ASEE
125 YEARS AT THE HEART OF ENGINEERING EDUCATION
2018 ASEE ANNUAL CONFERENCE & EXPOSITION
JUNE 24–27, 2018
SALT PALACE CONVENTION CENTER
SALT LAKE CITY, UT

Paper ID #21421

# Simpson's Paradox and Equity in a Classroom: When Dropping the Worst Homework is Prejudicial to Your Students

**Dr. Javier Rubio-Herrero, St. Mary's University, San Antonio**

Javier Rubio-Herrero, Ph.D., joined St. Mary's University in 2017 as an Assistant Professor of Industrial Engineering in the Department of Engineering.

Dr. Rubio-Herrero completed his Ph.D. in Operations Research at Rutgers University. Previously, he received a M.S. in Industrial Engineering from the University of Wisconsin-Madison and a B.S. in Industrial Engineering from the University of Seville (Spain).

Before joining St. Mary's, Dr. Rubio-Herrero worked at Pacific Northwest National Laboratory, where he applied optimization and machine learning techniques to a diverse array of areas such as national security, deep learning, and energy. He also has experience in industry (in the context of supply chain and manufacturing) and in the public sector, where he was part of the Transport and Energy unit at the Institute for Prospective Technological Studies.

Dr. Rubio-Herrero has published articles in journals and conference proceedings, and serves as an active reviewer in peer-reviewed publications. His research interests deal with the applications of optimization and operations research techniques to solve engineering problems.

# Simpson’s Paradox and equity in a classroom: when dropping the worst homework is prejudicial to your students

**Abstract**

Dropping the worst homework, or the homework with the lowest grade, is a common practice that instructors do when they intend to increase their students’ grades. The following article shows that if this measure is taken at some stage during the course, other than at the end of it, the grade of some students worsens after dropping their worst homework and the perception of their performance is biased. To illustrate this phenomenon, we provide an example in which this decrease can be of almost 4% and we find that this effect is more felt if a student has performed poorly in a midterm exam, i.e. it targets those to whom this policy is supposed to help. While in terms of equality this policy is usually extended to all the students, we conclude that its performance fails when it comes to assessing its equity. This is due to the effect of the so-called Simpson’s Paradox.

## Simpson’s Paradox in the literature

Although first properly introduced by Edward Simpson when working on contingency tables that did not show second-order interactions [1], Simpson’s Paradox (henceforth SP) has been known and observed for more than a century. In short, this paradox represents the mathematical phenomenon by which aggregated measures behave contrary to what was expected when observing and comparing their individual component measures. It has been spotted in many studies and, as a matter of fact, it has been recently calculated that the contingency tables presented by Simpson present this effect 1.67% of the times [2]. Probably the most famous study to date is the one that dismissed the possibility of sex bias in graduate admissions at the University of California-Berkeley despite what aggregated data seemed to indicate [3]. Indeed, of all men that applied for admission to graduate school starting in fall 1973, 44% were admitted versus only 35% of women. This indication of a bias at the aggregated level was found statistically significant by the appropriate chi-square nonparametric test. However, close examination of similar admission data done at a disaggregated level (i.e. department by department) showed that such bias did not exist, even unveiling cases where the admission rates were favorable to women. Also in the academic environment, a more recent study proposed a methodological refinement for studying the performance of underrepresented minority students in STEM classes, as they showed that SP may cause a misconception in the broad agreement that these students underperformed in such classes relative to their overall GPA [4]. SP is also responsible for the misconception that graduation rates of male athletes are lower than those of their peers [5].

Another example is that of recovery rates due to a drug treatment: even though recovery rates in a global population might be higher with a drug treatment than without, they can be lower if the data are analyzed in males and females separately [6].

The effect of this paradox has been noticed in many other fields. For example, when comparing income tax rates in 1974 and 1978, each income category saw a reduction in tax rate, yet the overall tax rate increased [7]. Even in sports analytics, SP may lead to incorrect conclusions: the National Basketball Association (NBA) reports, among other metrics, the field goal percentage (i.e. the ratio of field goals made to field goals attempted) of each of the two teams that play in a basketball game. This metric includes two-point and three-point field goals. Interestingly, in an NBA Playoffs game in 2011 between the Memphis Grizzlies and the San Antonio Spurs, the Grizzlies combined for a field goal percentage of 47.1% versus a 45.9% by the Spurs. Nevertheless, the Grizzlies shot worse than the Spurs both in two-point shooting and three-point shooting (49.4% and 30.0% versus 50.8% and 31.8% respectively) [8].

A similar effect can also appear even when the trends (increasing or decreasing) at both the aggregated level and the disaggregated level point to the same direction: SAT scores in the United States rose by 7 points between 1980 and 1984, but the increments at a disaggregated level were higher than that (8 points for white people, 15 points for non-white people) [9].

Obviously, the manifestation of this paradox may lead to important misperceptions, especially in sensitive settings, which makes this phenomenon of utter importance when it comes to interpreting results. Clearly, a misinterpretation of the results from Berkeley graduate admissions might have led to legal problems. A misunderstanding of recovery rates due to a specific drug may lead to incorrect or false advertising, also leading to legal and ethical issues.

It is my aim to provide in this article another example of SP that took place in an educational setting. The contents discussed here arose while reevaluating the grading scheme in a statistics course. This reevaluation consisted of a widely used measure when grading in the United States: dropping the worst grade from the homework of each student in class. Such practice intends to allow for a “bad day”, so this grade is not taken into account when computing the homework performance of a given student, consequently increasing his or her grade. In general, this policy gives way to grade inflation, a phenomenon that is well documented [10, 11, 12, 13, 14] and that has has been historically a concern because of its implications when it comes to assessing the reliability of a student’s performance. This inflation is even more prominent when grades are allocated into a small number of categories, as is usually the case with the typical A-to-F scheme. In particular, some authors have already analyzed the impact of this policy on the students’ behavior and their performance when it comes to taking an optional final exam or to preparing properly for a mandatory final test [15, 16, 17, 18]. Other authors have considered the problem of selecting a larger set of assignments and quizzes to drop such that the grade improvement is maximized [19]. However the goal of this paper is to explain why this practice, surely beneficial to the students’ grades in the long term (i.e. once the course finishes), might very well be prejudicial in the middle of the course. This is because it leads to the unexpected and to some extent counterintuitive outcome of a grade decrease if the score is calculated at some other stage during the course. This effect is another instance of the paradox that has just been introduced.

### How Simpson's Paradox appears in our classroom

Dropping the worst homework is a very common practice intended to increase the overall score of a course section. One assumes that doing so will help increase the grade of each and every student in the classroom. It is important to define that by score at a given stage we mean the ratio between the number of points attained up to that stage and the maximum number of points attainable up to that stage. Mathematically this score is

$$S(\%) = \frac{\# \, of \; points \; attained \; by \; student}{Maximum \, \# \, of \; points \; attainable} \times 100.$$

The maximum number of points attainable depends on the moment in the course that this score is computed. If we assume that the overall score of a course is 100 points, the maximum points attainable by any student at the end of the course is clearly 100. However, the maximum points attainable by student at week 5 will depend on the number of homework assignments, projects or midterm exams completed at that stage and their respective weights in computing the overall 100 points that the course is based on. For instance, in this statistics course the grading scheme was:

- Homework: 30% (5 assignments, 6% each assignment)
- Midterm exam: 30%
- Final exam: 40%

The deadline for dropping this course was the end of week 8 of the semester. By the end of that week, the students had finished 3 homework assignments and the midterm exam. The maximum number of points attainable at that stage of the course was 48: $100 \times 0.30$ points from the midterm and 18 from the 3 homework assignments (i.e. $100 \times 0.06 \times 3$). In order to increase the students' grades the worst homework was excluded from the calculations. Interestingly, I found that while most of the students saw their grades enhanced, some others were prejudiced by this measure and obtained poorer scores than they had before dropping their homework. The importance of this change lies obviously in the bias introduced when assessing their performance, only a few days before the deadline for dropping the course. Students that are bad informed about their grade status may withdraw, with the subsequent economic impact on their financial situation.

Hence, this is yet another instance in which the aggregated measure (overall score at the end of week 8) presents a reversal in its trend with respect to the disaggregated measures (scores in the midterm and the homework separately, at the end of week 8): after dropping the worst homework, all the students keep the same score in their midterms and improve their score in their homework; however, not all of them improve their overall score.

For the sake of the discussion that will follow, let $w_h$ be the weight assigned to the average homework score and $w_m$ be the weight assigned to the midterm score. These weights are relative to the overall score of the course (i.e. the average score of the homework has a weight of 0.3 in the final score) such that $0 < w_h, w_m < 1$ and $w_h + w_m \leq 1$. Let $n \geq 2$ be the total number of homework assignments in the course and $n'$ be number of homework assignments completed at the time the score evaluation is done (clearly $n' \leq n$). Moreover, let $m$ be score attained in the midterm exam and $h_i, i = 1, 2, ..., n'$ be the *ordered* scores of the homework completed at the time the score evaluation is done, such that $100 \geq h_1 \geq h_2 \geq ... \geq h_{n'} \geq 0$. In the case that is illustrated in this article, $n = 5, n' = 3, w_h = 0.3$, and $w_m = 0.3$.

In order to illustrate this phenomenon, Table 1 shows the scores of some students at the end of week 8 and how SP occurred in some cases. For privacy reasons, all the grades in this table as well as the names of the students are factitious and conceived for illustration purposes only.

| **STUDENT** | $\mathbf{\frac{\sum_{i=1}^{n'} h_i}{n'}}$ | $\mathbf{h_{n'}}$ | $\mathbf{\frac{\sum_{i=1}^{n'-1} h_i}{n'-1}}$ | **m** | $\mathbf{S_b}$ | $\mathbf{S_a}$ |
|---|---|---|---|---|---|---|
| John | 74.00% | 40.00% | 91.00% | 85.00% | 80.88% | 87.00% |
| Lauren | 87.33% | 80.00% | 91.00% | 78.00% | 81.50% | 82.33% |
| Sandra | 95.00% | 90.00% | 97.50% | 40.00% | 60.63% | 59.17% |
| Carlos | 93.66% | 93.00% | 94.00% | 0.00% | 35.13% | 31.33% |

Table 1: Effect of dropping the worst homework in some student's grades at the end of week 8

All the students see an increase in their homework average grade. However, contrary to what intuition may dictate, Sandra and Carlos see their scores reduced. The reason is easy to spot: a student's score before dropping the lowest grade is

$$S_b(\%) \quad = \quad \frac{\frac{\sum_{i=1}^{n'} h_i}{n'} w_h' + m w_m}{w_h' + w_m}, \tag{1}$$

where $w_h' = w_h n'/n$. A student's score after dropping the lowest grade is

$$S_a(\%) \quad = \quad \frac{\frac{\sum_{i=1}^{n'-1} h_i}{n'-1} \bar{w}_h' + m w_m}{\bar{w}_h' + w_m}, \tag{2}$$

where $\bar{w}_h' = w_h(n'-1)/(n-1)$. Since $n' < n$, the ratio $n/n-1$ is always smaller than the ratio $n'/n'-1$ and it follows that $\bar{w}_h'/w_h' < 1$ (i.e. $\bar{w}_h' < w_h'$). Therefore, the denominator of (2) is always smaller than the denominator of (1). In this sense $S_a$ should be greater than $S_b$. However, this is not guaranteed because the numerator of (2) could be smaller than the numerator of (1), even though the average after dropping the lowest homework does not decrease because $\sum_{i=1}^{n'-1} h_i/(n'-1) \geq \sum_{i=1}^{n'} h_i/n'$. Consequently, we cannot conclude that $S_a > S_b$.

An immediate question that arises after this analysis is: when is the impact of SP felt the most? In other words, when is $S_b - S_a$ maximized? Note that any positive value of this difference indicates a reversal, an occurrence of SP. For example, consider that the weights $w_m, w_h$, and the total number of homework asigments $n$ are given by the course syllabus and that a student has a homework average $\sum_{i=1}^{n'} h_i/n'$ after $n'$ assignments. Then, the effect of SP, if felt at all, will be directly proportional to the score of the dropped score and inversely proportional to the grade in his or her midterm exam. If we write $\sum_{i=1}^{n'-1} h_i/(n'-1) = (\sum_{i=1}^{n'} h_i - h_{n'})/(n'-1)$, the function $S_b - S_a$ is linear with respect to the worst homework $h_{n'}$ and the midterm grade $m$:

$$S_b - S_a \quad = \quad \alpha h_{n'} + \beta m + \gamma, \tag{3}$$

where

$$
\begin{aligned}
\alpha &= \frac{\bar{w}'_h}{(n'-1)(\bar{w}'_h + w_m)},\\
\beta &= \frac{w_m}{w'_h + w_m} - \frac{w_m}{\bar{w}'_h + w_m},\\
\gamma &= \sum_{i=1}^{n'} h_i \left( \frac{w'_h}{n'(w'_h + w_m)} - \frac{\bar{w}'_h}{(n'-1)(\bar{w}'_h + w_m)} \right).
\end{aligned}
$$

The slope $\alpha$ is strictly positive, whence it follows that, ceteris paribus, the quantity $S_b - S_a$ increases with the score of the dropped homework. This score is bounded above by $\sum_{i=1}^{n'} h_i/n'$ (i.e. the minimum score cannot be greater than the average) and below by $\max\{0, \sum_{i=1}^{n'} h_i - 100(n'-1)\}$ (i.e. the minimum score cannot be lower than 0 or the difference between the total score after $n'$ assignments and a perfect score in $n'-1$ assignments, whichever is greater). Similarly, the slope $\beta$ is strictly negative because $\bar{w}'_h < w'_h$, and therefore, again ceteris paribus, the quantity $S_b - S_a$ decreases with the score of the midterm exam. The score in the midterm exam may vary between 0% and 100%. Figure 1 shows the score differences for our four students as a function of their midterm and worst homework grades. Each plane corresponds to a student. Note that each plane is bounded by the values that $m$ and $h'_n$ may take. Since $\alpha$ and $\beta$ only depend on the course syllabus and the stage at which the grades are modified but not on the student's performance, all the planes will be parallel. The black points mark the coordinates for each student as given by Table 1.

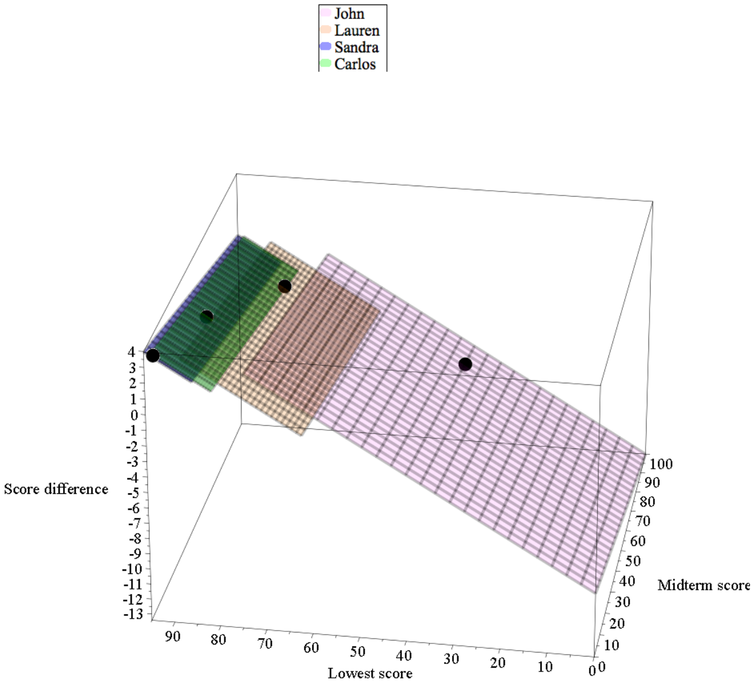

Figure 1: Score difference as a function of the midterm and the worst homework grades

The interpretation of the slopes $\alpha$ and $\beta$ give way to reconsider the equity that the policy of dropping the lowest homework introduces in our classroom. Indeed, there is a clear sense of equality because all the students “enjoy” the application of this policy to modify their grades. However, the equity of such a measure is very questionable, to the point that it can be prejudicial to many students, more concisely to those that did especially good in the homework or especially bad in the midterm exam. As a rule of thumb, Table 2 shows qualitatively the impact that dropping the lowest homework has on students depending on their performance during the course up to the stage $n'$ where the policy is applied. On average, students tend to perform better in the homework than in the midterm (which makes sense because when completing the homework they usually work in groups and seek some extra help). This is especially true in the case of poor performers, for which this dropping policy is typically applied, and therefore the equity and effectiveness of these actions are clearly compromised.

| | **High Score** | **Low Score** |
|---|---|---|
| **Worst homework** | ☹ | ☺ |
| **Midterm exam** | ☺ | ☹ |

Table 2: Qualitative assessment of the effectiveness of the policy of dropping the worst homework

Now consider that our four students have already done their midterm exams and we drop their worst homework assignment. As an example, Figure 2 shows how our four students’ lowest grade would affect their score. Each line corresponds to a student. The difference between $S_b$ and $S_a$ is calculated for the range allowed in each case. This difference is maximum when the worst homework obtained the highest possible grade (i.e. when $h_{n'} = \sum_{i=1}^{n'} h_i/n'$). This maximum difference can still be negative; students with this characteristic, will never decrease their overall score when dropping their lowest homework: SP will not affect them. This is the case of John, but not the cases of Lauren, Sandra, and Carlos. Lauren is a candidate for a reversal effect, but her worst homework’s grade (80.00%) is not high enough to enter the reversal zone. Sandra and Carlos, however, have grades that are within this zone and thus their overall score decreases after increasing their homework averages. In Carlos’s case, this difference can be almost 4%.

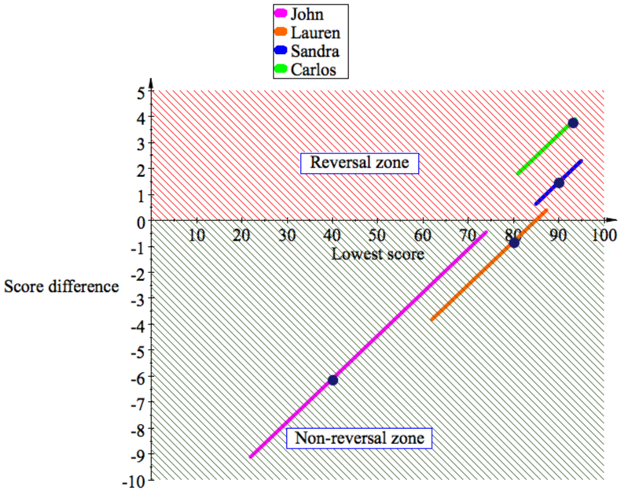

Figure 2: Occurrence and effect of Simpson’s paradox as a function of the lowest homework grade

Likewise, once all the $n'$ homework have been graded, the difference of scores is inversely proportional to the grade of the midterm. This is shown in Figure 3. John will never have SP impact his grade, regardless his score in the midterm. Lauren is not affected but she would have been should she had scored about 20 points less in her exam. Finally, Sandra and Carlos reduce their overall grade, but this would have not been the case should they had scored considerably more in the exam.

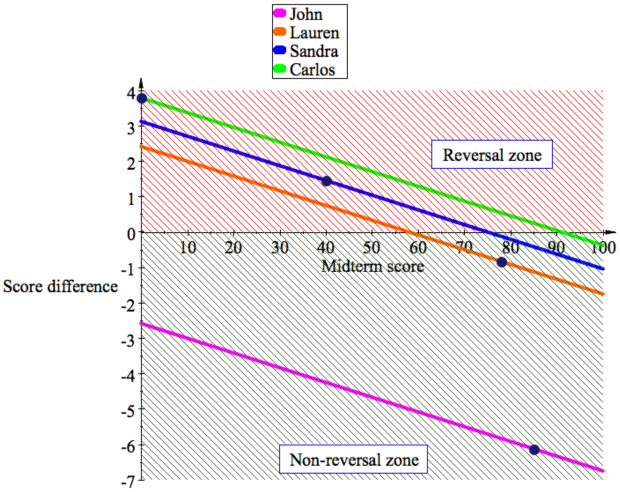

Figure 3: Occurrence and effect of Simpson's paradox as a function of the midterm grade

As commented in the first lines of this paper, SP may affect students when evaluating their performance at some stage during the course, but never at the end of the course. This can be seen by examining equations (1) and (2). The scores at that stage can be calculated as

$$\begin{aligned} S_b(\%) &= \frac{\dfrac{\sum_{i=1}^{n'} h_i}{n'} w'_h + m w_m + f w_f}{w'_h + w_m + w_f}, \\ S_a(\%) &= \frac{\dfrac{\sum_{i=1}^{n'-1} h_i}{n'-1} \bar{w}'_h + m w_m + f w_f}{\bar{w}'_h + w_m + w_f}, \end{aligned}$$

where $f$ is the score of the final exam and $0 < w_f < 1$ is its weight. It must hold that $w_h + w_m + w_f = 1$. However, at the end of the course $n' = n$ and therefore $w'_h = \bar{w}'_h = w_h$, and thus

$$\begin{aligned} S_b(\%) &= \frac{\dfrac{\sum_{i=1}^{n} h_i}{n} w_h + m w_m + f w_f}{w_h + w_m + w_f}, \\ S_a(\%) &= \frac{\dfrac{\sum_{i=1}^{n-1} h_i}{n-1} w_h + m w_m + f w_f}{w_h + w_m + w_f}. \end{aligned}$$

Given that the average of the homework does not decrease when dropping the worst of them, and that both denominators are identical, it is easy to see that $S_a > S_b$.

## Conclusions

In this paper we show that dropping the worst homework of our students may affect negatively their overall grades at some stage during the course, but never at the end. This phenomenon constitutes an instance of SP and its importance lies in how it introduces a bias in the students' perception towards their performance at that stage of the course and may even lead to undesired withdrawals. Such withdrawals obviously result in an economic burden on the students' economy, since they will have to register again for the same number of credits in this or in another course.

In general, we show that this policy may not produce the desired outcome at an intermediate stage of the course if students have low grades in their midterm or high grades in their worst homework. This is because the effect produced by SP is directly proportional to the grade of the worst homework and inversely proportional to the score attained in the midterm. Therefore, this reversal, if it appears at all, will have its greatest impact on students that obtained a 0 in their midterm exam and full score in all their homework assignments. This fact greatly questions the equity and effectiveness of this measure when evaluated in the middle of the course, as it is mainly taken to improve the grade of those students that do not perform well and yet it may punish them for their low midterm grades.

## References

[1] E. H. Simpson, “The interpretation of interaction in contingency tables,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 238–241, 1951.

[2] M. G. Pavlides and M. D. Perlman, “How likely is Simpson’s Paradox?” *The American Statistician*, vol. 63, no. 3, pp. 226–233, 2009.

[3] P. J. Bickel, E. A. Hammel, and J. W. O’Connell, “Sex bias in graduate admissions: Data from Berkeley,” *Science*, vol. 187, no. 4175, pp. 398–404, 1975.

[4] J. Tomkin, M. West, and G. L. Herman, “A methodological refinement for studying the STEM grade-point penalty,” in *Frontiers in Education Conference (FIE), 2016 IEEE.* IEEE, 2016, pp. 1–5.

[5] V. A. Matheson, “Research note: Athletic graduation rates and Simpson’s Paradox,” *Economics of Education Review*, vol. 26, no. 4, pp. 516–520, 2007.

[6] J. Pearl, “Simpson’s Paradox: An anatomy,” UCLA Cognitive Systems Laboratory, Tech. Rep., 2011.

[7] C. H. Wagner, “Simpson’s Paradox in real life,” *The American Statistician*, vol. 36, no. 1, pp. 46–48, 1982.

[8] Y. Z. Ma and A. M. Ma, “Simpson’s paradox and other reversals in basketball: examples from 2011 NBA playoffs,” *International Journal of Sports Science and Engineering*, vol. 5, no. 03, pp. 145–154, 2011.

[9] H. Wainer, “Minority contributions to the SAT score turnaround: An example of Simpson’s Paradox,” *Journal of Educational and Behavioral Statistics*, vol. 11, no. 4, pp. 239–244, 1986.

[10] R. Singleton and E. R. Smith, “Does grade inflation decrease the reliability of grades?” *Journal of Educational Measurement*, vol. 15, no. 1, pp. 37–41, 1978.

[11] J. Millman, S. P. Slovacek, E. Kulick, and K. J. Mitchell, “Does grade inflation affect the reliability of grades?” *Research in Higher Education*, vol. 19, no. 4, pp. 423–429, 1983.

[12] H. Zangenehzadeh, “Grade inflation: A way out,” *The Journal of Economic Education*, vol. 19, no. 3, pp. 217–226, 1988.

[13] K. Ogilvie and M. Jelavic, “Grade inflation in the US higher educational environment: a faculty perception study,” *International Journal of Management in Education*, vol. 7, no. 4, pp. 406–416, 2013.

[14] M. Elie, “Grade inflation in nursing education: Proposed solutions for an ongoing problem,” in *Nursing forum.* Wiley Online Library, 2015.

[15] E. Sewell, “Grade dropping: An empirical analysis,” *The Journal of Economic Education*, vol. 35, no. 1, pp. 24–34, 2004.

[16] L. Hadsell and R. MacDermott, “Grade dropping, strategic behavior, and student ’satisficing’,” *American Journal of Business Education*, vol. 3, no. 7, p. 57, 2010.

[17] R. MacDermott *et al.*, “The effects of dropping a grade in intermediate macroeconomics,” *New York Economic Review*, vol. 40, no. 1, pp. 40–50, 2009.

[18] R. J. MacDermott, “The impact of assessment policy on learning: Replacement exams or grade dropping,” *The Journal of Economic Education*, vol. 44, no. 4, pp. 364–371, 2013.

[19] D. M. Kane and J. M. Kane, “Dropping lowest grades,” *Mathematics Magazine*, vol. 79, no. 3, pp. 181–189, 2006.