assigning participants to control and experimental groups or by relying on multiple baseline scores. Random assignments to comparison groups expose students to the same RTM effects and ensure that any natural regression or statistical noise is distributed across both groups, canceling out the RTM effect when comparing pre- and post-test scores and making it easy to attribute differences in outcome to a given treatment. Another way to control the RTM effect during the design of a study is to take multiple baseline measurements so that instead of using one pretest score, the researcher will average two or more baseline scores to approximate the students' "true" starting points. This step matters because when extreme scores are influenced by random fluctuations, a single measurement can overstate or understate a student's ability due to random factors, such as test-day anxiety and occasional distractions (Marsden & Torgerson, 2012; Nielsen et al., 2007).

The third RTM control or mitigation strategy pertains to the study participant selection process itself. According to Marsden and Torgerson (2012), selecting students only on the basis of extreme scores is bound to produce RTM bias, because any scores at the extremes are more likely to move closer to the mean upon retesting, irrespective of intervention. One way to avoid this problem is to use an inclusive selection criterion. For example, rather than selecting only low-scoring students, researchers could select a mix of low-, medium-, and high-scoring students, or set a less extreme cutoff for the treatment group by selecting students who fall in the lowest 50% quartile in lieu of selecting them only from the lowest 25% quartile (Morton & Torgerson, 2005).

To control the effect of RTM during data analysis, several mathematical models have been proposed. According to Nielsen et al. (2007), educational researchers doing a pretest-posttest or longitudinal study can use this model to correct posttest scores: RTM effect = $(1 - \rho) \cdot X_{pre} + \epsilon$, where ρ is the correlation between pretest and posttest scores, X_{pre} is the pretest score of one student and ϵ accounts for statistical noise. This model enables researchers to adjust observed post-test scores and mitigate biases arising from RTM. Another way of controlling the effects of RTM during data analysis is to use analysis of covariance (ANCOVA) to adjust posttest scores according to pretest scores (Marsden & Torgerson, 2012). The following ANCOVA model estimates treatment effects while accounting for baseline differences: $Y_{post} = \alpha + \beta \cdot (X_{pre} - \overline{X}_{pre}) + \gamma \cdot \text{Group} + \epsilon$, where Y_{post} represents the *posttest score* for an individual participant, α is the baseline value before any adjustments/contributions from other variables, β is the weight of the relationship between the predictor variable X_{pre} and the dependent variable Y_{post}, $\overline{X}_{pre}$ is the *mean pretest score* for the entire population being analyzed (it centers the baseline scores and ensures that adjustments account for group-level variations), γ is the *coefficient for the group variable* (it quantifies the difference in posttest outcomes between groups after accounting for such factors as baseline differences), and Group is a categorical variable that shows the group assignment of each participant (i.e., Group 1 vs. Group 2). The function of this last term of the ANCOVA model is to isolate the effect of group membership (i.e., treatment vs. control) on Y_{post}, the posttest score.

Practical Applications

All the techniques or approaches reviewed thus far can be applied in various educational contexts. But using the dataset arising from the hypothetical 100 student population we have been working with, let us now apply Marsden and Torgerson's (2012) and Nielsen et al.'s (2007) recommendation. Table 2 provides a snapshot of that hypothetical dataset.

Table 2. Summary Dataset

Attribute	Value	Description
Population size	100	Total number of students
Population mean (μ)	75	Average test score of the population
Total standard deviation (σ_t)	10	Total variability in the population's test scores
Within-student standard deviation (σ_w)	6	Random fluctuations in scores due to test-taking conditions
Between-students variance (σ_b^2)	64	Derived from total variance equation $\sigma_t^2 = \sigma_w^2 + \sigma_b^2$
Between-students standard deviation (σ_b)	8	Square root of between-students variance
Cutoff for selection	Scores below 60 (z = 1.5)	Cutoff point for selecting students for the intervention
Selected students	6	Number of students scoring below the cutoff (60)
Pretest mean of selected students (X_{pre})	55.05	Average pretest score of the six selected students
Posttest mean of selected students (Y_{post})	60.30	Average posttest score of the six selected students
Observed improvement	5.25	Difference between posttest and pretest means
Correlation between pretest and posttest (ρ)	0.64	Ratio of true student ability differences to total variation
RTM scaling factor $C(z)$	1.85	Factor reflecting extremeness of the selection cutoff

Approach 1: Multiple Baseline Measurements

Marsden and Torgerson's (2012) recommendation can be applied in four steps. First, each student's pretest score is measured at least twice, introducing X_{pre1} and X_{pre2}, both drawn from N (55.05, 6). Second, calculate mean pretest scores for every student. For example, for a student whose X_{pre1} and X_{pre2} are 53 and 57, the mean pretest score $X_{pre\text{-}mean}$ is $\frac{\text{Xpre1} + \text{Xpre2}}{2} = \frac{53+57}{2} = 55$. Third, using the formula for the variability of the mean ($\sigma_{mean} = \frac{\sigma_w}{\sqrt{n}}$), where n is the number of baseline measures, calculate $\sigma_{mean} = \frac{6}{\sqrt{2}} \approx 4.24$. Fourth, use the RTM effect formula: $\sigma_w \cdot (1 - \rho) \cdot C(z)$ to recalculate the RTM effect as follows: $4.24 \cdot (1 - 0.64) \cdot 1.85 = 4.24 \cdot 0.36 \cdot 1.85 = 2.83$. Prior to adjustment, σ_w was 6 (see Table 2), so the RTM effect was $6 \cdot (1 - 0.64) \cdot 1.85 = 6 \cdot 0.36 \cdot 1.85 = 4.19$, which means that the RTM effect decreased from 4.19 to 2.83.

Approach 2: Post-Test Score Adjustments

Building on the outcomes of the first approach, this RTM mitigation process consists in applying Nielsen et al.'s (2007) correction formula: $Y_{\text{corrected}} = Y_{\text{post}}$ – RTM effect, where $Y_{\text{corrected}}$ is the posttest score after mathematically adjusting for RTM. As previously noted, without taking at least two baselines, σ_w was 6, the RTM effect was 4.19, and Y_{post} was 60.30. In plugging these values in the correction formula, we find that $Y_{\text{corrected}} = 60.30 - 4.19 = 56.11$. However, after collecting two baseline scores, σ_w and the RTM effect decreased to 4.24 and 2.83. And since $Y_{\text{post}} = 60.30$, $Y_{\text{corrected}} = 60.30 - 2.83 = 57.47$. The outcome of Nielsen et al.'s approach is captured in Table 3.

Table 3. Comparative Analysis of Improvements

Metric	Before correction	After correction
Pretest mean ($X_{\text{pre-mean}}$)	55.05	55.05
Posttest mean (Y_{post})	60.30	57.47
Improvement without correction ($Y_{\text{post}} - X_{\text{pre-mean}}$)	5.25	-
Corrected posttest mean ($Y_{\text{corrected}}$)	-	57.47
Improvement after correction ($Y_{\text{corrected}} - X_{\text{pre-mean}}$)	-	2.42

Applying Marsden and Torgerson's (2012) multiple baseline adjustments and Nielsen et al.'s (2007) formulaic corrections to our hypothetical 100-student population, we find that whereas the first approach is intended for use during study design, the second is designed for data analysis. The reason is that while Marsden and Torgerson's approach controls RTM during study *design* by reducing variability and enhancing baseline estimates, Nielsen et al.'s recommended approach deals with RTM during *data analysis* by removing its effects from observed results. In other words, the two approaches are complementary and provide educational researchers with the practical tools they need to control the effects of RTM both before and after data collection. But even with such a grasp of RTM and the processes by which to control its distorting effect, Nielsen et al. (2007) and Marsden and Torgerson's (2012) approaches carry two major flaws. The first is about how baseline scores are collected (Ledford & Gast, 2024). In concurrent multiple baseline designs, data collection is done simultaneously, and treatments are staggered over time (Kazdin, 2020), providing some experimental control. However, the method is not effective enough for reducing the RTM effect to its smallest expression because natural changes in test scores or human behavior may still appear as true treatment effects (Slocum et al., 2022). Conversely, nonconcurrent multiple baseline designs stagger data collection over time. But this method poses another problem, as overlapping baseline information and the intrusion of external factors make the tedium of isolating true treatment effects even harder (Kennedy, 2022).

The second flaw lies in the nature of formulaic corrections, which rely on formulas to enhance the quality of observations by adjusting for RTM during data analysis. Mathematically sound as this approach seems, it works on the implicit assumption that the relationship between pretest and posttest scores is linear and stable across samples, even though the noisy, often nonlinear patterns characteristic of real-world data seldom conform to formulas. Moreover, the consistency of formulaic corrections depends heavily on accurate estimates of parameters such as standard deviations, correlation coefficients, and scaling factors, all of which are potential sources of error. The implication is that while formulaic corrections and multiple baseline adjustments are useful, neither is the optimal solution for the RTM problem.

A Bayesian Model for Controlling RTM

If the flaws identified in the preceding section seem intricate, we can sort them out using Bayesian regression (Drugowitsch, 2013; Mara, 2019). We can indeed use this hybrid Bayesian model to better control RTM: $Y_{\text{adjusted}} = X_{\text{pre-mean}} + w\,(Y_{\text{post}} - X_{\text{pre-mean}}) + \gamma$ (RTM effect), where $X_{\text{pre-mean}}$ is the stabilized pretest mean score obtained by averaging multiple baseline scores for a student or participant, Y_{post} is the observed posttest score after treatment, and w is the weight assigned to the observed changes (posttest minus pretest mean). Irrespective of the weight the model gives to the observed data versus prior knowledge, w will typically lie between 0 and 1, and γ is the weight given to the prior RTM effect. This weight adjusts how much of the RTM effect is factored in the correction based on the strength of that prior knowledge, and the RTM effect is derived from the formula: $\sigma_w\,(1 - \rho) \cdot C(z)$, where σ_w is the within-student variability, ρ is the correlation between pretest and posttest scores, and $C(z)$ is the scaling factor we saw earlier.

This new model is hybrid for several reasons. First, instead of using a single pretest score, it averages multiple baseline measurements and, therefore, decreases the variability (σ_w) in the data, which in turn drives the RTM effect downward. Second, the model uses prior knowledge, such as known population means and variances, to adjust post-test scores. Third, the model uses the RTM equation to calculate how much of the observed change is attributable to RTM. Fourth, the weights (w and γ) are leveraged to balance the contributions of the observed data and the prior RTM effect. For example, if $w = 1$ on a scale of 0 to 1, this hybrid Bayesian model relies fully on the observed data (posttest scores minus pretest scores) and disregards all adjustments for prior RTM knowledge. But if $w = 0$, the model disregards the observed data and relies entirely on the prior RTM effect. Likewise, if $\gamma = 1$, the model assumes that the prior RTM knowledge perfectly explains the adjustment and applies it. And if $\gamma = 0$, the model ignores prior RTM knowledge and makes no corrections based on it. But the strength of this hybrid Bayesian model lies also in the fact that these weights are proportional and complementary, so that researchers can tweak them according to the reliability of the observed data or based on how much credence they give to the prior RTM estimates available. For example, whereas a w of 0.9 means that

90% of the observed improvement (the difference between pretest and posttest scores) is trusted as valid and used in the adjustment, a γ of 0.7 means that 70% of the estimated prior RTM effect (according to prior knowledge or assumptions about RTM in similar datasets) is factored into the adjustment.

Returning to the same hypothetical population parameters we worked with earlier: *Pretest mean* ($X_{\text{pre-mean}}$) = 55.05, *posttest mean* (Y_{post}) = 60.30, and *observed improvement* ($Y_{\text{post}} - X_{\text{pre-mean}}$) = 5.25, and assuming (*w)* = 0.9 and (γ) = 0.7, here is how the model works. It uses weighted contributions to integrate observed data and prior RTM information, so that the adjusted posttest score is calculated as follows: $Y_{\text{adjusted}} = X_{\text{pre-mean}} + w\,(Y_{\text{post}} - X_{\text{pre-mean}}) + \gamma$ (RTM effect), that is: $Y_{\text{adjusted}} = 55.05 + 0.9\,(60.30 - 55.05) + 0.7\,(4.19) = 55.05 + 4.73 + 2.93 = 62.71$. means that the true *improvement in this example is* $Y_{\text{adjusted}} - X_{\text{pre-mean}} = 62.71 - 55.05 = 7.66$. All told, this hybrid Bayesian model (a) adjusts the RTM weight based on known or prior data, (b) includes both observed data and prior RTM knowledge and (c) produces a more reliable adjusted score that reflects empirical observations and theoretical expectations. But as with the other methods that Nielsen et al. (2007) and Marsden and Torgerson (2012) have recommended, this innovative model has its own limitations. It is computationally complex and requires more advanced statistical analyses than traditional techniques such as multiple baseline adjustments or formulaic corrections. It also requires prior data, such as σ_w (i.e., known or previously documented variability or within-student standard of deviation), as well as ρ (i.e., known or previously documented correlations between the pretest and posttest scores of comparable student groups).

Conclusion

In attempting to assemble the puzzle of RTM, this paper has placed into sharp relief the ways in which RTM distorts data interpretation in educational research (Cochrane et al., 2020). RTM often arises in normal distributions, where extreme values are unlikely to reappear because, in subsequent observations, they regress toward the mean. Yet, in some cases, RTM effects are the result of confounding variables or external factors that are completely unrelated to interventions but can distort the interpretation of changes, especially in pretest-posttest and longitudinal studies. More broadly, RTM can lead to false causal inferences.

A textbook example is the illusion that criticism is far more productive than praise. As Pinker (2021) has observed, teachers often attribute a struggling student's improvement after a bad grade or harsh criticism to the efficacy of these practices, even if the rebound is statistically predictable. Similarly, an extraordinary performance followed by a dip in subsequent attempts is often treated as evidence that praise is counterproductive. These common misinterpretations and many others justify the need for more comprehensive statistical methods with which educational researchers can disentangle true treatment or intervention effects from those driven by RTM. In comparing the methods recommended by Marsden and Torgerson (2012) and Nielsen et al. (2007) to the hybrid Bayesian model introduced in this paper, certain important distinctions emerged. The multiple baselines method stabilizes pretest scores through averaging but may yield conservative estimates as it does not fully account for residual RTM effects. Although useful, Nielsen et al.'s (2007) formulaic corrections depend on the precision of input parameters, such as σ_w (within-student variability) and ρ (the correlation coefficient). In contrast, the hybrid Bayesian model dynamically balances observed data with prior knowledge about RTM to ensure that corrections do not overestimate or underestimate the true effect of various treatments (Mara, 2019; Murphy, 2012). It follows that this paper advances the discourse on RTM by incorporating Bayesian principles into educational data analysis. The model's versatility extends well beyond educational research, finding applications in various fields where RTM is a persistent concern. For example, in clinical trials, Bayesian methods can help distinguish true treatment effects from fluctuations in patient outcomes due to RTM, ensuring more accurate assessments of medical interventions. Environmental researchers can also rely on the model to disentangle RTM effects from true changes in ecological or climate datasets. Whether it is applied to large research studies or small classroom experiments, this hybrid Bayesian model optimally decreases the risks of taking RTM artifacts for true treatment effects. Even so, future research could focus on empirically testing the model in different educational conditions to evaluate its adaptability. Additional studies could also explore the model's integration with advanced machine learning techniques to automate the calibration of prior parameters and further enhance its precision. Finally, expanding the model's application to datasets in other spheres of academic research would provide great insights into its broader applications.

References

Asbury, C. A. (1974). Selected factors influencing over- and under-achievement in young school-age children. *Review of Educational Research*, *44*(4), 409-428. https://doi.org/10.3102/00346543044004409

Cochrane, K. M., Williams, B. A., Fischer, J. A., Samson, K. L., Pei, L. X., & Karakochuk, C. D. (2020). Regression to the mean: A statistical phenomenon of worthy consideration in anemia research. *Current Developments in Nutrition*, *4*(10), Article nzaa152. https://doi.org/10.1093/cdn/nzaa152

Drugowitsch, J. (2013). *Variational Bayesian inference for linear and logistic regression*. arXiv. https://doi.org/10.48550/ARXIV.1310.5438

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246-263. https://doi.org/10.2307/2841583

Illenberger, N., Small, D. S., & Shaw, P. A. (2019). *Regression to the Mean's Impact on the Synthetic Control Method: Bias and Sensitivity Analysis* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1909.04706

Isaac, S., & Michael, W. B. (1995). *Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences* (3rd ed.). EdITS. https://psycnet.apa.org/record/1995-98981-000

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251. https://doi.org/10.1037/h0034747

Kazdin, A. E. (2020). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kennedy, C. H. (2022). The nonconcurrent multiple-baseline design: It is what it is and not something else. *Perspectives on Behavior Science*, *45*, 647-650. https://doi.org/10.1007/s40614-022-00343-0

Ledford, J. R., & Gast, D. L. (Eds.). (2024). *Single case research methodology: Applications in special education and behavioral sciences* (4th ed.). Routledge.

Linden, A. (2013). Assessing regression to the mean effects in health care initiatives. *BMC Medical Research Methodology*, *13*, Article 119. https://doi.org/10.1186/1471-2288-13-119

Mara, T. A. (2019). *Linear regression in the Bayesian framework* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1908.03329

Marsden, E., & Torgerson, C. J. (2012). Single group, pre- and post-test research designs: Some methodological concerns. *Oxford Review of Education*, *38*(5), 583-616. https://doi.org/10.1080/03054985.2012.731208

Morton, V., & Torgerson, D. J. (2005). Regression to the mean: Treatment effect without the intervention. *Journal of Evaluation in Clinical Practice*, *11*(1), 59-65. https://doi.org/10.1111/j.1365-2753.2004.00505.x

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*(3), 622-637. https://doi.org/10.1037/0033-2909.88.3.622

Nielsen, T., Karpatschof, B., & Kreiner, S. (2007). Regression to the mean effect: When to be concerned and how to correct for it. *Nordic Psychology*, *59*(3), 231-250. https://doi.org/10.1027/1901-2276.59.3.231

Pinker, S. (2021). *Rationality: What it is, why it seems scarce, why it matters*. Viking.

Slocum, T. A., Joslyn, P. R., Nichols, B., & Pinkelman, S. E. (2022). Revisiting an analysis of threats to internal validity in multiple baseline designs. *Perspectives on Behavior Science*, *45*, 681-694. https://doi.org/10.1007/s40614-022-00351-0

Smith, G. (1997). Do statistics test scores regress toward the mean? *CHANCE*, *10*(4), 42-45. https://doi.org/10.1080/09332480.1997.10542064

Smith, G., & Smith, J. (2005). Regression to the mean in average test scores. *Educational Assessment*, *10*(4), 377-399. https://doi.org/10.1207/s15326977ea1004_4

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, *6*(2), 103-114. https://doi.org/10.1177/096228029700600202

Streiner, D. V. L. (2001). Regression toward the mean: Its etiology, diagnosis, and treatment. *The Canadian Journal of Psychiatry*, *46*(1), 72-76. https://doi.org/10.1177/070674370104600111

Thorndike, R. L. (1942). Regression fallacies in the matched groups experiment. *Psychometrika*, *7*(2), 85–102. https://doi.org/10.1007/bf02288069

Yu, R., & Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, *5*, Article 01574. https://doi.org/10.3389/fpsyg.2014.01574