Frontiers of Education Development and Innovation 教育发展与创新前沿, 2026, 4(4), 7-14
Published Online August 2026.
https://ufascience.com/fedi

均值回归效应在教育干预研究中的识别与校正：一种混合贝叶斯模型

郑偌洁

四川大学商学院，成都 610065，四川，中国

摘要：均值回归在数据分析中普遍存在，教育研究者却常把它误当作真实变化的证据，把随机波动错误地归因于处理效应。作为一种极端值在重复施测后自然向平均值靠拢的统计现象，均值回归在采用前测—后测设计或纵向设计的教育研究中尤其容易被误读：观察到的变化即便只是统计假象，也往往被认定为处理的效果。本文借助一个假想案例与两项真实研究，考察均值回归带来的技术难题，并提出一种混合贝叶斯模型；与多重基线校正、公式化校正等常规做法相比，该模型能更有效地削弱均值回归的影响。具体而言，该混合贝叶斯模型依靠多次基线测量把前测阶段与均值回归相关的失真降到最低，并利用标准差、总体均值等先验知识来细化后测数据的校正。由此，该模型为教育研究者准确评估干预效果、提升各类研究驱动的教育政策与实践的有效性，提供了一件新的工具。

关键词：贝叶斯回归；因果推断；前测—后测设计；均值回归

Identifying and Correcting Regression to the Mean in Educational Intervention Research: A Hybrid Bayesian Model

RuoJie Zheng

Business School, Sichuan University, Chengdu 610065, Sichuan, China

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

文章引用: 郑偌洁. 均值回归效应在教育干预研究中的识别与校正：一种混合贝叶斯模型[J]. 教育发展与创新前沿, 2026, 4(4): 7-14.
DOI: https://doi.org/10.61784/fedi35

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

一、引言

均值回归（regression to the mean, RTM）是一种纯粹的统计现象，它对数据集有着普遍的影响，但其中许多含义被广泛误解[1]。每当数据集中的极端值在重复施测后向平均值移动，研究者往往把这种移动误读为因果效应的证据，而没有认识到它们只是统计假象[2-4]。在教育研究中，这种把随机变化错误归因于处理效应的倾向不仅有碍于教师效能的评估，也有碍于教学方法与课程的评价。换言之，未能识别 RTM 会导致对某些干预、教学法以及课程改革的效果作出不准确的论断。在依赖前测—后测设计或纵向设计的学术研究中，这一问题尤为突出[5-6]。教育研究中的问题还因大量依赖标准化测验分数等本身就具有内在波动性的结果测量而更趋复杂，研究者往往因此忽略 RTM，或把随机波动当成真实的成绩变化。例如，Smith 与 Smith[4] 报告了若干案例：研究者把测验分数的提高归功于教学干预，而其中许多变化其实源自统计上的均值回归。Thorndike[7] 给出了另一个诠释失当的例子：在一项研究中，标准化测验分数高而平均学分绩点（GPA）低的大学生被视为「低成就者」，而 GPA 高但测验分数低的同学则一律被视为「超常成就者」。与其他案例一样，这些解释并未考虑 RTM 足以说明上述差异的可能性。此外，心理学家 Daniel Kahneman 与 Amos Tversky 强调，RTM 不仅扭曲数据解释，还会因忽视统计上的细微之处而导致误入歧途的政策决定。在一篇被广泛引用的论文中，Kahneman 与 Tversky[6] 描述了一群飞行学校教官推行正强化训练政策的经过。依照学校心理学家的建议，教官们对成功完成复杂动作的学员一贯予以表扬。然而他们后来发现，出色的表现之后往往跟着可度量的成绩下滑。教官们把这一模式误读为正强化无效的证据，索性彻底放弃了该政策。他们没有认识到的是，这些变化并非因果效应，而是 RTM 的可预期表现[6]。

本文有两个目的。其一，讨论 RTM 如何潜入以学生群体为对象的教育研究，以及在研究设计或数据分析阶段通常如何处理它。其二，借助一个假想数据集与两个实证案例来具体呈现 RTM，并提出一种能更有效削弱其失真效应的混合贝叶斯模型。全文的基本论点是：尽管 RTM 在教育研究中普遍存在，它并非无法对付。

二、理论框架

RTM 是一种反直觉的统计现象，最早由博学家高尔顿（Francis Galton，1822—1911）在两项重要实验中发现。第一项实验中，Galton[8] 通过种植不同大小的香豌豆种子来研究 RTM，按直径把种子分为七类，随后记录每株植物结出的子代种子，并计算每类各 100 粒种子的平均直径。结果显示出一种清晰的模式：最小的种子倾向于产出略大的子代，最大的种子则产出略小的子代，使得子代种子的大小向这 100 粒种子的平均直径回归。Galton 把这一模式称为均值回归，这个术语后来催生了现代统计学中的回归概念。第二项实验中，Galton 把对同一现象的研究扩展到人的身高，查阅了 205 对父母及其 928 名成年子女的家庭记录。由于男性平均身高比女性高 8%，Galton 在与男性身高比较之前，先把女性身高乘以 1.08 的系数加以换算。他随后取每位母亲与父亲身高的平均值，算出「中亲身高」，再把全部中亲身高分为九类，计算每一类中子女身高的中位数。

Galton[8] 的发现再次印证了 RTM：身高高于或低于平均水平的父母，其子女的身高向总体平均值靠拢。具体而言，父母身高每偏离均值 1 英寸，子女身高更可能只朝同一方向偏离 0.69 英寸。这些开创性研究以及其后大量研究结果的一致性，使 RTM 成为一个真实的统计问题。不过正如 Pinker[1] 所指出的，RTM 不仅适用于父母与子女的身高和智商，更普遍地适用于任何两个并非完全相关的变量：一个变量上的极端值，往往对

应着另一个变量上不那么极端的值。当然，这并不意味着最终所有孩子都会长成平均身高，也不意味着人口的智商会收敛到 100。它的含义是：每当一个极端值出现在钟形分布中，与之配对的任何其他变量都不大可能复制它的离群地位。问题在于，在为数过多的研究中，RTM 被误当成了变化的证据[9-12]。例如，在一项考察针对学困生的阅读干预的研究中，研究者最初记录到干预后阅读分数有显著提升；然而进一步分析表明，这些提升主要归因于 RTM——依据初始分数偏低而被选入的学生，其成绩提高主要是统计概率使然，而非干预的效果。在另一项考察小学行为干预的研究中，研究者发现最初被认定有扰乱行为的学生随时间推移出现了显著改善。然而正如 Streiner 所指出的，这些改善大多随后被归因于 RTM——在重复测量中，那些极端的或扰乱性的行为自然会向平均水平回归。在此基础上，Illenberger 等 进一步探讨了在使用合成控制法（一种通过对控制单元加权组合来逼近未处理反事实、从而估计处理效应的方法）进行政策评估时，RTM 如何同样使结论产生偏倚。通过在双重差分（DID）框架下的模拟，研究者把合成控制法与最近邻匹配作了比较，发现合成控制法会放大 RTM 偏倚，从而在实际并不存在效应时提高错误检出处理效应的可能性。究其原因，合成控制法汇聚了全部控制单元的信息，虽降低了估计量的方差，却同时增强了对有偏结果的信心。其结果是，在特定条件下合成控制法的第一类错误率可能接近其他方法的两倍。

三、RTM 效应的测度

为看清 RTM 的分量有多重，设想一项假想的纵向研究，其学生总体为 100 人，具有如下属性：平均测验分数为 75（0—100 分制）；标准差（分数在全体学生间的离散程度）为 10；学生内标准差 σ_{w}（因施测条件造成的分数随机波动）为 6。再设学生间标准差 σ_{b} 由总方差方程导出：

$$\sigma_{\mathrm{t}}^2 = \sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{b}}^2 \tag{1}$$

式中 σ_{t}^2 为总方差，σ_{w}^2 为学生内方差（同一名学生在重复施测中的分数变异），σ_{b}^2 为学生间方差，即学生在能力与准备程度上的真实差异。已知 $\sigma_{\mathrm{t}} = 10$（总标准差）、$\sigma_{\mathrm{w}} = 6$（学生内标准差），则 $\sigma_{\mathrm{b}}^2 = \sigma_{\mathrm{t}}^2 - \sigma_{\mathrm{w}}^2 = 10^2 - 6^2 = 100 - 36 = 64$，故学生间标准差 $\sigma_{\mathrm{b}} = \sqrt{64} = 8$。这一数值表明，学生在能力或准备程度上的真实差异，为整体分数分布贡献了 8 个变异点。

仍以该学生总体为例，进一步假设：（a）全体学生的测验分数服从正态分布；（b）只有得分低于 60 的学生被选入补救课程；（c）得分低于 60 的学生只有 6 人；（d）这 6 名学生的初测平均分为 55.05；（e）他们的随访即干预后平均分为 60.30，因而测验分数的平均变化（随访分减初测分）为 +5.25。如果这一变化意味着成绩显著提升，那么问题是：它是否也可能是 RTM 的假象？

要弄清这一点，可使用 RTM 期望值公式：

$$\mathrm{RTM} = \sigma_{\mathrm{w}} \cdot (1 - \rho) \cdot C(z) \tag{2}$$

式中 σ_{w} 为学生内标准差，反映一名学生的测验分数因随机因素（考试焦虑等）而产生的变动幅度；ρ 为初测与随访分数之间的相关系数，它也反映分数彼此预测的可靠程度，取决于学生真实能力差异与总体总变异之比；$C(z)$ 为反映选取截断点极端程度的标度因子——截断点（z 分数）越极端、越靠近分布的尾部，$C(z)$ 就越大。就我们所处理的这一假想学生总体而言，$C(z)$ 由统计用表按 z 分数 1.5（截断点 60、均值 75、标准差 10）查得，其值为 1.85。把已知量 $\sigma_{\mathrm{w}} = 6$、$\rho = \frac{\sigma_{\mathrm{b}}^2}{\sigma_{\mathrm{t}}^2} = 8^2/10^2 = 0.64$、$C(z) = 1.85$ 代入 RTM 公式，即得 $6 \times (1 - 0.64) \times 1.85 = 6 \times 0.36 \times 1.85 = 4.19$。把 4.19 分的 RTM 期望效应与前面算得的 5.25 分平均变化相比较，可知 RTM 期望效应占提升幅度的 79.81%（4.19 除以 5.25）。换言之，本例中平均变化的 79.81%（≈80%）可由 RTM 期望效应解释。

四、现实案例

在前文提到的案例之一中，Marsden 与 Torgerson 重新审视了一项既往前测—后测研究所积累的数据集。为查明 RTM 是否影响了所涉学生的变化分数，作者把这些变化分数对前测分数作图。若存在 RTM，就会出现负相关，因为平均而言前测分数高的学生所取得的增益会小于前测分数低的学生。图 1 显示，「前测分数与变化分数之间存在很强的负相关（−0.65，$p < .001$）」（，第 586 页）。

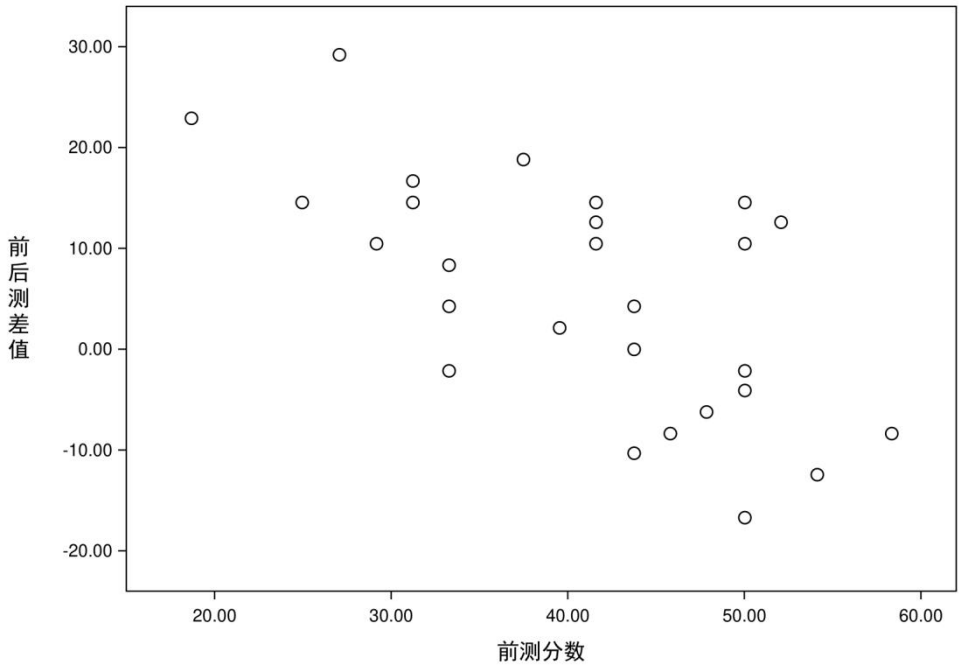

Figure 1 Pretest Scores Plotted Against Gains Between Pre- and Posttests

图 1 前测分数与前后测增益的散点图

在考察 RTM 效应时，Marsden 与 Torgerson 依据前测分数，就听、说、读、写四个具体技能领域，对低分四分位组与高分四分位组的 16 项前后测增益作了比较。在这 16 项比较中（即两组、四项结果测量下的前测至后测与前测至延迟后测），有 6 项在低分组显示出统计上显著的提升，另有 4 项处于统计边缘。但在几乎所有其余情形中，低分组被试所表现出的提升都远大于高分组的同伴。这些结果进一步说明，RTM 效应会造成成绩较差的学生获益更大的假象，并引发严重的误读。归根结底，在几乎所有教育测评中，分数变异中总有一部分来自随机误差；而这种误差在分布的两端更为突出，因为那里的测验分数离平均值更远。于是当学生再次接受测验时，分布尾部的分数比靠近中心的分数更可能向均值移动。尽管并非每一名学生的分数都会回归均值，但多数会，从而使两端的平均分在后续测验中向总体样本均值收敛。Marsden 与 Torgerson 的综述提供了实证依据：在涉及基线或干预前分数极低或极高的学生的研究中，RTM 效应带来的问题更为严重。

在另一项前测—后测研究中，Nielsen 等 考察了两届丹麦大学生在学年之初（t1）及一年之后（t2）学习风格的变化。为记录学生的初始学习风格得分，作者使用 14 种学习风格把每名学生归入低起点或高起点类别。考虑到 RTM 效应可能扭曲研究结论，Nielsen 等在分析所采集数据的变化之前先作了相应校正。RTM 效应本会在多大程度上扭曲该研究结果，见表 1：其中列出了各学习风格从 t1 到 t2 的平均变化，并分别给出对 RTM 效应作校正与未作校正的结果。该表显示，在学习风格初始水平较低的学生中，校正后的结果除「内部型」外变化都很小——「内部型」出现了显著的增加；校正后的结果显示学习风格得分随时间略有下降。与之相对，

同一批低起点学生的未校正平均变化却错误地显示所有学习风格都有所提高，且其中半数以上的提高看似显著。这一反差再次提示：RTM 会使研究者把统计假象误认为真实变化。

Table 1 Mean Differences in Learning Style Scores (t2 − t1) Corrected and Uncorrected for RTM Effect

表 1 学习风格得分的平均差值（t2 − t1）：对 RTM 效应校正与未校正的结果

学习风格	低起点·校正后（真实变化）	低起点·未校正（虚假变化）	高起点·校正后（真实变化）	高起点·未校正（虚假变化）
立法型	−0.76	3.14*	−0.18	−1.66*
执行型	0.98	3.63**	−0.30	−2.48**
司法型	−1.34	1.15	−2.32**	−3.59**
君主型	0.38	2.62**	0.19	−1.76*
等级型	−1.16	1.44	−0.98	−1.61**
寡头型	−0.87	1.05	0.64	2.10
无序型	−0.64	0.50	−0.04	−2.61*
民主型	−0.02	2.53**	0.90	−0.61
整体型	−0.07	2.56*	−0.60	−3.69**
局部型	0.10	3.16**	−0.75	−3.07**
内部型	−1.54*a	0.46	−0.74	−2.52**
外部型	0.44	2.07**	−0.31	−1.45**
自由型	−1.18	0.52	0.57	−2.05
保守型	1.30	3.00**	0.35	−2.28*

五、RTM 效应的控制

依照 Nielsen 等 与 Marsden 和 Torgerson 的看法，RTM 效应可以在研究设计阶段或数据分析阶段加以处理。两文作者建议，教育研究者可以通过把被试随机分配到对照组与实验组，或依靠多个基线分数，来改进研究设计。随机分配到比较组使各组学生暴露于相同的 RTM 效应，保证任何自然回归或统计噪声都分摊到两组之中，从而在比较前后测分数时相互抵消，也使结果差异易于归因于特定处理。在研究设计阶段控制 RTM 效应的另一种做法是进行多次基线测量：研究者不再使用单一的前测分数，而是取两次或更多次基线分数的平均值来逼近学生「真实」的起点。这一步之所以重要，是因为当极端分数受随机波动影响时，单次测量会因考试当天的焦虑、偶发的干扰等随机因素而高估或低估一名学生的能力。

第三种 RTM 控制或缓解策略关乎被试的遴选过程本身。按照 Marsden 与 Torgerson 的说法，仅依据极端分数遴选学生必然产生 RTM 偏倚，因为无论有无干预，处于两端的分数在复测时都更可能向均值靠拢。避免这一问题的办法之一是采用包容性的遴选标准。例如，研究者可以不只选取低分学生，而是选取低分、中等、高分学生的组合；或者设定不那么极端的处理组截断点，改从最低的 50% 四分位段而非仅从最低的 25% 四分位段中选取学生。

为在数据分析阶段控制 RTM 的影响，已有若干数学模型被提出。按照 Nielsen 等 的看法，开展前测—后测研究或纵向研究的教育研究者可用下面的模型校正后测分数：

$$\text{RTM 效应} = (1 - \rho) \cdot X_{\text{pre}} + \epsilon \tag{3}$$

式中 ρ 为前测与后测分数之间的相关系数，X_{pre} 为某一名学生的前测分数，ϵ 用以计入统计噪声。该模

型使研究者得以调整观测到的后测分数，缓解 RTM 带来的偏倚。在数据分析阶段控制 RTM 效应的另一途径，是用协方差分析（ANCOVA）依据前测分数来调整后测分数。下面的 ANCOVA 模型在计入基线差异的同时估计处理效应：

$$Y_{\text{post}} = \alpha + \beta \cdot (X_{\text{pre}} - \bar{X}_{\text{pre}}) + \gamma \cdot \text{Group} + \epsilon \quad (4)$$

式中 Y_{post} 表示单个被试的后测分数；α 是在其他变量作出任何调整或贡献之前的基线值；β 是预测变量 X_{pre} 与因变量 Y_{post} 之间关系的权重；$\bar{X}_{\text{pre}}$ 是所分析总体的前测平均分（它把基线分数中心化，保证调整时计入组间的整体差异）；γ 是组别变量的系数（它量化在计入基线差异等因素之后组间后测结果的差别）；Group 则是标示每名被试组别归属的分类变量（即第 1 组与第 2 组）。ANCOVA 模型最后这一项的作用，是把组别归属（即处理组与对照组）对后测分数 Y_{post} 的效应分离出来。

六、实际应用

迄今回顾的全部技术或方法都可用于各种教育情境。不过，让我们就前面一直使用的 100 名学生的假想总体所产生的数据集，来应用 Marsden 和 Torgerson 与 Nielsen 等 的建议。表 2 给出了该假想数据集的概貌。

Table 2 Summary Dataset

表 2 数据集概要

属性	取值	说明
总体规模	100	学生总人数
总体均值 μ	75	总体的平均测验分数
总标准差 σt	10	总体测验分数的总变异
学生内标准差 σw	6	因施测条件造成的分数随机波动
学生间方差	64	总方差减去学生内方差所得
学生间标准差 σb	8	学生间方差的平方根
遴选截断点	低于 60 分（z = 1.5）	选取学生进入干预的截断点
入选学生数	6	得分低于截断点（60）的学生人数
入选学生前测均分 Xpre	55.05	6 名入选学生的平均前测分数
入选学生后测均分 Ypost	60.30	6 名入选学生的平均后测分数
观测到的提升	5.25	后测均分与前测均分之差
前后测相关系数 ρ	0.64	学生真实能力差异与总变异之比
RTM 标度因子 C(z)	1.85	反映遴选截断点极端程度的因子

（一）方法一：多重基线测量

Marsden 与 Torgerson 的建议可分四步实施。第一步，每名学生的前测分数至少测量两次，得到 X_{pre1} 与 X_{pre2}，二者均取自 N(55.05, 6)。第二步，计算每名学生的前测平均分。例如某名学生的 X_{pre1} 与 X_{pre2} 分别为 53 与 57，则其前测平均分 $X_{\text{pre-mean}} = (53 + 57)/2 = 55$。第三步，用均值变异性的公式

$$\sigma_{\text{mean}} = \frac{\sigma_{\text{w}}}{\sqrt{n}} \quad (5)$$

（式中 n 为基线测量的次数）算得 $\sigma_{\text{mean}} = 6/\sqrt{2} \approx 4.24$。第四步，用式（2）重新计算 RTM 效应：4.24 × (1 − 0.64) × 1.85 = 4.24 × 0.36 × 1.85 = 2.83。调整之前 σ_{w} 为 6（见表 2），故 RTM 效应为 6 × (1 − 0.64) × 1.85 = 6 × 0.36 × 1.85 = 4.19；这意味着 RTM 效应由 4.19 降至 2.83。

DOI: https://doi.org/10.61784/fedi35

（二）方法二：后测分数校正

在第一种方法结果的基础上，这一 RTM 缓解过程在于运用 Nielsen 等 的校正公式：

$$Y_{\text{corrected}} = Y_{\text{post}} - \text{RTM 效应} \quad (6)$$

式中 $Y_{\text{corrected}}$ 为经数学调整以计入 RTM 之后的后测分数。如前所述，在不采集至少两次基线的情况下，σ_{w} 为 6，RTM 效应为 4.19，Y_{post} 为 60.30；把这些值代入校正公式，得 $Y_{\text{corrected}} = 60.30 - 4.19 = 56.11$。而在采集两次基线分数之后，$\sigma_{\text{w}}$ 与 RTM 效应分别降至 4.24 与 2.83；由于 $Y_{\text{post}} = 60.30$，故 $Y_{\text{corrected}} = 60.30 - 2.83 = 57.47$。Nielsen 等所提方法的结果见表 3。

Table 3 Comparative Analysis of Improvements

表 3 提升幅度的对比分析

指标	校正前	校正后
前测均分 Xpre-mean	55.05	55.05
后测均分 Ypost	60.30	57.47
未校正的提升（Ypost − Xpre-mean）	5.25	—
校正后的后测均分 Ycorrected	—	57.47
校正后的提升（Ycorrected − Xpre-mean）	—	2.42

把 Marsden 和 Torgerson 的多重基线校正与 Nielsen 等 的公式化校正应用于我们的 100 名学生假想总体，可以看出：前一种方法用于研究设计阶段，后一种方法则用于数据分析阶段。原因在于，Marsden 与 Torgerson 的方法通过降低变异性、改善基线估计来在研究设计阶段控制 RTM，而 Nielsen 等所建议的方法则通过从观测结果中剔除 RTM 的影响来在数据分析阶段加以处理。换言之，两种方法相互补充，为教育研究者提供了在数据采集前后控制 RTM 失真效应所需的实用工具。然而，即便对 RTM 及其控制流程有了这样的把握，Nielsen 等 与 Marsden 和 Torgerson 的方法仍有两大缺陷。

其一在于基线分数的采集方式。在同期多重基线设计中，数据采集同时进行，而处理则在时间上错开，由此获得一定的实验控制。但该方法在把 RTM 效应压到最小这一点上还不够有力，因为测验分数或人类行为的自然变化仍可能表现为真实的处理效应。反之，非同期多重基线设计把数据采集在时间上错开，但这种做法带来了另一个问题：基线信息的重叠与外部因素的介入，使分离真实处理效应这件本就繁琐的事情更加困难。

其二在于公式化校正的本质——它依赖公式，通过在数据分析阶段针对 RTM 作调整来提高观测质量。这一路径在数学上看似严谨，却隐含地假定前测与后测分数之间的关系是线性的、且在不同样本间保持稳定；然而现实数据所特有的、含噪且往往非线性的模式，很少符合公式的设定。此外，公式化校正的一致性在很大程度上取决于标准差、相关系数、标度因子等参数估计的准确性，而这些都是潜在的误差来源。这意味着，公式化校正与多重基线校正虽有用处，却都不是 RTM 问题的最优解。

七、结论

在试图拼合 RTM 这块拼图的过程中，本文着重揭示了 RTM 在教育研究中扭曲数据解释的种种方式。RTM 常出现于正态分布之中：极端值不大可能再度出现，因为在后续观测中它们会向均值回归。然而在某些情形下，RTM 效应源于与干预完全无关的混杂变量或外部因素，却仍会扭曲对变化的解释，在前测—后测研究与纵向研究中尤其如此。更一般地说，RTM 会导致错误的因果推断。

一个教科书式的例子，是「批评远比表扬有效」这一错觉。正如 Pinker[1] 所观察到的，教师往往把学困生在一次差评或严厉批评之后的进步归功于这些做法的效力，哪怕这种反弹在统计上本就可以预期。同样，一次出色表现之后出现的成绩下滑，也常被当作表扬适得其反的证据。诸如此类的误读，说明确有必要发展更为完备的统计方法，使教育研究者能够把真实的处理或干预效应与 RTM 驱动的效应区分开来。把 Marsden 和 Torgerson 与 Nielsen 等 所建议的方法同本文提出的混合贝叶斯模型加以比较，可以看出几点重要区别。多重基线法通过取平均稳定前测分数，但因未能完全计入残余的 RTM 效应，其估计可能偏于保守。Nielsen 等 的公式化校正虽然有用，却取决于 σ_{w}（学生内变异性）与 ρ（相关系数）等输入参数的精度。与之相对，混合贝叶斯模型动态地平衡观测数据与关于 RTM 的先验知识，保证校正既不高估也不低估各种处理的真实效应。由此，本文把贝叶斯原理引入教育数据分析，推进了关于 RTM 的讨论。该模型的适用性远不止于教育研究，在 RTM 长期构成困扰的诸多领域都有用武之地。例如在临床试验中，贝叶斯方法有助于把真实的处理效应与患者结局中由 RTM 引起的波动区分开来，从而更准确地评估医学干预。环境研究者同样可以借助该模型，把 RTM 效应与生态或气候数据集中的真实变化剥离开来。无论用于大型研究还是小型课堂实验，这一混合贝叶斯模型都能最大限度地降低把 RTM 假象当作真实处理效应的风险。即便如此，未来研究仍可着力于在不同教育情境中对该模型作实证检验，以评估其适应性。后续研究还可以探索该模型与先进机器学习技术的结合，使先验参数的标定自动化，进一步提高其精度。最后，把该模型的应用扩展到其他学术领域的数据集，将为认识其更广泛的适用性提供重要启示。

参考文献

[1] 刘昌亚. 义务教育课后服务长效管理机制. 基础教育, 2020(08): 14-18.

[2] 丁钢. 传统文化融入现代教育的路径探索. 华东师范大学学报 (教育科学版), 2021, 39(04): 1-10.

[3] 许晓东. 高校实验室育人功能挖掘. 实验技术与管理, 2022, 38(07): 210-215.

[4] Isaac S, Michael W B. 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, 1995.

[5] 沈建华. 劳动教育课程体系开发与实践. 中国教育学刊, 2021(09): 84-88.

[6] 彭刚. 通识教育课程体系优化建设. 清华大学教育研究, 2021, 42(03): 1-8.

[7] 孙杰远. 民族地区教育现代化推进策略. 广西师范大学学报 (哲学社会科学版), 2021, 59(03): 108-115.

[8] Linden A. Assessing regression to the mean effects in health care initiatives. BMC Medical Research Methodology, 2013, 13: Article 119.

[9] 焦建利. 微课赋能课堂教学创新. 中国信息技术教育, 2026(16): 18-22.

[10] 王璐. 境外基础教育改革经验借鉴. 外国教育研究, 2021, 48(07): 3-16.

[11] 田慧生. 区域教育治理现代化建设路径. 教育科学研究, 2021(08): 5-11.

[12] Pinker S. Rationality: what it is, why it seems scarce, why it matters. Viking, 2021.