Received: 30 April 2021 | Revised: 12 October 2021 | Accepted: 5 November 2021

DOI: 10.1111/medu.14694

RESEARCH ARTICLE

ASSESSMENT

Quantifying teaching quality in medical education: The impact of learning gain calculation

Silke Westphale | **Joy Backhaus** | **Sarah Koenig**

Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany

Correspondence
Sarah Koenig, Institute of Medical Teaching and Medical Education Research, University Hospital Würzburg, Würzburg, Germany.
Email: koenig_sarah@ukw.de

Abstract

Background: Student performance is a mirror of teaching quality. The pre-/post-test design allows a pragmatic approach to comparing the effects of interventions. However, the calculation of current knowledge gain scores introduces varying degrees of distortion. Here we present a new metric employing a linear weighting coefficient to reduce skewness on outcome interpretation.

Methods: We compared and contrasted a number of common scores (raw and relative gain scores) with our new method on two datasets, one simulated and the other empirical from a previous intervention study ($n = 180$) employing a pre-/post-test design.

Results: The outcomes of the common scores were clearly different, demonstrating a significant dependency on pre-test scores. Only the new metric revealed a linear relationship to the knowledge baseline, was less skewed on the upper or lower extremes, and proved well suited to allow the calculation of negative learning gains. Employing the empirical dataset, the new method also confirmed the interaction effect of teaching formats with specific subgroups of learner characteristics.

Conclusion: This work introduces a new weighted metric enabling meaningful comparisons between interventions based on a linear transformation. This method will form the basis to intertwine the calculation of test performance closely with the outcome of learning as an important factor reflecting teaching quality and efficacy. Its regular use can improve the transparency of teaching activities and outcomes, contribute to forming rounded judgements of students' acquisition of knowledge and skills and enable valuable feedforward to develop and enhance curricular concepts.

1 | INTRODUCTION

Key players in education focus on assessing whether teaching results in any measurable improvement in test performance.[1,2] To assess changes in learning, educators often collect data from student testing both prior to (pre-test) and after (post-test) the intervention.[3] The paired test data are then analysed using gain scores to quantify classroom learning gain as an average for the cohort as well as for the individual.[4] Despite the clear study design, the statistical problem still exists as to how to analyse the learning gain best, as ceiling effects and pre-test scores can markedly influence the calculated values.[5–7]

One basic method to measure changes in learning is to determine the raw gain, representing the absolute difference between the post-test and pre-test scores.[8] Its analysis is attractive in terms of

simplicity, but fails to account for the observation that higher pre-test scores result in disproportionately lower learning gain.[9] Controversies regarding the lack of controls cause further problems with its use.[10–12]

A number of modified gain scores have been introduced to assess learning gain in relation to the baseline knowledge of individual students or cohorts (relative gain scores).[4] In particular, many authors suggested the use of the normalised gain score,[4,13–15] which is defined as the ratio between the average gain from pre- to post-test and the maximum possible gain.[13,16–18] The compromise was that normalising the absolute gain by the maximum gain possible accounts for the fact that some student cohorts have a wider margin for improvement than others do. However, this method vastly favours higher pre-test scores and was primarily designed to compute gains at the group level (classroom/student cohort) with an overall improvement in knowledge. Nevertheless, the calculation method became popular in undergraduate science, technology, engineering and mathematics (STEM) education literature[19] and in medical education.[20]

There is a need for objective measures to quantify the teaching quality in higher education and medical education in particular, such that the calculated scores allow the meaningful comparison and interpretation of results. Although previously published gains were developed to deal with inhomogeneous pre-test levels,[4] little attention has been paid to reducing the systematic distortion in learning gain calculations. Therefore, we introduce a new measure that minimises even further the bias of starting levels at both extremes (low and high pre-test scores).

Given the current lack of knowledge,[21] the aims of our study were to

1. Introduce a new linear weighted gain accounting for individual knowledge baseline with the goal of minimising the pre-test score influence.
2. Compare this method with those already available in view of the dependency on pre-test scores and the raw difference between pre- and post-test scores.
3. Analyse the impact of the different metrics on the resulting learning gain computed from simulated and empirical test data on both the group and individual levels.

2 | MATERIAL AND METHODS

2.1 | Common methods of calculating learning gain

We employed several current methods to calculate learning gain ($G_0 - G_3$) in the literature to compare and contrast the calculated learning gain in both our simulated and empirical data.

In the metrics below, categorised as raw gain (G_0) and relative gains ($G_1 - G_3$), 'pre' represents the pre-test scores and 'post' the post-test scores:

$$G_0 = post - pre \quad (1)$$

$$G_1 = \frac{post - pre}{100\% - pre} \text{ if pre} < \text{post} \quad (2)$$

$$G_2 = \frac{post - pre}{post + pre} \quad (3)$$

$$G_3 = \frac{post - pre}{pre} \text{ if post} < \text{pre} \quad (4)$$

Thereby, G_0 is the absolute difference or raw gain. G_1 represents the normalised gain calculated on the mean group or cohort level.[13] G_2 is the difference in the post and pre-test scores divided by the sum of each mean score (cohort level) and has been described as being symmetric about the mean. It can therefore be termed the symmetric gain.[4] G_3, the normalised change score, was created as a student-level alternative and in addition to G_1.[22] It was proposed to account for the difficulty that arises in the unusual situation in which students' learning gain is negative (post-test score < pre-test score). Thus, the score is scaled by the possible number of points students could have lost.

2.2 | New metric

Here we introduce our new metric, which we label G_4 and refer to as the weighted gain score. This learning gain is based on the non-parametric Kraemer–Andrews estimator.[23] The weighted gain score is defined as the percentage of raw difference in test score multiplied by a weighting coefficient to adjust for pre-test variability. The multiplier serves as a constant linear weighting coefficient transforming results linearly, thus facilitating the comparison of learning gain, both in different student cohorts and interventions.

This new gain G_4 termed as weighted gain is defined as

$$G_4 = (post - pre) * \left(\frac{pre}{\mu}\right) \quad (5)$$

We introduce μ as the expected mean learning gain of the student cohort undergoing a specific intervention; here the value was set as 50, representing 50%.[24] Alternative values for μ may also be assumed, changing the gradient of the linear relationship. In other studies, a predefined target of 30% defined the minimum value at which the intervention could be regarded as effective.[5,13,25]

2.3 | Study design

We compared and contrasted all five metrics. We initially applied the learning gain functions G_0–G_4 to a simulated dataset. This dataset illustrated a theoretical relationship between pre-test scores, differences in test scores (raw points), and the results of the calculated learning gains G_0–G_4, respectively. Secondly, G_0 (raw gain) and G_1(normalised gain) were compared with the new G_4 (weighted gain),

employing an empirical dataset originating from an intervention study (see below).

2.4 | Simulated dataset

We combined different pre-test scores with different raw changes in test performance on a predefined 10-point scale (fictional maximum test score), generating a simulated dataset. Pre-test scores ranged from 1 to a maximum of 9, depending on the simulated improvement or deterioration in performance of between 1 and 4 points, which in turn represented 10% to 40% of the maximum score of 10. To escape ceiling effects, the analysis only included those combinations of pre-test scores and differences in test performance avoiding having scores below the minimum of 0 or above the maximum of 10. The simulated dataset comprised 56 pairs of pre-test/post-test data. Thirty pairs demonstrated an increase in test performance, whereas 26 pairs displayed a decrease.

2.5 | Empirical dataset

The real-world paired test-data were obtained from a previously published prospective intervention study comparing the impact of a traditional lecture with a recorded lecture presented on an e-learning platform.[26] The study ran over two consecutive semesters at the University Medical Centre in Göttingen, Germany, during the teaching module 'Operative Medicine' in the fifth year of the degree course in human medicine. The local institutional review and ethics board judged the project as not representing medical or epidemiological research on human beings. The project was approved without any reservation under the proposal number 1/11/14. A traditional live lecture without explicit interaction on the subject of goitre was held for the first semester and replaced by a matching video-recorded lecture during the second, in which students were able to review and repeat sections ad libitum. The subject of inguinal hernia in the same traditional lecture format and an untaught subject (cholelithiasis) served as controls in both semesters. A multiple-choice pre-test/post-test model of cognitive learning acquisition measured the learning gain. Students also filled out a questionnaire on their personal preferences with respect to information technology, which led to the recognition of two clear clusters that were termed 'traditional learners' versus 'digital natives'. Data from both test scores were linked for each single student. In total, we included matched scores from $n = 180$ students in our analysis, independent of the intervention or controls. Prior to further statistical analyses, we confirmed that the real-world dataset was normally distributed with the Shapiro–Wilk test.[27]

2.6 | Recalculation of previously published findings

In the previous paper, learning gain was calculated as raw gain G_0.[26] In the present study, we recalculated the findings by comparing G_0 with the normalised gain (G_1) and the new gain (G_4) to investigate possible effects on the computation of results.

2.7 | Statistical analysis

IBM SPSS Statistics for Windows version 25.0 (Armonk, NY, USA: IBM Corp.) and R version 3.6.1 (R Foundation for Statistical Computing, Vienna, Austria) were used to perform the analyses. Learning gains G_0–G_4 were computed for the simulated paired data. This procedure was reproduced for the calculation of G_0, G_1, and G_4 for the empirical data. We were thus able to compare the influence of pre-test scores on the calculation of learning gain pairwise.

Exploring the empirical dataset further, correlation analysis (Pearson correlation r) evaluated the relationship between pre-test scores and calculated learning gains G_0–G_4. To investigate the impact of baseline knowledge on calculated learning gain, students were divided into three quartile groups ('poor performers': <25%, 'medium performers': 25%–75% and 'high performers': >75%) on the basis of the raw gain G_0.

We conducted a two-sided analysis of variance (ANOVA) to assess the effects of the pre-test score on the gain calculations and to determine if the results differed statistically significantly on the group level.

3 | RESULTS

3.1 | Learning gain in simulated paired data

To visualise the extent of the variance, G_0 to G_4 were plotted as a function of the pre-test scores (Figure 1a,b). In both cases of increase and decrease in test performance, G_0 had zero slope, with identical gains regardless of the pre-test score. The learning gain G_1 was sensitive to higher pre-test scores and resulted in disproportionally higher gain values. G_2 was moderately influenced as a function of low pre-test scores, whereas G_3 displayed an even stronger bias with over-calculation of gain values. G_2 and G_3 displayed a similar negative slope especially favouring low pre-test scores. The plots of G_1–G_3 were exponential in nature and thus demonstrated highly distorting effects. G_4 plots were linear and thus not disproportionate at either the upper or lower extremes of the pre-test scores. G_4 gradient increased with greater absolute differences between post- and pre-test scores. For negative learning gains (Figure 1b), G_1 markedly overemphasised the decrease in learning gain for high pre-test scores. G_2 and G_3 also overemphasised this decrease, however this time for low pre-test scores. Again, G_4 was linear, the gradient here being negative, indicating negative learning gain.

In summary, all five metrics yielded different results. However, G_4 appeared to be the most robust in terms of reducing bias towards pre-test scores. Furthermore, the linear plot enables improved comparison of different educational interventions.

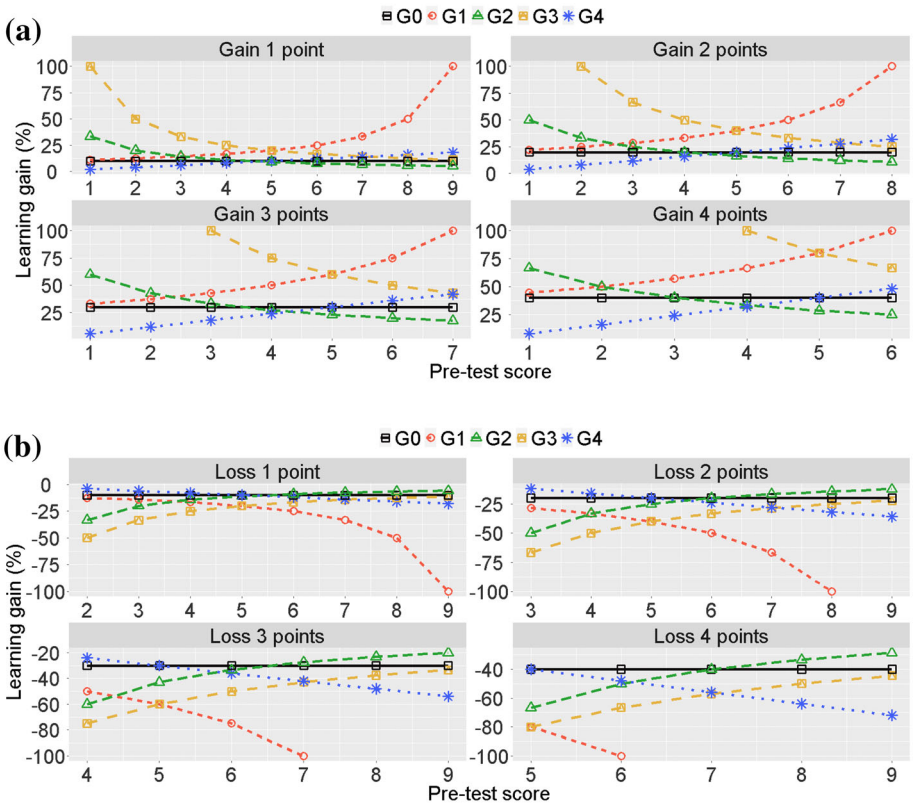

FIGURE 1 (a) Gain calculations G0–G4 as function of the theoretical pre-test scores using the simulated dataset. The simulation ran for a performance improvement (+1 to +4 points). (b) Gain calculations G0–G4 depicted as function of the theoretical pre-test scores using the simulated dataset. The simulation ran for a performance deterioration (−1 to −4 points) [Color figure can be viewed at wileyonlinelibrary.com]

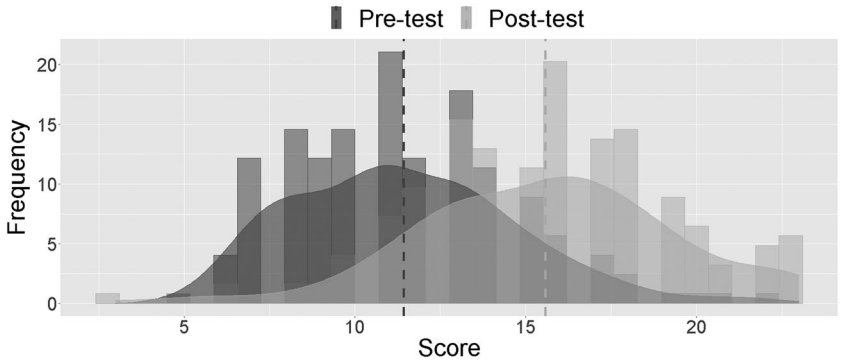

FIGURE 2 Distribution of pre-test (grey) and post-test scores (yellow) in the empirical dataset, reflecting the overall results from the intervention and controls; dashed vertical lines represent mean values

3.2 | Learning gain in empirical pre-test/post-test data

To determine reproducibility, learning gain was re-evaluated in an empirical dataset of 180 students (Figure 2). The Pearson correlation coefficient between pre-test and post-test scores was 0.57 ($p < 0.01$). The calculated mean pre-test score was 11.4 points. Following the teaching formats, the mean post-test score rose to 15.6 points. Thus, participants demonstrated significant increases ($p < 0.01$) in test scores (4.2 raw points of learning gain).

Using all five metrics, the calculated mean learning gain varied vastly, ranging from approximately 14% to 42% ($G_0 = 16.64\%$, $G_1 = 29.91\%$, $G_2 = 15.43\%$, $G_3 = 42.35\%$ and $G_4 = 14.06\%$). Two-sided ANOVA followed by a post-hoc Tukey test confirmed significant differences between all five gains.

Table 1 depicts a correlation coefficient matrix of calculated learning gains and pre-test scores. With respect to pre-test scores, significant negative correlations were determined for calculated learning gain using G_0 to G_3, whereas G_4 was independent. In addition, calculated gains using G_1 to G_4 correlated significantly with G_0. However, the strongest correlation among the equations was found between G_1 and G_4, as they both share the characteristic of overrating gain for students with an already high baseline score.

Taken together, our new G_4 demonstrates two important characteristics: the least dependency on pre-test score and a strong correlation with G_0.

We excluded G_2 and G_3 from further analysis. The latter was originally introduced as a supplement to the normalised gain G_1.[22] Furthermore, G_2 and G_3 both overemphasised learning gain for

	Pre-test scores	G_0	G_1	G_2	G_3	G_4
Pre-test scores	1					
G_0	−0.340**	1				
G_1	−0.202**	0.913**	1			
G_2	−0.491**	0.931**	0.795**	1		
G_3	−0.525**	0.900**	0.731**	0.961**	1	
G_4	−0.062	0.905**	0.979**	0.731**	0.651**	1

TABLE 1 Pearson correlation coefficient matrix of the different learning gain calculations and pre-test scores

**p < 0.01 (two-sided).

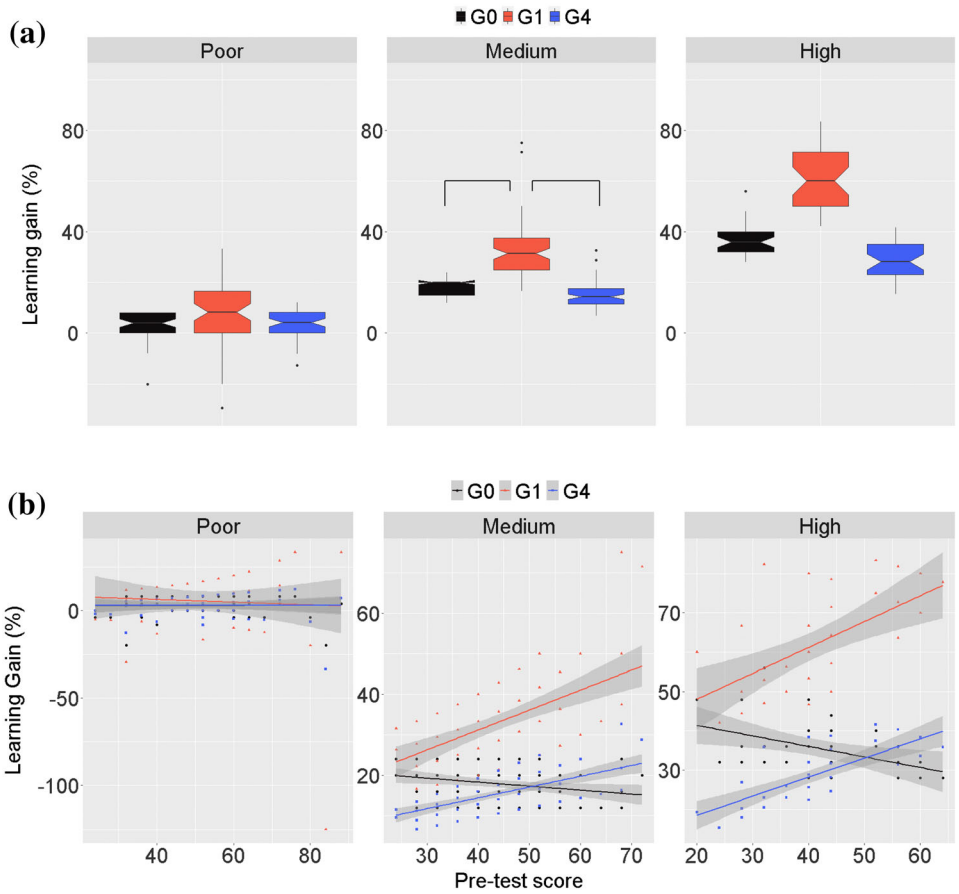

FIGURE 3 Notched boxplots (a) and scatter plot with regression line and 95% confidence interval (illustrated as grey areas) (b) to visualise learning gain in relation to the pre-test scores using *G*0, *G*1 and *G*4 in the three groups of performers (poor, medium, and high), **p < 0.01, ***p < 0.001 [Color figure can be viewed at wileyonlinelibrary.com]

students with low pre-test scores and were not eligible to differentiate performance towards the ceiling effects.

To visualise the differences in calculating learning gain in more detail, we separated the performance of students into three percentile groups based on their raw gain (G_0); G_1 and G_4 were assessed accordingly. The group of 'poor performers' comprised $n = 59$ students; $n = 84$ students were categorised as 'medium performers', with the remaining $n = 37$ students viewed as 'high performers'. Figure 3a summarises performance on the group level. There were no significant differences between the calculated gains in the 'poor' group. In 'medium performers', the gains resulting from methods G_0 and G_4 differed significantly from G_1. The gains calculated by each method (G_0, G_1 and G_4) for 'high performers' differed significantly from the other two, respectively. Figure 3b depicts calculated gains on the individual student level. In 'poor performers', the calculated gains from each method overlapped; there was no statistically significant difference between the metrics, $F(2, 174) = 0.703$, $p = 0.497$. However, the calculated learning gains using G_0, G_1 and G_4 differed significantly for 'medium performers', $F(2, 249) = 151.5$, $p < 0.001$, and was even more pronounced for 'high performers', $F(2, 108) = 132.3$, $p < 0.001$).

These results indicate that the use of weighted gain leads to diverging results, particularly for 'high performers', for whom possible ceiling effects are more likely. In contrast, the use of normalised gain score leads to a dissymmetrical high gain for students with high pre-test scores combined with high performance.

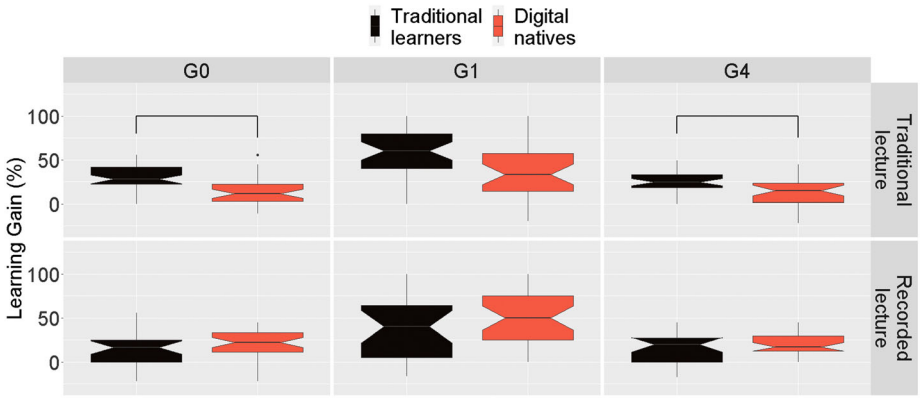

FIGURE 4 Notched boxplots of the data resulting from recalculation of percentage learning gains (G0,G1 and G4) for 'traditional learners' and 'digital natives' in a study on the effect of the teaching formats 'traditional lecture' and 'recorded lecture', *p < 0.05 [Color figure can be viewed at wileyonlinelibrary.com]

3.3 | Recalculation of the learning gain and differences in curricular effectiveness

To determine teaching quality, the gains G_0, G_1 and G_4 were applied to the empirical dataset from the module 'Operative Medicine' (Figure 4). In the former study,[26] there was an interaction effect between students' perceived affinity to new information technologies and method of knowledge delivery implemented. Digital natives were found to be at a significant disadvantage in the traditional live lecture. In the original paper, G_0 was employed to demonstrate the weaker performance of digital natives when compared with traditional learners. In our study, G_4 confirmed this significant difference, $F(1, 131) = 6.514$, $p < 0.05$. However, the interaction effect was lost if G_1 was chosen to determine the intervention effect, $F(1, 129) = 3.72$, $p = 0.056$. This result indicates that learning is not only dependent on the teaching format and characteristics of learners, but its measurement is highly sensitive to the method ultimately chosen to calculate learning gain.

4 | DISCUSSION

The outcome-based paradigm is an important educational model. Decisions concerning curriculum and teaching are driven by several student outcome parameters.[28] It is often difficult to prove the direct benefits of an intervention on knowledge delivery or even improvement in healthcare. Therefore, identifying teaching quality using objective measures of student learning gain is of major interest and importance. Of note, the burden of different competency curves also has to be considered, as knowledge baselines for individual students are inhomogeneous. The concept of learning gain is widely used in evaluating teaching methods in higher education.[29] This study used a pre-test/post-test design to assess the effects of different learning gain calculations. Aiming to compare knowledge gain across cohorts and semesters, we used percentages instead of actual test points in the metrics.

4.1 | Determination of learning using raw and relative gains

Common methods to calculate learning gain include computation of simple differences. Several authors name the raw gain G_0 as an unbiased estimate of the underlying true change.[30] However, some debate in the statistics community remains on the ability to measure change accurately. It is common knowledge that pre-test scores highly influence the calculated raw gain. Students with very high test scores have 'no room to grow', so changes in learning gain for teachers with 'high performers' will be depressed, even for highly effective teachers,[31] owing to the ceiling effect.[32] On the other hand, students with low pre-test scores have a wider margin for improvement than those with high pre-test scores. The limited use of G_0 is thus evident.

Hence, the relative gains (G_1, G_2 and G_3) were designed to overcome the difficulties associated with pre-test bias. Their scores set the absolute difference between pre- and post-test scores in relation to several reference variables or terms. Although these methods are popular in higher education, the calculated gains tend to distort students' knowledge systematically by overrating low pre-test levels (G_2 and G_3) or high pre-test levels (G_1) within their correction for pre-test bias. Mathematically, relative gains lead to exponential overemphasis at the extremes of the starting values. Whether this characteristic of the metric is precise depends on the educational scenario and whether student achievement is expected to range towards the middle of the score distribution.[31]

The most popular method, normalised gain (G_1), is known to yield disproportionately high outcomes in learning gain for high pre-test scores.[13,20] It distinguishes students at the very high (or low) end of the achievement spectrum poorly. Effects of this skewness on outcome interpretation have already been noted.[22] G_1 is not applicable, especially under two conditions: the pre-test score equals the maximum possible gain or when the pre-test score is higher than the post-test score (loss in performance). As such, the gain metric G_2 was specifically designed as an expansion of G_1, in order to overcome this negative-learning-gain limitation.

4.2 | Determination of learning using a weighted gain

Given the limitations of already published metrics, we proposed a new approach to calculating learning gain. We introduce a linear weighting coefficient that serves to minimise pre-test bias. The expected mean value μ of the weighting coefficient may be adapted to the pass grade of any test depending on difficulty. In our study, it was set to 50%.[24] Similar to the normalised gain metric G_1, our new metric G_4 decreases for low pre-test scores and raises scores for students with high pre-test scores.

However, G_4 was the only gain without any significant correlation with the pre-test scores in the empirical data. Although there were similarities and a high correlation between G_1 and G_4, the weighting effect of G_4 acts linearly, thus reducing skewness in computed results. In more detail, G_4 plots were linear and exhibited fewer disproportionate effects at either the upper or lower extremes of the pre-test scores. The gradient of G_4 increased with greater absolute differences between post- and pre-test scores; it was positive for improvements in performance and negative for drops in performance. Comparing G_1 with G_4, the distinguishing feature of our new weighted metric is that its use in negative learning gain is not limited mathematically.

The characteristics of the new metric G_4 presented here may improve comparison of the teaching quality of different interventions and learning gain between different study cohorts. Particularly in medical education, the outcome-based assessment of teaching activities covering a multitude of subjects taught during different semesters, inhomogeneous levels of baseline knowledge and divergence in students' abilities along the competency curves must be controlled at best before teaching quality at course level or even the entire curriculum can be judged. This is central to the performance-oriented allocation of financial resources at medical schools and funding in medical-education-based project proposals.[33,34] Moreover, the new metric G_4 does not depend on a normal distribution. The latter is important, because empirical test data from student cohorts often demonstrate *contamination* of a *symmetric* distribution[35] and parametric effect sizes such as Cohen's *d* lead to different outcomes.[36]

4.3 | Exploring learning gain differences in a teaching module

Reassessing previously published test data to measure teaching quality within the module 'Operative Medicine' revealed different results. G_0 was set as the original method, which we compared with G_1 and G_4 here. Using G_0, the authors found a significant interaction effect between students' affinity to information technologies and their benefit from a video-recorded versus traditional live lecture. This effect was confirmed by our new metric G_4. However, the distorting effect of G_1 was obvious in our dataset, and the interaction effect we discovered in the original work would never have been found as a result of implementing G_1. Thus, G_1 is prone to confounding findings as a result of its skewness.

One has to consider carefully the method of calculating learning gain, as deviating results may shed a totally different light on the outcome. Using G_1, with its particular shortcomings, may have led to a misleading conclusion of teaching quality in the context of new information technologies. This finding clearly illustrates the problem associated with selecting the appropriate metric and contributes to the discussion on how best to interpret test data in view of known test bias.[9]

4.4 | Limitations of the study

We used an empirical dataset from a pre-/post-test design and compared three different metrics on both group and individual levels, thereby also determining the impact of knowledge baselines. The value of this design as predictor or indicator of the underlying true change and its reliability has been discussed widely.[10,37,38] The pre-/post-test design allows a pragmatic approach to measuring the difference between two points in time, representing a period of change in knowledge. However, other scenarios using two reference points may also be considered. For example, instead of tests demonstrating knowledge, students might also estimate their level of competence based on learning objectives related to two points: rating of the present level (after completion of the teaching activity) and the retrospective level (before).[20] Of note, the approach of using two points or references is independent of the overall study design. To enhance objectivity, gains are usually calculated separately for the intervention as well as for the control group, as we did to compare the effectiveness of teaching formats employing the empirical data set.

Our teaching module data reflected a moderate increase in knowledge and no distinctive ceiling effect. Only a few participants demonstrated a high baseline knowledge (e.g. only three students achieved more than 80% in the pre-test) and none started on the maximum possible score. We were thus unable to demonstrate any distortion effects of the calculations at the upper extreme. No student attained the maximum score post-test. A dataset specifically including high starters as well as high achievers may prove more suitable to compare and contrast the metrics G_1 and G_4.

Using the empirical data, no significant correlation between pre-test scores and the metric G_4 was demonstrable on group level; however, a tendency was detectable. Furthermore, the pre-test score impacted significantly when employing the simulated data. This observation indicates that G_4 was highly capable of reducing the influence of pre-test scores; however the new metric was not completely independent of baseline knowledge.

5 | CONCLUSION

Given the distorting effects that various published learning gain calculations present, we suggest exercising caution when implementing current metrics to draw conclusions on teaching quality. Hence, we introduce a new linear weighted gain, developed to provide a realistic

and accurate method to document learner performance. We believe this new metric enables more meaningful comparison between educational interventions. Mathematically speaking, the metric is easy to calculate and broadly applicable on a routine level for anyone engaged in quality assurance and course/curriculum development. Its regular use can improve the transparency of teaching activities and outcomes, contribute to forming rounded judgements about students' acquisition of knowledge and skills and enable valuable feedforward to develop and enhance curricular concepts.

Further studies may confirm and build on our findings. The questions remain as to whether statistical analysis based on probabilistic test theory is superior to the classic gain calculations presented here. There are Rasch models that can be used to model learning gain in pre- and post-test settings directly.[39] This analysis of learning gain allows direct correlation between the improvement in knowledge and teaching content through calibrating the test items and students' abilities on the same scale. Applying such models to our data will be the subject of future research.

ACKNOWLEDGMENTS

We would like to express our deepest gratitude to all the students who participated in this study. Furthermore, we would like to thank Andrew Entwistle for providing comments on the draft version and his assistance with proofreading the manuscript.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CONFLICT OF INTEREST

The authors disclose that there are no conflicts of interest.

AUTHOR CONTRIBUTIONS

All authors were involved in the form and/or study design and contributed critically to the final preparation of this article, including approving the final version of the manuscript. In particular, S. K. conceived and designed the study, wrote the final study protocol and drafted the manuscript. S. W. ran the study, collected the results, analysed the data and developed the weighted gain metric. J. B. analysed the data and performed and verified the statistical analyses.

ETHICS STATEMENT

The empirical data analysed within this work was reviewed and judged by the local institutional review and ethics board as not representing medical or epidemiological research on human subjects and as such adopted a simplified assessment protocol. The project was approved without any reservation under the proposal number 1/11/14.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Silke Westphale https://orcid.org/0000-0002-8913-320X

REFERENCES

1. Spreckelsen C, Juenger J. Repeated testing improves achievement in a blended learning approach for risk competence training of medical students: results of a randomized controlled trial. *BMC Med Educ*. 2017;17(1):1-177. https://doi.org/10.1186/s12909-017-1016-y
2. Polkinghorne M, Roushan G, Taylor J. Considering the marketing of higher education: the role of student learning gain as a potential indicator of teaching quality. *J Mark High Educ*. 2017;27(2):213-232. https://doi.org/10.1080/08841241.2017.1380741
3. Cook DA, Beckman TJ. Reflections on experimental research in medical education. *Adv Health Sci Educ Theory Pract*. 2010;15(3):455-464. https://doi.org/10.1007/s10459-008-9117-3
4. Willoughby SD, Metz A. Exploring gender differences with different gain calculations in astronomy and biology. *Am J Phys*. 2009;77(7):651-657.
5. Colt HG, Davoudi M, Murgu S, Rohani NZ. Measuring learning gain during a one-day introductory bronchoscopy course. *Surg Endosc*. 2011;25(1):207-216.
6. Mellenbergh GJ, van den Brink WP. The measurement of individual change. *Psychol Methods*. 1998;3(4):470-485.
7. Williams B, Onsman A, Brown T. Exploratory factor analysis: a five-step guide for novices. *Australas J Paramedicine*. 2010;8(3):1-13.
8. Udovic D, Morris D, Dickman A, Postlethwait J, Wetherwax P. Workshop biology: demonstrating the effectiveness of active learning in an introductory biology course. *Bioscience*. 2002;52(3):272-281.
9. Brogt E, Sabers D, Prather EE, Deming GL, Hufnagel B, Slater TF. Analysis of the astronomy diagnostic test. *Astron Educ Rev*. 2007;6(1):25-42.
10. Cronbach LJ, Furby L. How we should measure "change": or should we? *Psychol Bull*. 1970;74(1):68-80.
11. Hake R. Should we measure change? Yes. In: *Evaluation of Teaching and Student Learning in Higher Education, New Directions in Program*. American Evaluation Association; 2010.
12. Meltzer DE. Normalized learning gain: a key measure of student learning. 2002.
13. Hake RR. Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am J Phys*. 1998;66(1):64-74.
14. Maier-Riehle B, Zwingmann C. Effektstärkevarianten beim eingruppen-prä-post-design: eine kritische betrachtung. *Rehabilitation*. 2000;39(04):189-199.
15. Bonate PL. *Analysis of Pretest–Posttest Designs*. Chapman and Hall/CRC; 2000.
16. Raupach T, Schiekirka S, Münscher C, et al. Implementierung und erprobung eines lernziel-basierten evaluationssystems im studium der humanmedizin. *GMS Z Med Ausbild*. 2012;29(3):1-14.
17. Hake RR. Design-based research in physics education: a review. *Diakses Pada Tanggal*. 2007;26:8-9.
18. Prather EE, Rudolph AL, Brissenden G, Schlingman WM. A national study assessing the teaching and learning of introductory astronomy. Part I. The effect of interactive instruction. *Am J Phys*. 2009;77(4):320-330.
19. Hake RR. Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization. 2002:1-14.
20. Schiekirka S, Reinhardt D, Beibarth T, Anders S, Pukrop T, Raupach T. Estimating learning outcomes from pre-and posttest student self-assessments: a longitudinal study. *Acad Med*. 2013;88(3):369-375.
21. Pickering JD. Measuring learning gain: comparing anatomy drawing screencasts and paper-based resources. *Anat Sci Educ*. 2017;10(4):307-316.
22. Marx JD, Cummings K. Normalized change. *Am J Phys*. 2007;75(1):87-91.
23. Kraemer HC, Andrews G. A nonparametric technique for meta-analysis effect size calculation. *Psychol Bull*. 1982;91(2):404-412.

24. Higham DJ. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Rev*. 2001;43(3):525-546.
25. Prather EE, Rudolph AL, Brissenden G. Teaching and learning astronomy in the 21st century. *Phys Today*. 2009;62(10):41-47.
26. 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.
27. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika*. 1965;52(3/4):591-611.
28. Harden RM. AMEE guide no. 14: outcome-based education: part 1—an introduction to outcome-based education. *Med Teach*. 1999;21(1):7-14.
29. McGrath CH, Guerin B, Harte E, Frearson M, Manville C. *Learning Gain in Higher Education*. Santa Monica, CA: RAND Corporation; 2015.
30. Schiekirka S, Anders S, Raupach T. Assessment of two different types of bias affecting the results of outcome-based evaluation in undergraduate medical education. *BMC Med Educ*. 2014;14(1):1-9.
31. Resch A, Isenberg E. How do test scores at the ceiling affect value-added estimates? *Statistics and Public Foreign Policy*. 2018;5(1):1-6. https://doi.org/10.1080/2330443X.2018.1460226
32. Šimkovic M, Träuble B. Robustness of statistical methods when measure is affected by ceiling and/or floor effect. *PLoS ONE*. 2019;14(8):e0220889. https://doi.org/10.1371/journal.pone.0220889
33. Gottlieb M, Lee S, Burkhardt J, et al. Show me the money: successfully obtaining grant funding in medical education. *West J Emerg Med*. 2019;20(1):71-77. https://doi.org/10.5811/westjem.2018.10.41269
34. Brähler E, Strauss B. [Performance-oriented allocations of financial resources at medical schools: an overview]. Leistungsorientierte mittelvergabe an medizinischen fakultäten: eine aktuelle ubersicht. *Bundesgesundheitsbl Gesundheitsforsch Gesundheitsschutz*. 2009;52(9):910-916. https://doi.org/10.1007/s00103-009-0918-1
35. Micceri T. The unicorn, the normal curve, and other improbable creatures. *Psychol Bull*. 1989;105(1):156-166.
36. Hedges LV, Olkin I. Nonparametric estimators of effect size in meta-analysis. *Psychol Bull*. 1984;96(3):573-580.
37. Marsden E, Torgerson CJ. Single group, pre- and post-test research designs: some methodological concerns. *Oxf Rev Educ*. 2012;38(5):583-616.
38. Dimitrov DM, Rumrill PD Jr. Pretest–posttest designs and measurement of change. *Work*. 2003;20(2):159-165.
39. Pentecost TC, Barbera J. Measuring learning gains in chemical education: a comparison of two methods. *J Chem Educ*. 2013;90(7):839-845.

How to cite this article: Westphale S, Backhaus J, Koenig S. Quantifying teaching quality in medical education: The impact of learning gain calculation. *Med Educ*. 2022;56(3):312-320. doi:10.1111/medu.14694