Original Paper

Comparing the Weighted Gain Score and a Rasch-Based Approach for Estimating Learning Outcomes in Medical Education: Quantitative Study

Rauf Aliyev[1], MD; Joy Backhaus[1], MSc; Silke Hammer[2], MD; Sarah König[1], MME, MD

[1]Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany
[2]Institute of Diagnostic and Interventional Radiology, University Hospital Würzburg, Würzburg, Germany

Corresponding Author:

Sarah König, MME, MD
Institute of Medical Teaching and Medical Education Research
University Hospital Würzburg
Josef-Schneider-Str. 2/D6
Würzburg 97080
Germany
Phone: +49 931 201 55210
Fax: +49 931 201 655213
Email: Koenig_Sarah@ukw.de

Abstract

Background: Pretest-posttest designs are widely used to estimate learning gain in studies evaluating educational interventions in medical education. The Weighted Gain Score (WGS) was proposed to reduce bias associated with differences in baseline performance.

Objective: This study evaluated the statistical and inferential properties of the WGS by comparing it to Rasch Learning Gain (RLG) across 3 datasets.

Methods: The WGS implements a weighting coefficient that includes the parameter μ, which linearly rescales the difference between pretest and posttest percentage scores. We examined the effect of varying μ (30, 50, and 70) on learning gain calculations and compared the results with those obtained using RLG. The following three datasets were analyzed: (1) a small illustrative dataset demonstrating the mathematical behavior of the WGS, (2) an empirical dataset from a previous educational evaluation study, and (3) a randomly generated binomial dataset designed to examine the metric under larger sample conditions.

Results: Changing the parameter μ in the WGS affected the magnitude of the calculated learning gains: lower μ-values produced larger gain estimates, whereas higher μ-values produced smaller estimates. Despite these differences in scale, the WGS and RLG correlated strongly in both the empirical dataset (r=0.93; P<.001) and the simulated dataset (r=0.92; P<.001); variation in μ did not alter the inferential results. Both methods identified the same interaction effect in the empirical dataset.

Conclusions: The WGS produced results highly consistent with those of RLG while requiring substantially lower computational complexity. The metric can be applied to both small and large datasets and allows μ to function as an adjustment coefficient for calibrating learning gain estimates across cohorts without altering inferential conclusions.

JMIR Med Educ 2026;12:e75516; doi: 10.2196/75516

Keywords: medical education; teaching quality; curriculum evaluation; learning gain; pretest-posttest design; Rasch model; Weighted Gain Score

Introduction

Teaching quality in medical education is a complex construct encompassing curriculum design, instructional methods, teaching expertise, learner engagement, and assessment practices [1-4]. High-quality teaching in this context contributes to the development of competent physicians and thereby influences the quality of patient care [5,6].

Among the various aspects of teaching quality in medical education, student learning outcomes represent one

measurable indicator frequently used in program evaluation and educational research [7-11]. However, interpreting learning outcomes as indicators of teaching effectiveness requires caution, as they are influenced by multiple factors beyond instructional quality. These include student motivation, prior knowledge, learning strategies, teacher enthusiasm, and learning activities occurring outside the formal curriculum [12,13]. To account for these influences, educational research often focuses not only on absolute performance but also on changes in performance over time. The concept of learning gain represents a widely used approach to capturing students' learning progress. In educational research, learning gain is commonly operationalized by assessing students before (pretest) and after (posttest) an educational intervention. The difference between pretest and posttest scores is then interpreted as an indicator of learning gain attributable, at least in part, to the educational intervention [14-16]. However, calculating learning gain is not trivial, as simple difference scores may lead to biased estimates depending on students' baseline knowledge. One simple approach is raw gain, which is calculated as the arithmetic difference between posttest and pretest scores. However, raw gain scores exhibit a negative correlation with baseline performance (ie, pretest scores) and are also affected by ceiling effects, meaning that students with lower pretest scores may appear to exhibit larger gains simply because they have more room for improvement [17-19].

To address these limitations, several modified gain metrics have been proposed. One widely used approach is the normalized gain introduced by Hake [20], which expresses the observed pretest-posttest gain relative to the maximum possible gain. Although this metric has been applied extensively in educational research [21,22], it also has important methodological limitations. It remains dependent on baseline performance, may inflate gains for students with high pretest scores, and behaves inconsistently when posttest scores fall below pretest scores or when pretest scores approach the maximum value [16,23].

Taken together, existing gain metrics may distort estimates of learning gain, particularly in cohorts with heterogeneous baseline knowledge. Many of these metrics either remain strongly dependent on baseline performance or require complex psychometric modeling. This highlights the need for approaches that provide statistically robust yet practically applicable estimates of learning gain in educational evaluation.

A recently proposed metric developed by our workgroup, the "Weighted Gain Score" (WGS), aims to address these limitations by applying a weighting coefficient that adjusts gain calculations according to students' baseline performance [16]. However, the statistical and inferential properties of this metric have not yet been systematically investigated. To address this gap, we evaluated the WGS by comparing it with Rasch Learning Gain (RLG), a Rasch model–based approach for estimating learning gain that served as the benchmark in our study [24]. Specifically, we addressed the following research questions:

- Does the WGS produce inferential results comparable to those produced by RLG?
- Can the parameter μ in the WGS be adjusted for different cohorts to calibrate learning gain calculations without altering inferential conclusions?

Through this analysis, we aimed to clarify the statistical behavior of the WGS and explore its potential applicability for the evaluation of educational interventions.

Methods

Metric WGS

The mathematical foundation of the WGS lies in the use of the weighting coefficient "pre/μ," which linearly transforms the difference between pretest and posttest percentage scores (denoted as "pre" and "post" in equation 1), thereby adjusting for pretest variability [16]. Formally, the WGS is defined as:

$$WGS = (post - pre) \times (pre/\mu) \quad (1)$$

To illustrate the computation, consider a hypothetical student with a pretest score of 40% and a posttest score of 70%.

For μ=50, the WGS is calculated as: WGS = (70 – 40) × (40/50) = 30 × 0.8 = 24.

If μ is increased to 70, the same performance yields: WGS = (70 – 40) × (40/70) = 30 × 0.57 = 17.14.

This example illustrates that increasing μ reduces the magnitude of the calculated gain while preserving the relative ordering of observations. When posttest scores fall below pretest scores, the WGS assumes negative values, indicating a decrease in performance.

Originally, the parameter μ used in the weighting coefficient was defined as the average pretest score of a cohort. It was constrained to integer values between 1 and 100, consistent with the percentage format of "pre" and "post." In the original formulation, its value was set at 50 as a default reference value [16]. In this study, μ is interpreted as an adjustment coefficient that functions as a scaling parameter for learning gain calculations. Changing its value proportionally rescales the calculated gain scores: higher values of μ lead to smaller gain estimates, whereas lower values produce larger gain estimates. Importantly, this modification represents a linear transformation of the calculated values and therefore does not alter the underlying statistical relationships among observations.

To examine the influence of this parameter on the stability of the WGS, we tested 3 calibration levels in our datasets: μ=30, μ=50, and μ=70. These values represent 3 nonextreme points within the possible range of 1 to 100, allowing us to evaluate the behavior of the WGS across low, moderate, and high scaling conditions.

Rasch Model and RLG

The Rasch model is a fundamental concept in modern psychometric measurement. The probability that a student

answers a specific item correctly depends on 2 key factors: the student's ability and the difficulty of the item. In the Rasch framework, a student's latent ability is denoted by θ, whereas item difficulty is represented by β. When a student's ability exceeds the difficulty of an item, the probability of answering correctly increases, and vice versa [25]. Because the Rasch model allows the estimation of individual students' abilities independently of the specific test items used, it is widely applied in educational measurement and medical education research [26]. With this in mind, we selected RLG as a reference method for evaluating the WGS.

We applied the dichotomous 1-parameter logistic Rasch model. Item parameters were estimated using conditional maximum likelihood estimation. On the basis of the fitted model, person abilities were subsequently calculated using maximum likelihood estimation separately for the pretest (θ_{pre}) and posttest (θ_{post}) data [27].

As indicated in equation 2, RLG was defined as the difference between the estimated posttest and pretest abilities. This difference represents the change in latent ability on the Rasch measurement scale and serves as an estimate of individual learning gain across the instructional intervention [24,28].

$$RLG = (\theta_{post} - \theta_{pre}) \quad (2)$$

To ensure the validity of Rasch-based ability estimates, we examined global model fit indicators. Item infit and outfit statistics ranged between 0.7 and 1.3, which is generally considered acceptable for the Rasch model. In addition, person reliability exceeded 0.8, and separation indices were >2, indicating satisfactory measurement precision.

Datasets

Three datasets were used to examine the behavior of the WGS under different analytical conditions:

1. The illustrative dataset (n=10): a small artificial dataset designed to illustrate the mathematical behavior of the parameter μ within the WGS metric
2. The empirical dataset (n=170): a dataset consisting of real-world data derived from a previously published educational evaluation study [29], used to examine the behavior of the WGS under authentic educational conditions and to perform inferential statistical analyses
3. The simulated dataset (n=1000): a randomly generated binomial dataset designed to mirror the structure of the empirical dataset while providing a larger sample size, allowing the behavior of the parameter μ to be examined independently of the empirical data

The Illustrative Dataset

Following the design of the simulated dataset in our previous study [16], we created an artificial dataset by combining different pretest scores with varying levels of raw gain in test performance, defined as the absolute difference between posttest and pretest scores. Pretest scores ranged from 1 to 10 points, and the gain in performance was simulated by increasing test scores by 1 to 4 points. To avoid potential ceiling effects, the analysis included only combinations in which the sum of pretest scores and the simulated gains did not exceed the maximum of 10 points. The sample size of the illustrative dataset was set at 10. RLG was not applicable here, as Rasch model–based estimation requires larger sample sizes to obtain stable parameter estimates [30].

The Empirical Dataset

The empirical dataset originated from a prospective educational study conducted at the University Medical Center Göttingen in Göttingen, Germany [29]. The study compared the learning gain of students attending a traditional lecture on goiter with that of students using a corresponding video podcast (vodcast) within the teaching module "Operative Medicine." The study was conducted over 2 consecutive semesters using a pretest-posttest design based on 9 multiple-choice test items. A total of 170 students participated. Students were additionally surveyed regarding their learning dispositions, which resulted in the classification of participants into 2 groups: "traditional learners" and "digital natives." A total of 35 students (20.59%) could not be clearly assigned to either group and were therefore excluded from group-based analyses. Consequently, 135 (79.41%) students were included in the 2-way ANOVA examining the interaction between teaching format and learning disposition (Table 1).

Table 1. Distribution of students according to teaching format and learning disposition in the empirical and simulated datasets.

Datasets and teaching formats	Traditional learners, n (%)	Digital natives, n (%)
Empirical dataset (N=135)		
Lecture	38 (28.15)	34 (25.19)
Vodcast	28 (20.74)	35 (25.93)
Simulated dataset (N=1000)		
Lecture	259 (25.9)	210 (21)
Vodcast	250 (25)	281 (28.1)

The Simulated Dataset

The simulated dataset was generated using a random binomial distribution, assuming a 50% probability of correctly answering a hypothetical examination item. This probability was applied to 9 multiple-choice items in both the pretest and the posttest scores, reflecting the structure of the "empirical dataset." Apart from the larger sample size, the primary difference between the empirical and simulated datasets was the random allocation of group variables. Two variables

were simulated: teaching format and learning disposition. Both variables were coded dichotomously. For consistency in labeling, the simulated variables were named analogously to those in the empirical dataset, although they represent random group assignments rather than actual instructional formats or learning characteristics. Each simulated student had a 50% probability of being assigned to each category (Table 1). The sample size for the simulated dataset was 1000.

Statistical Analysis

All simulations and statistical analyses were conducted using the R software suite (version 4.1.2; R Foundation for Statistical Computing) [31]. Rasch modeling was performed using the *eRm* package [32].

To examine the relationship between the 2 learning gain metrics, Pearson correlation coefficients were calculated between the WGS and RLG scores.

To investigate potential interaction effects between teaching format and learning disposition, we conducted a 2-way ANOVA. Post hoc comparisons were performed using Bonferroni-adjusted contrasts. Effect sizes were reported as partial η^2, and 95% CIs were calculated where appropriate.

Normality of the dependent variables was assessed using the Shapiro-Wilk test and visual inspection of Q-Q plots. Minor deviations from normality were observed, which are common in bounded percentage scores (0% to 100%) frequently used in educational assessments. Given the present sample sizes and the absence of influential outliers, ANOVA was considered sufficiently robust to moderate violations of the normality assumption.

Homogeneity of variances across groups was evaluated using the Levene test and the Brown-Forsythe test, both of which indicated no statistically significant differences in variance between groups. All statistical tests were 2-sided, and a significance level of $P<.05$ was applied.

Ethical Considerations

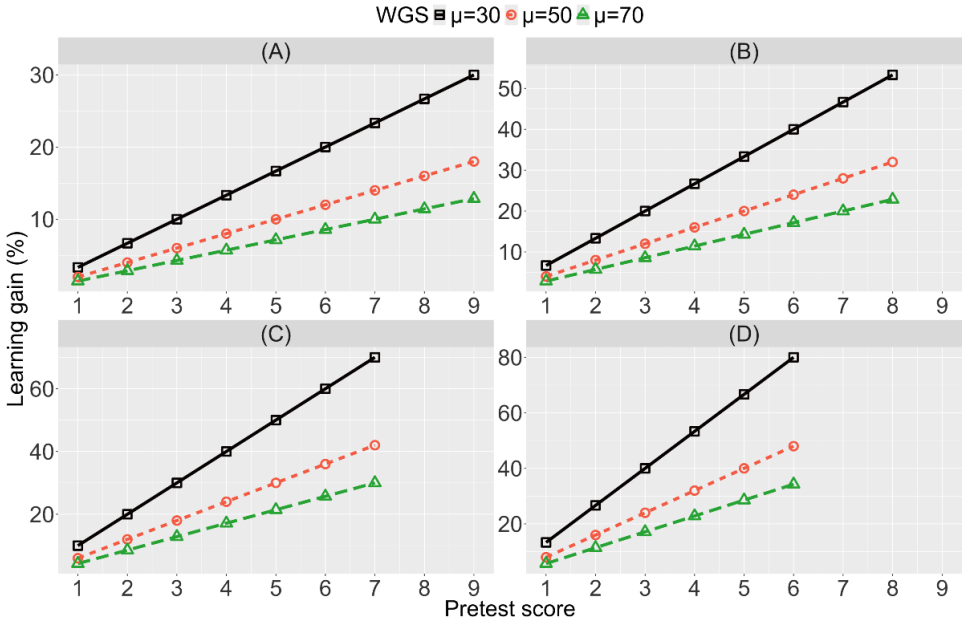
The empirical data analyzed within this work were reviewed and judged by the local institutional review and ethics board (Medical Ethics Committee, University Medical Center Göttingen) as not representing medical or epidemiological research on human participants and, therefore, were assessed using a simplified assessment protocol. The project was approved without any reservation under proposal number 1-11-14.

Results

Effect of the Parameter μ on WGS Learning Gain Estimates

The illustrative dataset demonstrates the mathematical effect of varying μ (30, 50, and 70) on the WGS. Changes in μ systematically altered the slope of the WGS learning gain plots (Figure 1). Each subplot represents a different raw gain scenario, ranging from 1 to 4 points. As the μ-value increased, the slope of the learning gain curve decreased, resulting in smaller WGS values for the same pretest score. For example, with a gain of 1 point and a pretest score of 6, the WGS was approximately 20% for μ=30 and decreased to <10% for μ=70. This pattern remained consistent across all 4 gain scenarios, illustrating that increasing μ reduces the magnitude of the calculated learning gain while preserving the relative ordering of observations.

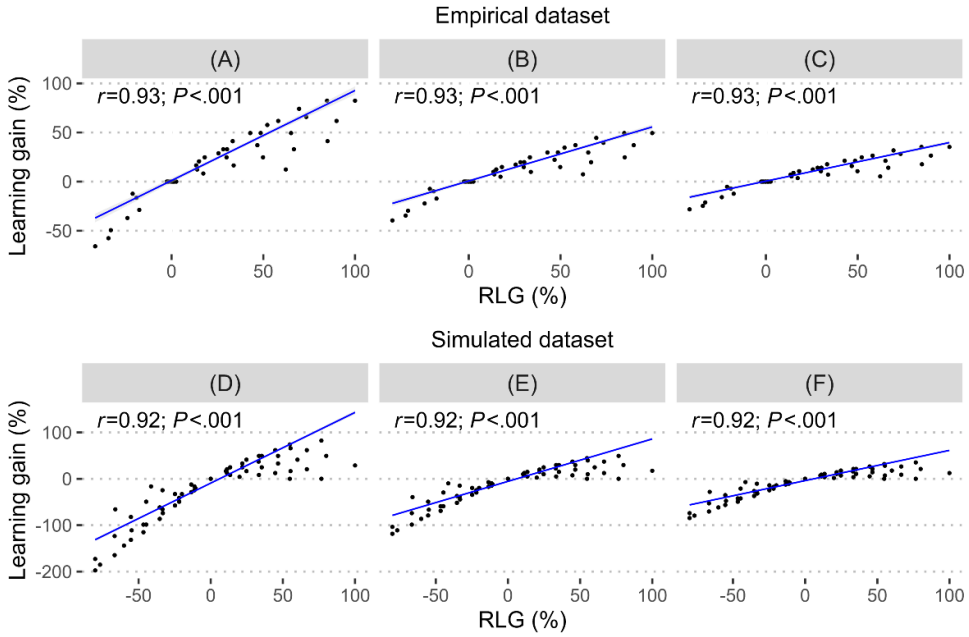
Figure 1. Effect of varying μ (30, 50, and 70) on Weighted Gain Score (WGS) learning gain estimates in the illustrative dataset. (A) Gain of 1 point, (B) gain of 2 points, (C) gain of 3 points, and (D) gain of 4 points.



Correlation Analysis Between WGS and RLG

The WGS demonstrated a strong positive correlation with RLG across all tested μ-values (Figure 2). In the empirical dataset, the Pearson correlation coefficient was consistently high (r=0.93; P<.001). A similarly strong relationship was observed in the simulated dataset (r=0.92; P<.001). The correlation coefficients remained identical across the tested μ-values (30, 50, and 70) in both datasets.

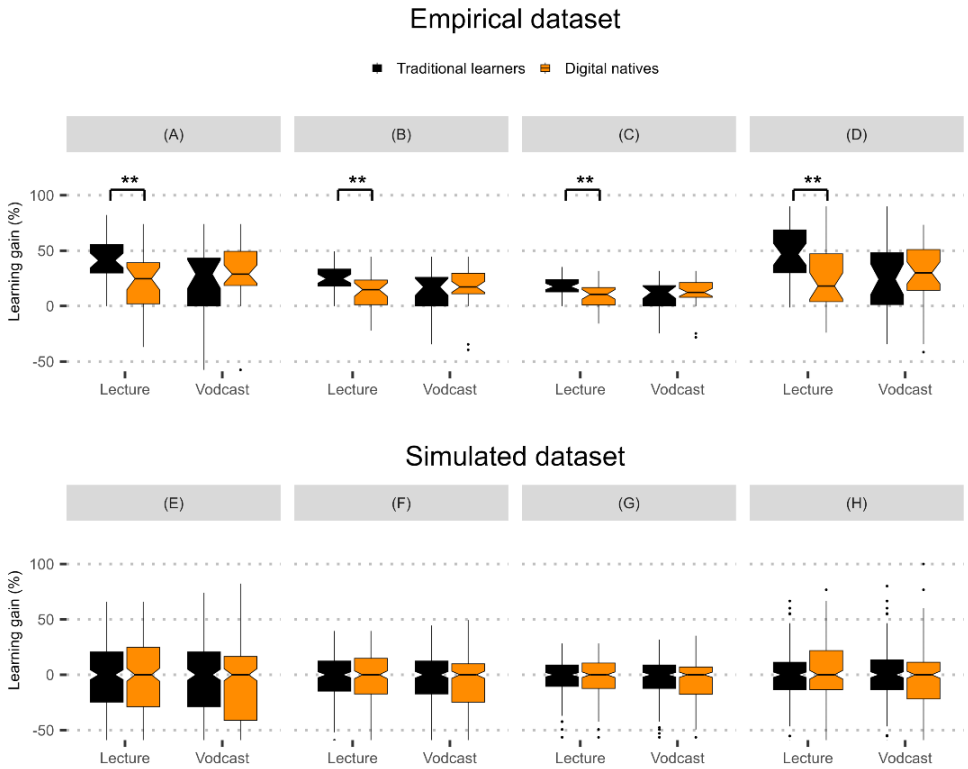
Figure 2. Correlation between Weighted Gain Score (WGS) and Rasch Learning Gain (RLG) in the empirical and simulated datasets. (A) Empirical dataset with μ=30, (B) empirical dataset with μ=50, (C) empirical dataset with μ=70, (D) simulated dataset with μ=30, (E) simulated dataset with μ=50, and (F) simulated dataset with μ=70.



Analysis of Interaction Effects Using WGS and RLG

All 3 calibrations of the WGS (μ=30, μ=50, and μ=70) detected a significant interaction effect between teaching format and learning disposition in the empirical dataset (Figure 3). Traditional learners displayed higher learning gains in the lecture format than digital natives ($F_{1,131}$=6.51; P=.01; partial η^2=0.05). For μ=50, the mean difference was −11.64 (95% Bonferroni-adjusted CI −21.46 to −1.83; P=.01). Corresponding estimates were −19.41 for μ=30 (95% Bonferroni-adjusted CI −35.80 to −3.04; P=.01) and −8.32 for μ=70 (95% Bonferroni-adjusted CI −15.33 to −1.31; P=.01).

Figure 3. Learning gain estimates calculated using Weighted Gain Score (WGS) and Rasch Learning Gain (RLG), depicting the interaction between teaching format and learning disposition in the empirical and simulated datasets. (A) Empirical dataset with WGS (μ=30), (B) empirical dataset with WGS (μ=50), (C) empirical dataset with WGS (μ=70), (D) empirical dataset with RLG, (E) simulated dataset with WGS (μ=30), (F) simulated dataset with WGS (μ=50), (G) simulated dataset with WGS (μ=70), and (H) simulated dataset with RLG. **Indicates statistical significance at *P*=.01.



RLG also detected this interaction effect ($F_{1,131}$=6.75; *P*=.01; partial η^2=0.05) with a mean difference of −19.91 (95% Bonferroni-adjusted CI −36.80 to −3.05; *P*=.01), confirming the interaction pattern observed in the original study from which our empirical dataset was derived [16,29].

In the simulated dataset, no significant interaction between teaching format and learning disposition was observed when learning gains were calculated using the WGS, regardless of the μ-value applied ($F_{1,996}$=0.39; *P*=.53; partial η^2<0.001; Figure 3). Similarly, RLG did not reveal any significant difference in performance between the groups ($F_{1,996}$=1.10; *P*=.29; partial η^2=0.001). Because teaching format and learning disposition were randomly assigned in the simulated dataset, we did not necessarily expect any interaction effect.

Discussion

Inferential Behavior of WGS Compared With RLG

A robust method for calculating learning gain is essential for capturing students' learning progress following an educational intervention and for providing interpretable indicators of educational effectiveness. Such a method should be statistically sound, transparent, and practically applicable within evaluation processes.

This study evaluated the statistical behavior of the WGS, a method designed to estimate learning gain in a way that is both methodologically robust and straightforward to implement. The first research question examined whether the WGS yields inferential results comparable to those obtained with RLG. Our findings demonstrated a strong inferential correspondence between the 2 methods. The WGS produced learning gain estimates that correlated highly with those derived from RLG, while also identifying the same interaction effect in the empirical dataset as the Rasch model–based approach. Importantly, these inferential conclusions remained stable across all tested μ-values (30, 50, and 70). The identical correlation coefficients between the WGS and RLG and the unchanged ANOVA results indicate that modifying the parameter μ linearly rescales learning gain estimates. Consequently, varying μ changes the magnitude of WGS values but does not affect statistical inference.

Robustness of WGS Under Nonnormally Distributed Data

Neither the empirical nor the simulated dataset fully satisfied the assumption of normality, although no substantial skewness was observed. In medical education research, deviations from normality are common, particularly in pretest-posttest designs [17,33]. A ceiling effect occurs when pretest scores approach the maximum possible value, limiting the measurable improvement, whereas a floor effect arises when pretest performance is concentrated near the minimum score in a difficult test. Very easy items tend to produce ceiling effects, whereas very difficult items may lead to floor effects. Despite deviations from normality, the WGS demonstrated stable inferential behavior across the empirical and simulated datasets, suggesting a degree of robustness. This finding is consistent with previous research indicating that parametric methods such as ANOVA and correlation

analyses are generally robust to moderate violations of normality, particularly in samples of the size examined in this study [34-36]. Nevertheless, future research is needed to examine the behavior of the WGS across a broader range of distributional scenarios to better establish its reliability.

Applicability of WGS in Small Samples

The illustrative dataset demonstrates that the WGS yields interpretable results even with very small sample sizes. In contrast, Rasch model–based approaches typically require substantially larger samples to ensure stable estimation of item parameters and person abilities [24,26]. This distinction is particularly relevant in educational settings with small cohorts, such as specialized teaching modules, pilot courses, or resource-intensive instructional interventions. In such contexts, the WGS may represent a practical alternative method for estimating learning gain because it does not rely on complex parameter estimation.

More broadly, transparent feedback on learning outcomes supports the continuous development of teaching practices, as evidence suggests that feedback on educational performance encourages educators to engage in reflective improvement of their teaching [37-39].

The Role of μ as an Adjustment Coefficient

The second research question examined whether the parameter μ in the WGS can be adjusted across different cohorts to calibrate learning gain calculations without altering inferential outcomes. In the original study introducing the WGS, μ was defined as the average pretest score of a cohort. Our findings suggest that the role of μ can be understood more broadly. Rather than representing solely the cohort mean, μ functions as a scaling parameter that allows calibration of the learning gain metric. To reflect this role more accurately, we interpret μ in this study as an adjustment coefficient that can be modified depending on the analytical purpose of the evaluation. On the basis of the results of this study, 3 conceptual adjustment strategies can be distinguished: absolute adjustment, relative adjustment, and routine evaluation. A decision framework for selecting μ is provided in Multimedia Appendix 1.

Absolute Adjustment: Monitoring of Cohort Learning

Absolute adjustment refers to the use of a fixed μ-value to estimate learning gain within a stable scaling framework. When μ remains constant, differences in learning gain across courses, time points within the curriculum, or different cohorts can be interpreted without recalibration of the metric, thereby ensuring cross-cohort comparability. This approach supports standardized monitoring of educational outcomes, for example, when evaluating curricular developments over time or comparing modules within a program. Observed differences in learning gain may arise from multiple factors, including instructional design, assessment characteristics, or cohort composition. Maintaining a fixed μ ensures that such differences remain visible and can be attributed to substantive factors.

Relative Adjustment: Evaluation of Teaching Interventions

Relative adjustment enables comparison of teaching interventions across cohorts with heterogeneous characteristics. In educational practice, cohorts often differ in characteristics such as demographics, motivation, workload, or external contextual influences [13,40]. When learning gain is used to compare instructional formats, such heterogeneity may affect the interpretation of outcomes. Under a relative adjustment strategy, a μ-value may be calibrated separately for each cohort, allowing the scaling of learning gain calculations to reflect cohort-specific baseline conditions. Although this approach does not eliminate potential confounding factors, it may reduce systematic bias associated with heterogeneous starting conditions. This strategy is particularly useful when learning gain is evaluated without strict requirements for cross-cohort comparability, but with a focus on fair comparison of teaching interventions within specific cohorts or instructional contexts.

Routine Evaluation: Selecting μ in Practice

In routine applications, when learning gain is estimated without strict requirements for cross-cohort comparability or cohort-specific calibration, μ may be selected pragmatically based on the cohort's mean pretest performance. For example, cohorts with mean pretest scores approximately 50% of the maximum achievable score may be assigned μ=50, whereas cohorts with substantially higher or lower baseline knowledge may be assigned correspondingly higher (eg, ≥70) or lower (eg, ≤30) μ-values. This pragmatic approach enables straightforward estimation of learning gain while preserving a transparent and easily interpretable scaling of the WGS metric.

Limitations

One limitation of the WGS arises when a student obtains a pretest score of zero, which results in a calculated learning gain of zero regardless of posttest performance. In practice, such cases are unlikely in multiple-choice assessments because guessing and prior knowledge increase the probability of obtaining at least 1 correct response [41]. In the empirical dataset, no student recorded a pretest score of zero, and in the simulated dataset, a negligible number (3 out of 1000 students) achieved zero points on the pretest score. One possible strategy is to exclude such observations from the analysis. However, this may reduce statistical power and introduce bias if students with low baseline scores are systematically underrepresented. Alternatively, a small positive offset (pseudocount) could be added to avoid undefined computations, analogous to continuity corrections used in categorical data analysis [42,43]. The implications of such adjustments should be examined in future methodological studies, for example, through sensitivity analy-

ses comparing different handling strategies for zero-baseline observations [44].

A further limitation concerns the sample size of the empirical dataset (n=170). Although cohort sizes of this magnitude are common in single-semester cohorts at German medical faculties, they are slightly below commonly cited recommendations for stable Rasch parameter estimation, which often suggest sample sizes of approximately 150 to 200 participants or more [45]. Nevertheless, global model fit indicators in the empirical dataset (infit and outfit statistics, person reliability, and separation indices) were within acceptable ranges, supporting the interpretability of the RLG-based estimates despite the moderate sample size.

Another limitation relates to the test length used in the simulated dataset, which consisted of 9 multiple-choice items to mirror the empirical dataset. Because measurement reliability generally increases with test length [46-50], the limited number of items may reduce measurement precision and restrict the generalizability of the findings. Therefore, future research should examine the performance of the WGS in assessments with larger item sets that more closely reflect the scope of medical examinations.

Finally, both datasets exhibited deviations from normality, although homogeneity of variances across groups was supported by the Levene and Brown-Forsythe tests, and no influential outliers were observed. Previous methodological research indicates that ANOVA and Pearson correlation are generally robust to moderate violations of normality, particularly in samples of the present size [34-36]. Therefore, we consider the impact of nonnormality on the inferential conclusions to be limited.

Conclusions and Future Research

This study evaluated the WGS as a method for estimating learning gain in pretest-posttest educational designs. Our findings indicate that the WGS provides robust and easily interpretable estimates while remaining computationally simple. Rather than replacing established psychometric models, the WGS may complement existing approaches, particularly in routine educational evaluations.

Future research should further develop the WGS as a broadly applicable evaluation instrument. In particular, establishing a methodologically sound calibration framework for μ will be essential, including empirically grounded decision models that guide μ-selection according to the evaluation purpose, such as cohort monitoring or comparative evaluation of teaching interventions. In addition, integrating the WGS into structured program evaluations, including longitudinal monitoring across courses, will be important for assessing its generalizability across educational contexts.

Future work may also explore the integration of the WGS within Bayesian test-theoretical frameworks [51]. By incorporating prior information and updating gain estimates as new data become available, Bayesian approaches could further improve the precision and contextual sensitivity of WGS-based learning gain estimates. Further studies should also examine the behavior of the WGS under different distributional conditions to better establish its robustness.

Acknowledgments

The authors sincerely thank Simone Kann and Michael Schuler for their valuable insights and thoughtful suggestions, as well as Andrew Entwistle for his contribution to the revision of this manuscript.

Funding

This research did not receive funding from any specific grant provided by public, commercial, or not-for-profit agencies.

Data Availability

The data supporting the findings of this study are provided as a multimedia appendix to facilitate full reproducibility.

Authors' Contributions

All authors were involved in the conception and/or design of the study and contributed critically to the final preparation of this study, including approving the final version of the manuscript. In particular, SK conceived and designed the study, wrote the final study protocol, and drafted the manuscript. RA conducted the study, collected the results, and analyzed the data. SH and JB analyzed the data and performed and verified the statistical analyses.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Decision framework for selecting the calibration parameter μ in Weighted Gain Score calculations according to the evaluation objective (absolute adjustment, relative adjustment, or routine evaluation).
[DOCX File (Microsoft Word File), 194 KB-Multimedia Appendix 1]

References

1. Charalambous CY, Praetorius AK, Sammons P, Walkowiak T, Jentsch A, Kyriakides L. Working more collaboratively to better understand teaching and its quality: challenges faced and possible solutions. Stud Educ Eval. Dec 2021;71(1):101092. [doi: 10.1016/j.stueduc.2021.101092]

2. Gibson KA, Boyle P, Black DA, Cunningham M, Grimm MC, McNeil HP. Enhancing evaluation in an undergraduate medical education program. Acad Med. Aug 2008;83(8):787-793. [doi: 10.1097/ACM.0b013e31817eb8ab] [Medline: 18667897]
3. Litzelman DK, Stratos GA, Marriott DJ, Skeff KM. Factorial validation of a widely disseminated educational framework for evaluating clinical teachers. Acad Med. Jun 1998;73(6):688-695. [doi: 10.1097/00001888-199806000-00016] [Medline: 9653408]
4. Noor C, Hozan CT, Vîlceanu N, Bonțea MG. A review of the effectiveness of the role of various components in medical education. Arch Pharm Pract. 2023;14(4):155-159. [doi: 10.51847/LrElkFGJAO]
5. McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Medical education featuring mastery learning with deliberate practice can lead to better health for individuals and populations. Acad Med. Nov 2011;86(11):e8-e9. [doi: 10.1097/ACM.0b013e3182308d37] [Medline: 22030671]
6. Gould BE, Grey MR, Huntington CG, et al. Improving patient care outcomes by teaching quality improvement to medical students in community-based practices. Acad Med. Oct 2002;77(10):1011-1018. [doi: 10.1097/00001888-200210000-00014] [Medline: 12377677]
7. Schiekirka-Schwake S, Anders S, von Steinbüchel N, Becker JC, Raupach T. Facilitators of high-quality teaching in medical school: findings from a nation-wide survey among clinical teachers. BMC Med Educ. Sep 29, 2017;17(1):178. [doi: 10.1186/s12909-017-1000-6] [Medline: 28962568]
8. Schiekirka S, Reinhardt D, Beißbarth T, Anders S, Pukrop T, Raupach T. Estimating learning outcomes from pre- and posttest student self-assessments: a longitudinal study. Acad Med. Mar 2013;88(3):369-375. [doi: 10.1097/ACM.0b013e318280a6f6] [Medline: 23348083]
9. Gruppen LD. Outcome-based medical education: implications, opportunities, and challenges. Korean J Med Educ. Dec 2012;24(4):281-285. [doi: 10.3946/kjme.2012.24.4.281] [Medline: 25813324]
10. Harden RM. AMEE guide no. 14: outcome-based education: part 1-an introduction to outcome-based education. Med Teach. Jan 1999;21(1):7-14. [doi: 10.1080/01421599979969]
11. Haverkamp N, Barth J, Schmidt D, Dahmen U, Keis O, Raupach T. Position statement of the GMA committee “teaching evaluation”. GMS J Med Educ. 2024;41(2):Doc19. [doi: 10.3205/zma001674] [Medline: 38779701]
12. Fraenkel JR, Wallen NE, Hyun HH. How to Design and Evaluate Research in Education. 8th ed. McGraw-Hill; 2012.
13. Cook DA, Beckman TJ. Reflections on experimental research in medical education. Adv Health Sci Educ Theory Pract. Aug 2010;15(3):455-464. [doi: 10.1007/s10459-008-9117-3] [Medline: 18427941]
14. Colt HG, Davoudi M, Murgu S, Zamanian Rohani N. Measuring learning gain during a one-day introductory bronchoscopy course. Surg Endosc. Jan 2011;25(1):207-216. [doi: 10.1007/s00464-010-1161-4] [Medline: 20585964]
15. McGrath C, Guerin B, Harte E, Frearson M, Manville C. Learning gain in higher education. RAND Corporation. 2015. URL: https://www.rand.org/pubs/research_reports/RR996.html [Accessed 2026-05-25]
16. Westphale S, Backhaus J, Koenig S. Quantifying teaching quality in medical education: the impact of learning gain calculation. Med Educ. Mar 2022;56(3):312-320. [doi: 10.1111/medu.14694] [Medline: 34767274]
17. Šimkovic M, Träuble B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. PLoS One. 2019;14(8):e0220889. [doi: 10.1371/journal.pone.0220889] [Medline: 31425561]
18. Bereiter C. Some persisting dilemmas in the measurement of change. In: Harris CW, editor. Problems in Measuring Change. University of Wisconsin Press; 1963:3-20.
19. Prieler J, Raven J. Problems in the measurement of change (with particular reference to individual change [gain] scores) and their potential solution using IRT. In: Uses and Abuses of Intelligence: Studies Advancing Spearman and Raven’s Quest for Non-Arbitrary Metrics. Royal Fireworks Press; 2008:173-210.
20. Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. Am J Phys. Jan 1998;66(1):64-74. [doi: 10.1119/1.18809]
21. Coletta VP, Phillips JA. Interpreting FCI scores: normalized gain, preinstruction scores, and scientific reasoning ability. Am J Phys. Dec 2005;73(12):1172-1182. [doi: 10.1119/1.2117109]
22. Nissen JM, Talbot RM, Thompson AN, Van Dusen B. Comparison of normalized gain and Cohen’s d for analyzing gains on concept inventories. Phys Rev Phys Educ Res. 2018;14:010115. [doi: 10.1103/PhysRevPhysEducRes.14.010115]
23. Marx JD, Cummings K. Normalized change. Am J Phys. 2007;75(1):87-91. [doi: 10.1119/1.2372468]
24. Pentecost TC, Barbera J. Measuring learning gains in chemical education: a comparison of two methods. J Chem Educ. Jul 2013;90(7):839-845. [doi: 10.1021/ed400018v]
25. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. 3rd ed. Routledge; 2015.

26. Downing SM. Item response theory: applications of modern test theory in medical education. Med Educ. Aug 2003;37(8):739-745. [doi: 10.1046/j.1365-2923.2003.01587.x] [Medline: 12945568]
27. Embretson SE, Reise SP. Item Response Theory. Psychology Press; 2013.
28. Wallace CS, Bailey JM. Do concept inventories actually measure anything? Astron Educ Rev. 2010;9(1). [doi: 10.3847/AER2010024]
29. Backhaus J, Huth K, Entwistle A, Homayounfar K, Koenig S. Digital affinity in medical students influences learning outcome: a cluster analytical design comparing vodcast with traditional lecture. J Surg Educ. 2019;76(3):711-719. [doi: 10.1016/j.jsurg.2018.12.001] [Medline: 30833205]
30. Chen WH, Lenderking W, Jin Y, Wyrwich KW, Gelhorn H, Revicki DA. Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. Qual Life Res. Mar 2014;23(2):485-493. [doi: 10.1007/s11136-013-0487-5] [Medline: 23912855]
31. R Core Team. The R Project for Statistical Computing. R Foundation for Statistical Computing. 2013. URL: http://www.R-project.org [Accessed 2026-05-25]
32. Mair P, Hatzinger R. Extended Rasch modeling: the eRm package for the application of IRT models in R. J Stat Soft. 2007;20(9):1-20. [doi: 10.18637/jss.v020.i09]
33. Micceri T. The unicorn, the normal curve, and other improbable creatures. Psychol Bull. 1989;105(1):156-166. [doi: 10.1037/0033-2909.105.1.156]
34. Bishara AJ, Hittner JB. Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. Psychol Methods. Sep 2012;17(3):399-417. [doi: 10.1037/a0028087] [Medline: 22563845]
35. Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. Behav Res Methods. Dec 2021;53(6):2576-2590. [doi: 10.3758/s13428-021-01587-5] [Medline: 33963496]
36. Havlicek LL, Peterson NL. Robustness of the Pearson correlation against violations of assumptions. Percept Mot Skills. Dec 1976;43(3_suppl):1319-1334. [doi: 10.2466/pms.1976.43.3f.1319]
37. Boerboom TBB, Stalmeijer RE, Dolmans DHJM, Jaarsma DADC. How feedback can foster professional growth of teachers in the clinical workplace: a review of the literature. Studies in Educational Evaluation. Sep 2015;46:47-52. [doi: 10.1016/j.stueduc.2015.02.001]
38. Scheeler MC, Ruhl KL, McAfee JK. Providing performance feedback to teachers: a review. Teach Educ Spec Educ. 2004;27(4):396-407. [doi: 10.1177/088840640402700407]
39. Evans C, Howson CK, Forsythe A. Making sense of learning gain in higher education. Higher Education Pedagogies. 2018;3(1):1-45. [doi: 10.1080/23752696.2018.1508360]
40. Ewert A, Sibthorp J. Creating outcomes through experiential education: the challenge of confounding variables. Journal of Experiential Education. Jan 1, 2009;31(3):376-389. [doi: 10.5193/JEE.31.3.376]
41. Kubinger KD, Gottschall CH. Item difficulty of multiple choice tests dependant on different item response formats—an experiment in fundamental research on psychological assessment. Psychol Sci. 2007;49(4):361-374.
42. Weber F, Knapp G, Ickstadt K, Kundt G, Glass Ä. Zero-cell corrections in random-effects meta-analyses. Res Synth Methods. Nov 2020;11(6):913-919. [doi: 10.1002/jrsm.1460] [Medline: 32991790]
43. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. May 15, 2004;23(9):1351-1375. [doi: 10.1002/sim.1761] [Medline: 15116347]
44. Aung NM, Jurak I, Mehmood S, Axon E. Sensitivity analysis in meta-analysis: a tutorial. Cochrane Evid Synth Methods. Jan 2026;4(1):e70067. [doi: 10.1002/cesm.70067] [Medline: 41497796]
45. O'Neill TR, Gregg JL, Peabody MR. Effect of sample size on common item equating using the dichotomous Rasch model. Appl Meas Educ. Jan 2020;33(1):10-23. [doi: 10.1080/08957347.2019.1674309]
46. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. Sep 2004;38(9):1006-1012. [doi: 10.1111/j.1365-2929.2004.01932.x] [Medline: 15327684]
47. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. Jun 27, 2011;2:53-55. [doi: 10.5116/ijme.4dfb.8dfd] [Medline: 28029643]
48. de Vet HC, Mokkink LB, Mosmuller DG, Terwee CB. Spearman-Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. J Clin Epidemiol. May 2017;85:45-49. [doi: 10.1016/j.jclinepi.2017.01.013] [Medline: 28342902]
49. Brown W. Some experimental results in the correlation of mental abilities. Br J Psychol 1904-1920. Oct 1910;3(3):296-322. [doi: 10.1111/j.2044-8295.1910.tb00207.x]
50. Spearman C. Correlation calculated from faulty data. Br J Psychol 1904-1920. Oct 1910;3(3):271-295. [doi: 10.1111/j.2044-8295.1910.tb00206.x]

51. Rindskopf D. Overview of Bayesian statistics. Eval Rev. Aug 2020;44(4):225-237. [doi: 10.1177/0193841X19895623] [Medline: 31894697]

Abbreviations

WGS: Weighted Gain Score
RLG: Rasch Learning Gain

Edited by Awsan Bahattab; peer-reviewed by Muhammad Saeed Shafi, Shaikha Alzaabi, T A Valencia-Perez; submitted 18.May.2025; final revised version received 19.Mar.2026; accepted 22.Apr.2026; published 16.Jun.2026

Please cite as:
Aliyev R, Backhaus J, Hammer S, König S
Comparing the Weighted Gain Score and a Rasch-Based Approach for Estimating Learning Outcomes in Medical Education: Quantitative Study
JMIR Med Educ 2026;12:e75516
URL: https://mededu.jmir.org/2026/1/e75516
doi: 10.2196/75516