 doi:10.20944/preprints202211.0338.v1

Communication

Assigning fair and defensible grades to clinical interns

Brian Foutch [1,*], Amy Ferguson [1,*], Russell Coates [1], Yutaka Maki [1], and Janet Summers [1]

[1] Rosenberg School of Optometry, University of the Incarnate Word, 4301 Broadway, CPO #31, San Antonio TX, USA; coates@uiwtx.edu (R.C.), maki@uiwtx.edu (Y.M.), jesummer@uiwtx.edu (J.S.)
* Correspondence: foutch@uiwtx.edu (B.F.); aferguso@uiwtx.edu (A.F.)

Abstract: Background: Grades in clinical courses matter. They are often used to determine clinical academic awards, scholarships, and—most importantly—interns' suitability for graduate medical education opportunities. Aware of these stakes, clinic preceptors may feel pressure to grade too leniently or uniformly. A fair method of adjusting for differences in preceptor bias is then needed. Approach: The authors propose a technique that employs the advantages of both criterion- and normative-based grading to adjust for differences in both grader leniency and uniformity. Evaluation: The technique produces fair adjustments to any raw assign grades, and the authors demonstrate how easily this process can be administered in any clinical setting where multiple preceptors are evaluating interns. Implications: This work provides a grading framework that is transparent to all stakeholders but places responsibilities at the appropriate level. That is, clinic performance evaluations are left to clinic preceptors but grading to clinic academic managers.

Keywords: normative grading; criterion-based grading; clinic evaluations; clinic interns

1. Introduction

Those who assign grades in clinical courses know the challenges very well. First, there is no consensus regarding the value of assigning letter grades [1-8], but students care about and are motivated by grades [2-3]. However, grades are mostly cultural, can be psychologically harmful, and can inhibit learning [4-6]. For example, an important aspect of healthcare is teamwork [7] which may be discouraged by sorting students by merit [8]. Second, students want grades to be a fair representation of their work [9], but academic managers want grades that predict subsequent success [10]. The third challenge is transforming evaluations from multiple clinic preceptors in various settings into a final course grade.

To meet these challenges, we developed a hybrid (criterion- and normative-based) technique that numerically transforms raw numerical evaluations of clinical performance into adjusted grades using the individual and overall grader mean and standard deviations. Our experience suggests this technique fairly adjusts for most differences in preceptor biases toward grading and may prove useful in most training contexts.

2. Approach

2.1. Individual evaluations

Both third- and fourth-year optometry interns at our in-house clinics receive a single grade per day from their assigned clinic preceptor. We have observed that some preceptors are unwilling to grade students differently, limiting the spread of their grades. In addition, some preceptors are biased toward grading leniently or stringently. To combat these, all interns and supervising preceptors are supplied with scoring criteria. We use five grading dimensions (Table 1), and preceptors assign either *exceeds*, *meets*, or *below expected* for each dimension and enter the grades into a customized commercial medical education system (Meditrek®, HSoft Corporation, Morrisville, PA). If an intern *meets expected* levels for all dimensions, their grade is 85%. For each individual dimension, intern grades are adjusted up (if exceeds) or down (if below) for that dimension. These adjustments are shown in the headers in Table 1.

Table 1. Scoring criteria available to all clinic preceptors and interns

Intern level	***Speed** Exceeds: +2 pts Below: -7 pts	***Skills** Exceeds: +4 pts Below: -12 pts	**Knowledge/Analysis** Exceeds: +4 pts Below: -12 pts	***Communication** Exceeds: +3 pts Below: -10 pts	**Charting/Billing** Exceeds: +2 pts Below: -7 pts
Level 1; 3rd year; Summer	- All minimum elements completed within 90-120 minutes	- Accurately performs all core exam elements *without guidance*	- Knows normal/expected values; Recognizes/describes abnormal findings; formulates plans for low complexity conditions *with guidance*	- Documents all abnormal conditions - Accurately communicates exam findings to attending - Communicates plan to patient *with guidance*	- Charting and billing are complete and accurate *with limited guidance*
Level 2; 3rd year; Fall	- All minimum elements completed within 80-100 minutes	- Accurately and efficiently performs all core exam elements *without guidance* - Refines clinical history based on exam findings	- Develops tentative diagnoses from case history; Selects appropriate additional tests from history and/or exam findings; Identifies and formulates plan for low/moderate complexity conditions *with limited guidance*	- Efficiently documents all abnormal conditions - Efficiently communicates exam findings to attending - Communicates plan to patient *with very limited guidance*	- Charting and billing are complete and accurate *without guidance*
Level 3; 3rd year; Spring	- All minimum elements completed within 60-90 minutes	- Develops accurate tentative diagnoses from clinical history - Accurately performs additional tests *without guidance*	- Prioritizes tentative diagnoses; Selects appropriate additional tests from history or exam findings; Formulates plans for low/moderate complexity conditions *without guidance*	- Effectively communicates exam findings to attending; Communicates plan to patient *without guidance*; Incorporates patient education into treatment plan *with guidance*	- Charting and billing are complete and accurate *without guidance*
Level 4; 4th year; Summer through first half of Fall	- All minimum elements completed within 50-75 minutes	- Accurately performs all core exam elements *without guidance* - Efficiently performs additional tests *without guidance*	- Knows normal/expected values - Recognizes/describes abnormal findings - Formulates plans for low complexity conditions *with guidance*	- Efficiently documents findings; Effectively communicates exam findings to attending; Communicates plan to patient *without guidance* - Incorporates patient education into treatment plan *with very limited guidance*	- Charting and billing are complete and accurate *without guidance*
Level 4; 4th year; Second half of fall through Spring	- All minimum elements completed within 45-60 minutes	- Accurately performs all core exam elements *without guidance* - Efficiently performs additional tests *without guidance;* Identifies visual needs of non-communicative patient (*if applicable*)	- Formulates plans for low complexity conditions *with guidance* - Improvises examination appropriate to exam complexity; Independently formulates diagnosis list and plan (to include coordination of care and treatment)	- Effectively communicates exam findings to attending - Communicates plan to patient *without guidance* - Incorporates patient education into treatment plan *without guidance*	- Charting and billing are complete and accurate *without guidance*

* Assumes cooperative/communicative patient

2.2. Arithmetic adjustments

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 November 2022 doi:10.20944/preprints202211.0338.v1

Our technique to adjust grades is simple and best demonstrated by a sample of hypothetical grades (Table 2). We have only described a few *types* of graders, but they are sufficient to demonstrate the technique. Grades given by preceptors to each student for each day in clinic are shown as "raw" grades and are transformed into standardized (z-) scores by Z = (raw grade - M)/SD, where M = preceptor mean and SD = preceptor standard deviation. The z-scores are subsequently transformed back into adjusted grades by the reverse by using the grand mean (GM; 88.32) and the overall standard deviation (OSD; 6.33). Therefore, the adjusted grade = GM + z-score x OSD.

Table 2. Hypothetical grade data sorted by graders.

Grader:	Student:	Raw grades:	Mean ± SD:	Z-scores:	Adjusted grades:
1 – Lenient, Equal	Student-1	98		-0.447	85.49
1 – Lenient, Equal	Student-1	98		-0.447	85.49
1 – Lenient, Equal	Student-1	98	98.40 ± 0.89	-0.447	85.49
1 – Lenient, Equal	Student-1	100		1.789	99.64
1 – Lenient, Equal	Student-1	98		-0.447	85.49
2 - Lenient, Varied	Student-4	95		1.467	97.61
2 - Lenient, Varied	Student-4	92		0.550	91.80
2 - Lenient, Varied	Student-2	88	90.20 ± 3.27	-0.673	84.06
2 - Lenient, Varied	Student-2	87		-0.978	82.13
2 - Lenient, Varied	Student-2	89		-0.367	86.00
3 - Average, Equal	Student-1	87		0.447	91.15
3 - Average, Equal	Student-1	87		0.447	91.15
3 - Average, Equal	Student-5	85	86.60 ± 0.89	-1.789	77.00
3 - Average, Equal	Student-5	87		0.447	91.15
3 - Average, Equal	Student-5	87		0.447	91.15
4 - Stringent, Varied	Student-2	78		-0.863	82.86
4 - Stringent, Varied	Student-2	76		-1.274	80.26
4 - Stringent, Varied	Student-3	85	82.20 ± 4.87	0.575	91.96
4 - Stringent, Varied	Student-3	87		0.986	94.56
4 - Stringent, Varied	Student-3	85		0.575	91.96
5 - Stringent, Equal	Student-5	83		-1.095	81.39
5 - Stringent, Equal	Student-5	83		-1.095	81.39
5 - Stringent, Equal	Student-4	85	84.00 ± 1.10	0.730	92.94
5 - Stringent, Equal	Student-4	85		0.730	92.94
5 - Stringent, Equal	Student-4	85		0.730	92.94

To fully understand the technique, it is also useful to show a few examples. When calculating z-scores, one is forcing a normal distribution (i.e., Gaussian bell-shaped curve) to fit observed data. Figure 1 shows normal distributions fit to data for Graders 1 and 4. Average grades are 98 and 82 for graders 1 and 4, respectively, and transform to z-scores of 0.0. The adjusted grade is then taken from the black curve which represents a normal distribution fit to all raw grades given by all graders. A z-score of 0.0 is then assigned the average grade of all graders (i.e., grand mean; 88.3).

Normal distributions for graders 2 and 5 are shown in Figure 2. For graders 2 and 5, grades of 87 and 83, respectively, both transform to z-scores of -1.10. A z-score of -1.10 is assigned an adjusted grade of 1.10 standard deviations lower than the grand mean of all graders (i.e., 88 – 1.1 x 6.33 = 81.4).

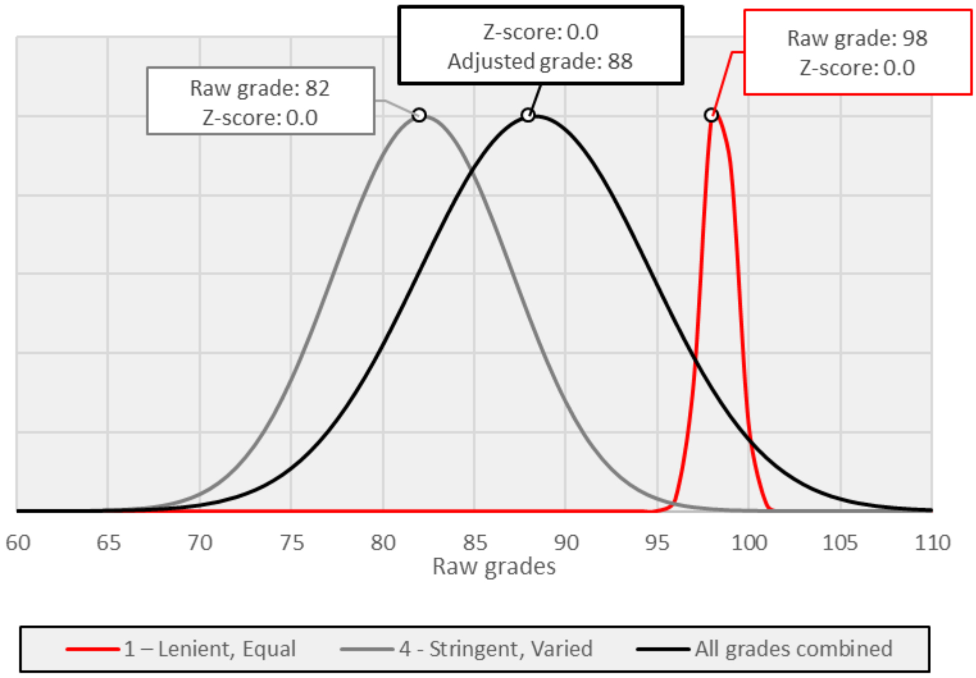

Figure 1. Normal distributions fit to observed data for a lenient grader with limited spread in grades (i.e., Grader 1; red line) and a more stringent grader with more spread in grades (i.e., Grader 4; gray curve). For graders 1 and 4, average grades of 98 and 82, respectively, both transform to z-scores of 0.0. The adjusted grade is then taken from the black curve which represents a normal distribution fit to all raw grades given by all clinic graders. A z-score of 0.0 is the average grade (i.e., grand mean) of all graders (i.e., 88.3 for our hypothetical data).

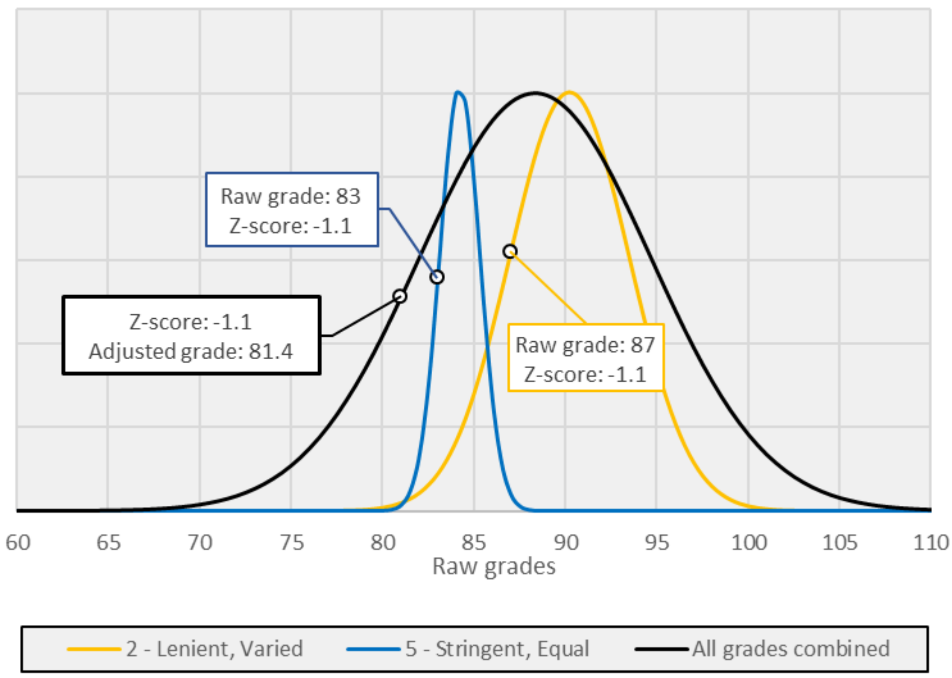

Figure 2. Normal distributions fit to observed data for a lenient grader with more spread in grades (i.e., Grader 2; orange line) and a more stringent grader with less spread in grades (i.e., Grader 5; blue curve). For graders 2 and 5, grades of 87 and 83, respectively, both transform to z-scores of -1.10. The adjusted grade is then taken from the black curve which represents all clinic graders. A z-score of -1.10 represents a grade 1.10 standard deviations lower than the overall average grade (i.e., grand mean = 88) of all graders (i.e., 88 – 1.1 x 6.33 = 81.4).

3. Evaluation of technique

As the grand means of the raw and adjusted grades will always be equal, mean comparisons are meaningless. To evaluate the individual differences between techniques, we first found it useful to compare raw and adjusted grades graphically (shown in Figure 3). While the means are equal, the raw and adjusted grades are not highly correlated (r = .318, p = .121).

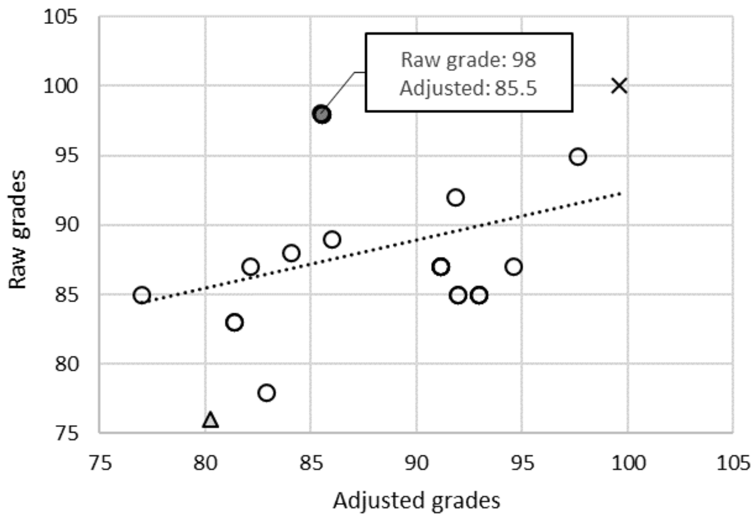

Figure 3. Scatterplot of raw grades vs. adjusted grades. The filled circle (labeled) and 'x' represent grades (98 and 100, respectively) given by the hypothetical lenient grader with a limited spread in grades (i.e., Grader 1 shown in Table 2). The filled triangle represents a raw grade of 76 given by a stringent, varied grader (i.e., Grader 4 from Table 2).

This scatterplot is useful in identifying outliers such as Grader 1's raw grade of 98 represented by the filled circle. While it is a high raw grade, a grade of 98 is 0.447 standard deviations *below average* for Grader 1 and therefore adjusts to a grade 0.447 overall standard deviations below the grand mean (i.e., 88.32 – 0.447 x 6.33 = 85.49; see Table 2 and Figure 1). In effect, grader 1 is saying "I gave Student 1 all 98s, but they are slightly below average for me." However, Grader 1's raw grade of 100 (represented by 'x') hardly adjusts; it is approximately two standard deviations (Z = 1.789) above Grader 1's mean and adjusts to a grade 1.789 overall standard deviations (OSDs) above the grand mean (i.e., 88.32 + 1.789 x 6.33 = 99.64; see Table 2).

In our simple plot of raw v. adjusted grades, a raw grade of 76 given by the stringent, varied grader (i.e., Grader 4 from Table 2) adjusts to 80.3 and appears to be an outlier. However, it is a low grade for Grader 4 that adjusts to a low grade overall. This is a problem with comparing raw and adjusted methods using a simple scatter plot. The comparisons may be better addressed by plotting the grade differences (i.e., raw – adjusted) against the averages of the techniques (i.e., [raw + adjusted]/2). These are plotted against each other in Figure 4 according to a technique described by Bland & Altman [11]. In Bland-Altman (B-A) plots, lines representing the mean difference—as well as upper and lower bounds of acceptable differences—are plotted. Any group of standardized scores will always have a mean z-score equal to 0, and the final step of our adjustment technique is to multiply z-scores by the overall standard deviation then add the grand mean. Therefore, in our technique, the mean of all adjusted grades will be equal to the grand mean of the raw grades (which is 88.3 in our hypothetical data). The mean difference line will then *always* be along the x-axis (i.e., mean differences = 0). While Bland & Altman do not prescribe the upper and lower bounds, it is suggested that all points should lie within the 95% confidence interval (CI) of the mean difference. All our points lie within these bounds. The B-A plot can also highlight whether raw or adjusted grades are systematically too

 doi:10.20944/preprints202211.0338.v1

high or low. If so, then all points would be above (if raw > adjusted) or below (if raw < adjusted) the zero line. Since the points on the sample B-A plot in Figure 4 are scattered equally above and below the x-axis, it suggests no consistent bias of either method.

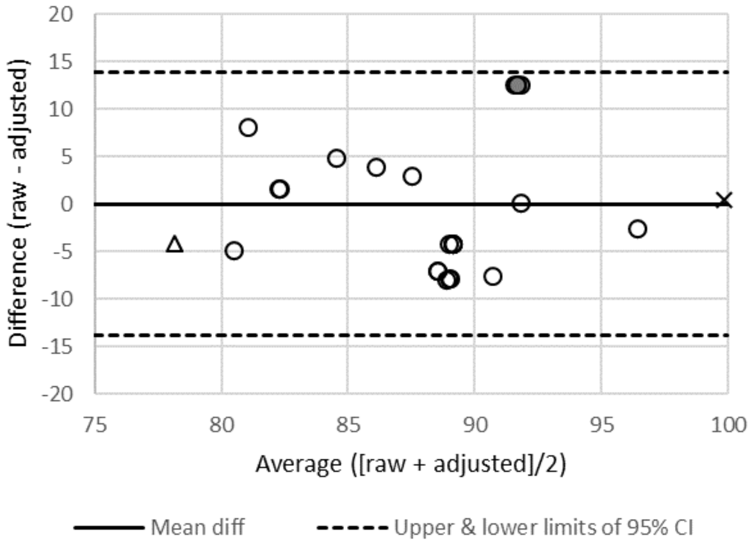

Figure 4. Bland-Altman plot comparing raw and adjusted grades. The data appear equally scattered above and below the mean difference line, which indicates that neither raw nor adjusted methods are biased toward high or low grades. The filled circle, 'x', and filled triangle represent the same points as in Figure 3.

There are other extreme cases not shown in our hypothetical data that could arise and are worth mentioning. Many of these cases would involve preceptors who give a limited spread of mediocre (i.e., "your work is adequate") grades. A simple example of this to handle is a grader who awards the same grade to every student. In these cases, z-scores are 0.0, and—regardless of the preceptor mean—all their grades are adjusted to the grand mean for all grades. However, there are, in our experience, graders who award mostly the same grade but a limited number of generous or below average grades. For example, assume a preceptor awards fifty grades of 85, one grade of 70, and one grade of 100. The mean ± SD = 85.00 ± 2.29, and z-scores for the two "exceptional" grades are Z(70) = (70 - 85.00)/2.29 = -6.56 and Z(100) = (100 - 85)/2.29 = 6.56. Assume—by example—our grand mean (GM) of 88.32 and overall SD (OSD) of 6.33, and these grades adjust to adjusted(70) = z(70) x OSD + GM = (-6.56 x 6.33) + 88.32 = 46.80 and adjusted(100) = (6.56 x 5) + 88 = 129.84.

In our experience, the technique produces fair adjustments to any raw grade, but it may create a perception of norm-referenced grading if the criteria are not properly and/or consistently applied. Barriers to doing so remain the subject of future investigations.

4. Implications

Overall, clinical education is less defined and organized than classroom education [10]. Interns are faced with managing many different conditions of variable complexities, and it makes grading their efforts challenging.

While our technique does not eliminate all the challenges of assigning grades in clinic courses, managing extreme grades is the largest impact. While some extreme graders are stringent graders, most graders are, in our experience, lenient graders. Leniency may be influenced by how "private" clinic grading is or how much they delay grading. One study revealed that individual graders were more likely to forget details than groups, but groups were more lenient graders after that same delay [12]. However, in our clinical settings, students work with an individual grader, so this source of leniency is unavoidable. One adaptable or modifiable source of leniency may be lack of training, as Ogden et al.

[13] demonstrate that training shifts preceptors toward more stringent, criterion-based grading. They also suggest another probable cause of grade inflation—the tension between the mentor and grader roles of clinic faculty. That is, clinical faculty struggle with the shift from advocate during patient care encounters to judge following the encounters. However, we believe this is one of the strengths of our technique. It allows clinic faculty to provide honest, legitimate, and useful feedback to interns but puts the responsibility of calculating the final grade on the course coordinators.

5. Conclusions

In summary, the variability of clinical encounters makes assigning final grades to student clinicians particularly challenging. While we cannot draw any inferences about clinic grading at all other institutions or training settings, we do present a reproducible method of converting clinic performance evaluations into fair and defensible grades. The technique adjusts for between- and within-grader differences in leniency and uniformity, while interns still receive criterion-based preceptor feedback they can use to improve their clinical performance. We challenge fellow clinic academic managers to honestly examine their grading practices and consider a similar technique.

Author Contributions: Conceptualization, B.F. and Y.M.; methodology, B.F., A.F. and Y.M.; formal analysis, B.F., A.F., R.C. and Y.M.; resources, B.F., A.F. and J.S.; writing—original draft preparation, B.F. and A.F.; writing—review and editing, A.F., R.C., Y.M. and J.S.; project administration, B.F. All authors have read and agreed to the published version of the manuscript.

Funding: This investigation received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Angelo, T.A. Doing assessment as if learning matters most. *AAHE Bulletin* **1999**, *51*, 3-6.
2. Baer, J.; Kaufman, J.C. Creativity research in English-speaking countries. In *The International Handbook of Creativity.* Kaufman, J.C., Sternberg, R.J. (Eds.), Cambridge University Press: Cambridge, UK, 2006.
3. Walvoord, B.E.; Johnson Anderson, V. *Effective grading: a tool for learning and assessment in college*, 2nd ed.; John Wiley & Sons, Inc.: San Francisco, CA, USA, 2010.
4. Romanowski, M.H. Student obsession with grades and achievement. *Kappa Delta Pi Record* 2004;40(4):149-151. https://doi.org/10.1080/00228958.2004.10516425
5. Kohn, A. *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes.* Houghton Mifflin Harcourt: New York, NY, USA , 1999.
6. Schinske, J.; Tanner, K. Teaching More by Grading Less (or Differently). *CBE-Life Sci. Ed.* **2014**, *13*, 159-166. https://doi.org/10.1187/cbe.CBE-14-03-0054.
7. Rosen, M.A.; DiazGranados, D.; Dietz, A.S.; Benishek, L.E.; Thompson, D.; Pronovost, P.J.; et al. Teamwork in healthcare: Key discoveries enabling safer, high-quality care. *Am Psychol* **2018**, *73*, 433–450. https://doi.org/10.1037/amp0000298.
8. Nilson, L.B. Grading student assessments. In *Teaching at its best: a research-based resource for college instructors*, 4th ed.; Nilson, L.B., Ed.; John Wiley & Sons, Inc.: San Francisco, CA, USA, 2016; pp. 303-316.
9. Suskie, L.A. *Assessing Student Learning: A Common Sense Guide*, 2nd ed.; John Wiley & Sons, Inc.: San Francisco, CA, USA, 2009.
10. Durning, S.J.; Hemmer, P.A. Commentary: grading: what is it good for? *Acad. Med.* **2012**, *87*, 1002-1004. https://doi.org/10.1097/ACM.0b013e31825d0b3a.
11. Bland, J.; Altman, D. Measuring agreement in method comparison studies. *Stat. Methods Med. Res.* **1999**, *8*, 135-160. https://doi.org/10.1177/096228029900800204.
12. Martell, R.F.; Borg, M.R. A comparison of the behavioral rating accuracy of groups and individuals. *J. Appl. Psychol.* **1993**, *78*, 43-50. https://doi.org/10.1037/0021-9010.78.1.43
13. Ogden, P.E.; Edwards, J.; Howell, M.; Via, R.; Song, J. The effect of two different faculty development interventions in third-year clerkship performance evaluations. *Fam. Med.* **2008**, *40*, 333-338. PMCID: 18465282.