# Rubric Scales: How a 'Zero' Scoring Option May Alter Rater Choices

**James Woodley**
**Ramapo College of New Jersey**

**Kathryn Yeaton**
**Ramapo College of New Jersey**

**Teresa Hutchins**
**Ramapo College of New Jersey**

*This article presents findings for an empirical study of how faculty raters respond to the scale values adopted for programmatic assessment rubrics. Analysis highlighted a statistically significant increase in the frequency with which the lowest score was assigned when the lowest score was one, rather than zero. Since the use of zero as a scoring option is common for rubrics used in assessment carried out in higher education, questions are raised about the potential impact on assessment results of relatively subtle changes in rubric design and, ultimately, on closing the loop activities proposed in response to programmatic assessment results.*

## INTRODUCTION

As institutions of higher education in the United States have been under pressure to justify rising costs and to deflect claims that students are not learning, they have worked to establish formalized processes that measure learning outcomes (Feuerstein, 2015). Thus, the validity and reliability of efforts to measure learning outcomes emerges as a matter of no small consequence. Against this backdrop, the use of scoring rubrics for assessment has become widespread, with rubrics generally accepted as a legitimate and useful tool for measuring learning outcomes and for providing insights that can support efforts to fix problems observed with student learning. Despite their widespread use, many studies indicate that while 'rubrics enhance and enrich assessment of student work, the validity of the tool is not without debate' (Rezaie and Lovorn, 2010, p.18). This debate arises from a variety of factors, including the scale utilized to evaluate student work.

This article aims to contribute to the larger, ongoing conversation about how to design rubrics in ways that lead to better measurement of learning outcomes by exploring the potential influence of rubric scale values on rater choices. The key issue examined here is whether raters are more or less likely to assign the lowest rating to student work when the lowest rating is zero, compared to when the lowest rating is not zero. This question became salient after observing the results of a change in rating scales for program assessment rubrics in a business school. The scale initially ranged from "0" through "5" and was later changed to "1" through "6", thereby eliminating a zero scoring option from the choices available to raters. Faculty ratings of student work were compared before and after this change to determine the potential

impact of changing the lowest score from zero to one. See Table 1 below for a quick summary of the descriptive findings.

**TABLE 1**
**COMPARING RATER SCORING CHOICES ACROSS TIMES PERIODS**

| Time period | Number of raters (*n*) | Raters never choosing the lowest rating (%) |
|---|---|---|
| Zero is lowest score | 38 | 47 % |
| One is lowest score | 36 | 25 % |

*N* = 74 raters

As Table 1 highlights, the percentage of raters never assigning the lowest rating dropped by almost half in the time frame in which one was the lowest score, instead of zero. To more carefully examine whether a lowest score of zero may alter rater choices, the second section of this manuscript reviews the literature. The third section presents hypotheses. The fourth section describes methodology. The fifth section presents the results of statistical analysis. The sixth and final section presents conclusions from this study, suggestions for practitioners, and suggestions for future research.

## LITERATURE REVIEW

Rubrics are tools which can be used as scoring guides (Suskie, 2004) and they are used in numerous educational settings on a worldwide basis. Each rubric normally includes: 1) a list which describes criteria or traits to be evaluated, and 2) guidelines used to evaluate each criterion. The guidelines used for evaluation typically take the form of rating scales used to indicate the degree to which students have achieved the desired criteria (i.e., performance levels). Current literature on the use of rubrics in post-secondary education has concentrated on the creation of rubrics and the scoring of student work samples when using rubrics. Relevant issues with the creation of rubrics are: establishment of linkages between learning goals and the evaluation criteria, plus the definition of quality differences and the adoption of a rating scale (Reddy and Andrade, 2010; Reddy, 2011).

A second area of research has focused on the scoring of student work samples when using rubrics, with the most common subject being how to establish reliability. Moskal and Leydens (2000) defined assessment reliability as the

> …consistency of assessment scores. For example, on a reliable test, a student would expect to attain the same score regardless of when the student completed the assessment, when the response was scored, and who scored the response. (Reliability section, para. 1)

To delve further into the nature of reliability, there are two forms of reliability that are generally discussed. These two types of reliability are: 1) interrater reliability, which refers to the consistency of scores assigned by two independent raters, and 2) intrarater reliability, which refers to the consistency of scores assigned by the same rater over time. Moskal and Leydens (2000) propose that the appropriate development and use of rubrics should mitigate errors from both of these types of reliability, as has Stemler (2004) and Jonsson & Svingby (2007).

On a related subject, Suskie (2004, p.141) states "No matter how carefully a rubric is constructed, rubric scores remain essentially subjective and thus prone to unintentional scoring errors and biases. Rating scale rubrics and holistic scoring guides are especially prone to scoring errors and biases ..." Thus, according to Suskie, despite the best efforts of those involved scoring errors and biases can unintentionally be introduced into the assessment process.

Given the possibility of errors and bias when using rubrics, it is helpful that there are numerous journal articles and books written to guide in the development and construction of rubrics for use in educational settings. Accordingly, there exists a rich literature about rubric scales. But, a dearth of literature exists examining the implications of specific values that might be assigned to a rubric scale. The closest line of research analyzes the use and impact of rating augmentation (e.g., Penny, Johnson, and Gordon, 2000). Rating augmentation involves expanding the scale of a rubric by giving raters the option of assigning half increments or pluses and minuses, rather than a single whole number. A rater first decides on the level of proficiency (i.e., Unacceptable, Acceptable or Exemplary), and then the rater decides where the student work sample falls within that level of proficiency. For example, a piece of student work may be scored as 5 or 6 after it is identified as Exemplary, if 5 and 6 are the top two scores available.

From among the guides that specifically address or provide examples of rubric scales, many incorporate scales which include zero as a scoring choice. A zero scoring option has various definitions across rubrics (e.g., Baryla, Shelley, and Trainor, 2012; Mertler, 2001; Moskal, 2000; Suskie, 2009; Walvoord & Anderson, 1998; Wiggins, 1998), with the common theme of a zero score being the lowest score on the scale. Guides also display consistency in how specific numeric scoring options are simply provided without a rationale given for the numbers selected. Finally, there is a very limited range of numbers used over and over in scales for rubrics, with negative values and large values essentially never being used. One can reasonably infer that those who design rubrics believe some potential scoring values would be vexing for raters (ie: negative values, infinity, one million, …). And yet, the literature implies that raters are likely to be indifferent to the specific numbers selected for a rating scale, at least within the numerical ranges usually suggested and used[1].

Distinctions can be observed across rubric construction guides and examples in how a zero score is defined when it is included as a scoring option. Authors employ a zero score in sample rubrics to denote "lowest level of performance" (Moskal, 2000; Suskie, 2004), "the absence of something" (Mertler, 2001), or both (Allen, 2006), depending on the rubric. One would hope that the absence of a clearly observable pattern for what a zero score represents has resulted from a careful examination of rater responses to scale structure, with empirical research showing that raters pay no attention to whether or not zero is a scoring option and instead focus solely on the verbiage given to describe different levels of student performance. However, at least in the context of programmatic assessment previous empirical research has not examined whether a zero scoring option might lead to bias in ratings given for student work, nor the circumstances under which the introduction of bias might become more or less likely.

## HYPOTHESES

The suggestion implicit in the literature that all of the frequently selected numeric scoring options for rubrics are equivalent in the eyes of raters is the key underlying issue examined in this section. If all of the frequently selected numeric scoring options for rubrics are equivalent in the eyes of raters, then there should be no statistically significant differences in ratings selected when rating scales shift up or down to different numerical values within the range of frequently selected values. This observation forms the basis for a proposition that raters are indifferent to the numerical values assigned to rubric scoring. Two hypotheses that facilitate testing this assertion are given below:

> Hypothesis 1 (H1): Raters using rubrics to score student work samples will be indifferent to the specific numerical value assigned to the lowest score.
>
> Hypothesis 2 (H2): Raters using rubrics to score student work samples will be indifferent to the specific numerical value assigned as the highest score.

The "Hypothesis 1" version of the proposition of rater indifference to specific numerical scoring options will likely not be supported by statistical testing if a change in scores assigned by raters is influenced by a

change of the lowest value from zero to a higher value. This becomes the primary focus of statistical testing presented later.

## METHODOLOGY

The programmatic assessment data utilized in this study was generated by School of Business faculty members working at a selective public comprehensive college in New Jersey. The faculty members involved in the assessment and evaluation process represented each of seven disciplines (Accounting, Economics, Finance, Information Technology, International Business, Management, and Marketing). In all cases, faculty evaluators were provided with a rating scale rubric designed to address a specific learning goal. Interrater reliability (i.e., norming) sessions were held for all evaluations in both two year time periods. Assessments for each learning goal were undertaken every second year, for a total of two assessments per learning goal over a four year period. The six learning goals being assessed were reasoning, integration, written communication, oral communication, perspectives, and ethics. Zero scores in time period 1 denoted "lowest level of performance", and not "the absence of something", for the overwhelming majority of ratings analyzed for this study.

The data utilized in this study consists of 4,192 evaluations of student work by 74 raters, collected during the period Fall 2009 through Spring 2013. Because the subject of interest was the possible effect of specific rubric scale values on the choices of *individual raters*, data was aggregated to the "*per rater*" level[2]. Thus, for example, whether or not a rater ever assigned the lowest score in all the work they rated, for a given assessment, was tracked as one variable (i.e., NoLowest = 1 means a rater never assigned the lowest score, for a given assessment).

An examination of 7 years of SAT scores for incoming students, beginning 3 years prior to the start of the analysis summarized later, suggested a very high degree of consistency in the preparedness of entering students. Also, grade point average (GPA) figures for graduating business school students changed little during the four years in question, with no discernible trend upward or downward.

The data was divided into two subsets: the period in which 0 was the lowest score and the period in which 1 was the lowest score. Because students are juniors or seniors when contributing work to assessments carried out, it is almost never the case that a student would have their work assessed more than once for the same learning goal. Also, even though the pool of raters was relatively stable over the four year time frame, as faculty turnover was not high, faculty raters usually assessed different goals across the two time periods. The learning goals, the traits, and the guidelines to evaluate each trait all remained essentially the same. The key substantive difference between time periods was the rubric scale values, which were 0-5 during the first two year period and 1-6 during the second time period.

Table 2 describes variables that were tested through statistical analysis. Table 3 presents descriptive statistics for each of the variables in Table 2, as well as Spearman's RHO values.

**TABLE 2**
**VARIABLE DEFINITIONS**

| Variable | Measurement scale |
|---|---|
| NoLowest | 1 = Rater never gives the lowest rating, 0 = Otherwise |
| NoHighest | 1 = Rater never gives the highest rating, 0 = Otherwise |
| Average* | The average rating given by each rater (minimum value = 0, maximum value = 5) |
| CanZero | 1 = time period in which the rubric scale was 0 to 5, 0 means "lowest quality work" |
| | 0 = time period in which the rubric scale was 1 to 6, 1 means "lowest quality work" |

* Average ratings had one point subtracted when the rating scale was 1 – 6 to make average ratings more easily comparable across the different time frames in which either one of the two rating scales outlined here was used.

**TABLE 3**
**DESCRIPTIVE STATISTICS AND SPEARMAN'S RHO**

| | | Minimum | Maximum | Mean | $\sigma$ | | 1 |
|---|---|---|---|---|---|---|---|
| 2 | 3 | | | | | | |
| 1 – NoLowest | 0 | 1 | 0.36 | 0.49 | | | |
| 2 – NoHighest | 0 | 1 | 0.32 | 0.47 | -.165 | | |
| 3 – CanZero | 0 | 1 | 0.51 | 0.50 | .232* | -.076 | |
| 4 – Average | 1.39 | 4.07 | 2.48 | | 0.55 | .321* | -.491* |
| .017 | | | | | | | |

$N = 74$
* Correlation is significant at the .05 level or better (2-tailed test)

## RESULTS

The statistical findings reported in Table 4 show a statistically significant increase in the frequency with which the lowest score will be selected when zero is not the lowest score (NoLowest, $p = 0.046$). In other words, raters were more likely to assign the lowest score when the lowest score was one, rather than zero. This statistical result does not support hypothesis 1 (H1) that raters are indifferent to the numerical value assigned to the lowest score. It seems noteworthy that many faculty assigned at least one zero score to student work in the first time period, just as many did not, which suggests the possibility of differences across raters in their reaction to a zero scoring option.

**TABLE 4**
**COMPARING SAMPLE MEANS ACROSS TIME PERIODS**

| Hypothesis Variables (Supported?) | CanZero = 1 ( $\sigma$ ) | CanZero = 0 ( $\sigma$ ) | $\chi 2$ | t |
|---|---|---|---|---|
| NoLowest H1 (No) | .47 (.51) | .25 (.44) | 3.991* | |
| NoHighest H2 (Partial) | .29 (.46) | .36 (.49) | 0.433 | |
| Average | 2.49 (.53) | 2.46 (.57) | | 0.223 |

N = 74
Significance level - * $p \leq 0.05$
All statistics are for 2-tailed tests
$\chi 2$ = chi-squared test statistic
t = t statistic

In contrast, no statistically significant difference was found across time frames in the frequency with which the highest rating is selected (NoHighest, $p = 0.511$). The statistical findings for NoHighest support hypothesis 2 (H2) that raters are indifferent to the specific numerical score chosen for the highest score. However, the level of support provided for hypothesis 2 (H2) must be judged as partial because only one possible combination was tested for the highest numerical score ("5" versus "6").

Even though findings provide only partial support for H2, the findings for NoHighest (H2) undermine a possible counter-hypothesis that variation over time in the occurrence of errors of central tendency may explain changes in the observed willingness of raters to assign the lowest score across the two time periods. As further support in favor of the given interpretation of findings about Hypothesis 1, findings presented in Table 4 do not show a statistically significant difference in average ratings over time (Average, $p = 0.824$). Thus, the data do not support a counter-hypothesis that the higher percentage of raters assigning the lowest score in the second time period was driven by lower quality student work during the second time frame.

## CONCLUSION

It is noteworthy that different faculty may have had different levels of comfort with assigning a zero rating to student work even though a norming session was held for all assessments. This suggests that the comfort level of individual raters with assigning a zero score to student work may not be easily changed. Thus, a key practical implication of these findings could be that a zero score at the lowest end of a rubric scoring structure may reduce the overall frequency with which scores at the lowest end of the scale are selected, with inconsistencies across raters in how they respond to a zero scoring option. These effects can, in turn, reduce interrater reliability and may even lead to a pattern of under-reporting problems with student learning outcomes.

To offer a tentative suggestion to practitioners, perhaps a zero scoring option will cause fewer problems when it signifies the absence of something. Then our commonly shared association between the concepts of 'zero' and 'nothing' may do no harm to the assessment process. In fact, a zero scoring option may draw raters to score student work at the lowest end of the scale when it is meant to signify the absence of something, thereby helping to overcome possible errors of central tendency. But, some faculty may be reluctant to assign a 'zero' score when a student has "done more than nothing." For example, some faculty may be reluctant to assign a zero score to an aspect of a student's reasoning skills when asked to assign scores based on a two to four page essay. Thus, whenever the lowest rating that can be

assigned when rating student work signifies low quality, rather than the absence of something, perhaps a number other than zero would be a better choice for the lowest end of the rating scale.

In closing, findings reported here support the suggestion of Suskie (2004) that subjective judgments are difficult to completely eliminate when using rubrics. Findings reported here provide a concrete example of Suskie's suggestion by highlighting the potential impact of relatively subtle changes in rubric design on assessment results and, ultimately, on any closing the loop activities proposed in response to assessment results. Given the importance of measuring learning outcomes well, further investigation of the possible influence of rubric scale choices on the outcomes of programmatic assessment seems warranted.

## ENDNOTES

1. An informal check of numerous rubrics in use and sample rubrics given in guides suggests that values of zero to ten may easily outnumber all other choices combined in the frequency of their selection as scoring options.
2. Analysis was also undertaken at the level of one data point per row in each rubric, for each student. Statistical findings were the same for both hypotheses for both levels of analysis (ie: for both "per rater" data and "each row in each rubric, for each student" data).

## REFERENCES


Allen, Mary J. (2006). *Assessing General Education Programs*. San Francisco, CA: Anker Publishing Company, Inc.

Baryla, Ed, Shelley, Gary, and Trainor, William. 2012. Transforming rubrics using factor analysis. *Practical Assessment, Research & Evaluation*, 17(4). Retrieved November 11, 2016 from http://pareonline.net/getvn.asp?v=17&n=4

Feuerstein, A. (2015). Rituals of verification: Department chairs and the dominant discourse of assessment in higher education. *Journal of Theory and Practice in Higher Education*, 15(6), 38-51.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2(2), 130-144.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25). Retrieved November 25, 2015 from http://pareonline.net/getvn.asp?v=7&n=25

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation*, 7(10). Retrieved November 25, 2015 from http://PAREonline.net/getvn.asp?v=7&n=10

Moskal, B. M (2000). Scoring rubrics: What, when, and how? *Practical Assessment, Research & Evaluation*, 7(25). Retrieved November 23, 2015 from http://PAREonline.net/getvn.asp?v=7&n=3

Penny, J., Johnson, R. L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric. *Assessing Writing*, 7(2), 143-164.

Reddy, M. Y. (2011). Design and development of rubrics to improve assessment outcomes. *Quality Assurance in Education*, 19(1), 84-104.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, 15(1), 18-39.

Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4). Retrieved November 25, 2015 from http://PAREonline.net/getvn.asp?v=9&n=4

Suskie, L. (2004). *Assessing student learning: A common sense guide*. San Francisco, CA: Anker Publishing Company, Inc.

Suskie, L. (2009). *Assessing student learning: A common sense guide*, 2nd Ed. San Francisco, CA: Jossey-Bass.

Walvoord, B. & Anderson V. J, (1998). *Effective grading: A tool for learning and assessment*. San Francisco, CA: Jossey-Bass. Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.