Research Note

Resetting the Score: Scores as Measures of Learning

Alaina Tackitt, *University of South Florida*

David Eubanks, *Furman University*

Structured Abstract

- **Aim:** Our research note focuses on the interpretation of instructor-assigned rubric scores from a large sample of student writing. A longitudinal study reveals that raw changes in average scores should not be interpreted as an adequate measure of learning. We explore the impact of our finding, share our responses, and suggest directions for future research.
- **Problem Formation:** Ideally, educational stakeholders would like to interpret changes in average scores as a reflection of writing ability; if the scores increase, students must be learning. Our findings suggest that the situation can be more complicated and that we cannot take increases in score averages at face value.
- **Information Collection:** The data set we analyzed is comprised of rubric scores assigned by 128 instructors to 52,001 essays over three years across a two-course sequence of first-year writing courses at a large, public R1 university. Score patterns revealed that averages increased across the three projects in both courses with varying amounts of measured growth in areas scored by instructors using the analytic rubric. Focusing on the scores of 1,887 students who completed all six projects across the two-semester sequence revealed that averages rise within each course but dip, or reset, between semesters. After the reset, scores from the second course in the sequence reach nearly the same level where they ended the first term.
- **Interpretation:** A simple narrative suggesting that the increasing score averages signify student learning would force us to assume that students

regress over the winter break, as indicated by the scoring reset, and then slowly regain the skill level they had achieved at the end of fall. We consider alternative interpretations related to rater behavior, extrarubric criteria, scale limitations, and assignment difficulty.

- **Conclusions:** Even scores generated with a standardized, analytic rubric used by trained instructors with a common curriculum designed as a scaffolded sequence cannot be assumed to provide valid longitudinal tracking of student learning. As a result, score use and interpretation, specifically in relation to classroom grading and external measures of learning, needs to be contextualized with confounding factors. In particular, we question the wisdom of using the same rubric scores for assessment data and grading.
- **Directions for Further Research:** The relationship between rubric scores and assignment or course grades deserves more attention. Since writing assessments cannot be assumed to measure changes in complex ability over time, further work is needed to understand the types of data gathering and analysis that facilitate this goal.

Keywords: institutional assessment, interpretation and use arguments, score reset, rubrics, extrarubric criteria, grading, score gain, VALUE rubric

1.0 Aim

Writing assessment scores affect students through admission, placement, and intervention decisions (Haswell & Elliot, 2019). Institutionally, score averages can impact state performance-based funding (State University System of Florida, 2020) and national college rankings (Morse et al., 2019). Score interpretations can be included in program evaluations or as part of larger assessment and accreditation efforts (Pagano et al., 2008). In addition to determining course grades, instructor scores can be included in research on patterns in student writing ability that inform curricular design and delivery and impact faculty development and classroom instruction (White et al., 2015). Given the many uses of scores, it is essential to attend to the validity of interpretations.

According to Kane's (2013) approach to validity arguments, "[p]ublic claims require public justification" (p. 1). To provide such justification, Kane emphasized "attention to the network of inferences and assumptions leading from test performances to conclusions and decisions" based on scores (p. 2). As a relatively new field of research, writing analytics can contribute to conversations about the interpretation and use of scores in terms of how scores are generated and the observed patterns, as well as the web of inferences made and conclusions drawn about students and their development.

Our interest is the measurement of student learning by analytic and holistic rubric scores. Our goal is to investigate score averages over time and question whether changes in score averages

can be taken at face value and used to make or support claims about student development. Specifically, we analyze and contextualize changes in average scores within a two-course sequence of first-year writing and ask: Can these averages be taken to represent trends in student learning?

2.0 Problem Formation

Measuring student learning appears to be straightforward. Educational structures are built on the belief that students absorb and practice material that results in learning and that the demonstrable improvements should be represented by an upward-sloping graph of increasing scores that reflect the learning. The expectation is that if learning is measured at multiple points, the averages should increase, and the difference signifies a quantification of student learning. The work of writing analytics often involves attempts to validate such assumptions.

Analyzing student writing samples as part of the first-year writing program at the University of South Florida (USF, the study site) revealed a pattern of increasing average scores over time. If such increases represent learning and development over time, it is evidence for effectiveness of the two-semester sequence. However, closer inspection of scores across these two courses complicates the narrative and casts doubt on a narrow or naïve interpretation of scores.

We will show that while scores increase across each term and course, they *reset* to a lower level between terms, which suggests that instead of increasing uniformly over time and terms, learning in writing courses, even a sequence of courses, is subject to declines and restarts. Complicating (or rejecting) the growth narrative renders the measures suspect and forces us to question the relationship between learning and scoring.

The act of writing and the assessment of the product are multidimensional. Extending the work of Kellogg (1993) and Flower (1994), MacArthur and Graham (2016) noted that "[f]rom a cognitive perspective, proficient writing is a complex, goal-directed problem-solving process that makes substantial demands of writers' knowledge, strategies, language skills, and motivational resources" (p. 36). A single measure based on the output of this process, as viewed through the lens of a rubric, seems unlikely to capture everything we might describe as *learning to write*.

Even if we just consider the quality of the written samples, ignoring students' writing processes and attitudes, exogenous variables can confound score interpretation. Sullivan and McConnell (2018) considered a study in which the Critical Thinking and Written Communication VALUE Rubrics were used to rate samples of student writing from courses taken during their first year, at the midpoint, and near graduation. Results from multiple institutions revealed that the scores of college seniors were higher on average than the scores of first-year students but lower than the scores of sophomores and juniors, producing more of an inverted U than a growth curve upward.

When Sullivan and McConnell (2018) reexamined the data and included the variable of assignment difficulty, they found that it impacted that outcome significantly and concluded that when students were challenged to do more, they did more. Conversely, when asked to do less, students did less—or at least their scores were lower. If this idea about assignment difficulty is

true, the use of VALUE rubrics is not valid unless assignment difficulty is included as an explanatory variable in addition to time. For our purposes, the finding implies that traditional growth narratives can be confounded by covariates and suggests that just as an upward slope may not indicate learning, a decreasing slope, or perhaps a dip such as our reset, also requires contextualization.

Similar concerns with score interpretation were raised by Rourke and Zhou (2018) in the absence of evidence of overall gains upon completion of a university writing seminar. Based on the assumptions of the simple growth narratives, the lack of increasing scores suggested a lack of learning, which alerted Rourke and Zhou to the possibility of complications with their controlled design. Similarly, the score reset in the present study can illuminate the relationship between rubric-based measurement and the presumed underlying reality of learning accomplishment, which we cannot directly observe.

A related study on the concept of reset, published in this issue of *The Journal of Writing Analytics*, revealed that developmental (not analytic) rubric ratings of writing are strongly linked to grade averages, but even when upward sloping graphs occur, significant learning gaps can still exist (Eubanks & Vanovac, 2020). Functionally, the difference between the developmental approach and the analytic rubric approach is the difference between absolute and relative measures. With growth that is studied over time, the analysis is primarily relative (to other students), suggesting that absolute (to an ideal level of performance) measures may not be well-suited for that purpose. In addition to reinforcing the limitations of score averages as signifiers of learning and recognizing the need to consider confounding variables, these considerations have considerable potential to impact teaching and learning.

3.0 Information Collection

The USF corpus comprises 52,001 intermediate and final drafts from 7,722 students and 128 instructors, spanning 2 courses, 7 terms, 3 years, and over 482 sections. The two-course sequence was designed to scaffold requisite knowledge and skills developed in the first (fall) course to be refined and advanced in the second (spring) course. For the present study, this corpus was restricted to the final drafts of 1,887 students who took both 1101 and 1102 as a fall-spring sequence and completed all six major assignments.

The six major writing assignments across the two-course sequence were scored by instructors using a common analytic rubric with five criteria (Focus, Evidence, Organization, Style, and Form), three of which distinguish between basic and critical thinking for a total eight traits from basic skills to complex reasoning (Moxley & Eubanks, 2016). Course content was centrally managed, and instructor training was provided in relation to content delivery and assignment evaluation using the analytic rubric. The scores assigned using the rubric were also used to determine the grade on the assignment. Given the structure of the course sequence, the design of the distributed content, and the use of an analytic rubric across sections and assignments, the fall-spring curriculum provides a structured environment that should support the creation of a corpus of scores that can be trusted.

The scores shown here are the average of the eight sub-component scores from the rubric. The use of an average score is justified because (1) student grades are based on a similar average, and (2) a principal components analysis showed that the sub-scores were highly correlated (Moxley & Eubanks, 2016). Figure 1 shows the longitudinal average scores.

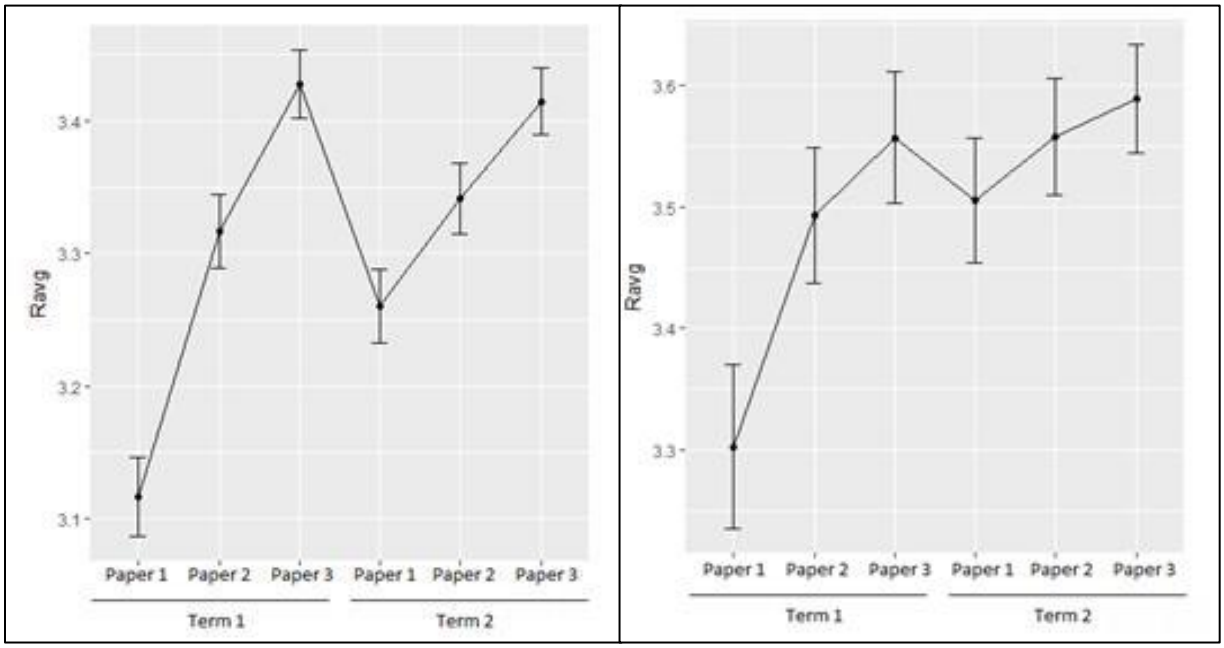

Figure 1

Average Scores for Students Over a Fall (Term 1) and Spring (Term 2) Term, With Three Standardized Major Project Assignments Each Term, Showing Two Standard Errors of the Estimate

The left graph (N = 1,887) in Figure 1 includes only students who had data for all six papers, and the right graph (N = 892) further limits the data to only those students who also had the same course instructor both terms to account for student-instructor biases that may exist.

The score averages in both graphs of Figure 1 show an increase in scores across the three standard assignments for the fall term, followed by what we term a *reset* to a lower value for the spring, after which growth resumes. The separation of the error bars shows statistically significant differences for all students from one assignment to the next (left graph). When the data are filtered to students who had the same instructor for both terms, the averages statistically plateau at the end of the first term.

Two patterns are clear: a pattern of measured increases within both courses and a reset between the two courses.

4.0 Interpretation

At face value, the score averages in Figure 1 suggest that the spring course may be unnecessary. Students finish the second term at the same average level they finished the first term. Worse, the reset finding could be interpreted as *undoing* prior learning. Given the complexity of teaching and learning in general and specifically in first-year writing, it is unlikely that those raw interpretations of the graphs are reasonable, but they make the important point that score averages cannot be taken at face value. We identify four factors that could cause the score reset.

First, we recognize that rating styles impact scores (Wind & Englehard, 2013). The reset could indicate an initial norming of grading practices and expectations between students and instructors or raters. The presence of the reset for students with the same instructor across both terms suggests that the norming would extend beyond an initial calibration and exist as a structure within each section and term. A narrative template indicated by the reset could reflect instructors who begin the term as difficult graders and then ease off across the term as a result of their own workload or fatigue. Some raters may increase scores across the term intentionally to improve student satisfaction in order to produce stronger student evaluations or to avoid grade grievances. The change could also reflect instructor efforts to develop confidence in students or indicate new raters developing confidence in their own scoring abilities.

A second possibility is that student learning is not what is being measured, or at least not all that is being measured. Instructors could be giving credit for extrarubric considerations such as attitude and effort or the perseverance and grit demonstrated by continuing to submit work. Since scores contribute to grades in this study, it is likely that instructors want to align scores and grades even if it means subverting the rubric. The limitations of the scoring criteria included in the rubric are readily apparent when we view standardized trait and holistic scores in light of the cognitive model of writing presented by MacArthur and Graham (2016).

A third option is a ceiling effect, which is most evident in the right graph of Figure 1 where the same instructors rated the same students. The ceiling effect suggests that scores may be clipping at the top end, causing raters to be impacted by the scale limitations. Extending the scoring scale may or may not impact the ceiling effect; however, the 4.0 grading model creates inherent limitations regardless of rubric or scores. A ceiling effect would serve to hide an even greater reset than we see in Figure 1.

Finally, assignment difficulty, as suggested in Sullivan and McConnell (2018), could contribute to the contextual use of the analytic rubric. The sequence structure is focused on process and genre and designed to scaffold skills, but while a complexity increase is intended, our data set does not consider this variable explicitly. Assignment difficulty, even as relevant to each student, could also impact rater mediation, especially when compounded by considerations such as effort and attitude.

Other possibilities are likely, and multiple factors could be impacting the scores. What is clear is that assumptions about score use, slopes, and rubrics all come into question through contextualization of scores beyond the too-simple trope that averages measure learning.

5.0 Conclusion

The score reset deserves replication and extended investigation, but we have identified preliminary lessons connected to score interpretation and use. To begin, a score interpretation and use argument should not be made based on raw score averages. Summaries of scores and rubric ratings require contextualized interpretation and analysis, which may include information about assignment difficulty and student demographics among other confounding factors, but any use deserves a general recognition of unidentified confounding factors that will always impact the scores.

Rather than simple averages like the ones in Figure 1, statistical models are needed (Eubanks & Vanovac, 2020). For example, hierarchical models can account for rater styles when estimating student abilities and can be used on incomplete data sets, where not every student has every observation recorded. Additionally, to the extent possible, researchers should align individual aims to the measurement tools, especially considering the relative versus absolute goals.

For an assessment to determine placement, absolute measure of competency is necessary. For measuring longitudinal change, however, relative measures may be more useful. One complication is that we suspect many putative absolute measures that depend on human raters inadvertently measure relative differences between students observed in a way that contaminates the aim. In other words, it is difficult to assess a piece of written work solely against an abstract standard without comparing the work to that of other students, which, again, highlights the broad impact of context and contextualization.

While such calibration is relevant in a classroom setting, it also impacts broad, institutional assessments, such as general education. The potential implications of a reset should be considered within conversations around prerequisites, sequences, and transfer. Understanding the trajectory of learning across and beyond the term and clarifying the role of scores and samples in determining growth and evaluating development is important for students and instructors, as well as programs, departments, and institutions. Conversations around norming may also need to expand to include these considerations. Such a complication is compounded when we include cognitive frameworks for writing that acknowledge multiple domains and interpersonal factors such as self-efficacy and motivation.

Finally, data extraction, which impacts score interpretation and use, also requires contextualization. Whether the result of rater behavior, scaffolded difficulty, or other complicating factors, *when* scores are pulled impacts *what* scores are pulled. Claims made based on score interpretation and use would vary depending on whether the scores for our students were pulled during the reset or at the end of the term.

In addition to the preliminary findings that speak to interpretation and use, we also recognized actionable information related to score generation. Our interpretation of the data confirmed that the generation of scores impacts interpretation and use and must be reconsidered. The claim that inference requires crucial background is not novel (Xiao et al., 2017), but

supporting such score generation and constructing sets of scores that provide the needed context requires significant modification of standard practices.

In response to our findings, specifically to implications of conflating grades and assessments, our program redesigned student evaluations to segregate rubric use from grades. Stronger data is one intended outcome, but actionable feedback for students is the main goal. When scoring is grading, instructors may feel limited in the feedback they can offer because students see the feedback primarily as a justification for the score instead of a learning opportunity. Rater-graders may feel that to provide critical, formative feedback requires assigning a lower grade than they believe the work deserves. While the distinction may be clear from where we sit, students would be confused by what they could see as competing or disconnected scores, grades, and feedback.

Why would we expect students not to accept the traditional narrative of scores as measurements of learning? If extensive positive feedback were accompanied by scores and/or grades that did not align with the feedback narrative, the message being sent to students would be lost in translation. Similarly, if students saw their scores and/or grades continue to increase despite their decreasing efforts and perhaps even progressively critical feedback, the increasing scores may simply disincentivize responding to feedback.

Our redesign also introduced the inclusion of extrarubric criteria within grading practices, which allows instructors to value student competencies and characteristics that contribute to their learning and to the learning of their peers and to discuss their impact and importance. If instructors were trying to incentivize positive attitudes, encourage effort, reward persistence, or build confidence by considering extrarubric criteria in their grading practices, we believe they were right to do so and that doing so explicitly and transparently can support student learning and improve instructor scoring. Making space for extrarubric criteria allows us to include their demonstration and development within the goals and outcomes for the course**.**

Reimagining the curriculum also expanded relevant content for instructor development and training. Acknowledging the segregation between scoring and grading can create space for pedagogical conversations that question and clarify the goals and roles of both, and these conversations can take place between administrators and instructors and between instructors and students. Further, recognizing that increasing scores do not necessarily signify learning allows us to entertain the possibility that an absence of high or increasing scores does not necessarily signify the absence of learning.

Exploring new ways to measure learning can inspire new ways of learning and teaching and help both teachers and students recalibrate their narratives around scoring and learning. Ultimately, contextualizing our findings and potential interpretations informed our curricular redesign and encouraged us to expand the scope of learning we value and focus on delivering meaningful and fair, formative feedback aimed at developing, in addition to demonstrating, student learning.

6.0 Directions for Future Research

Scores are used to make claims about individual students that ground forecasting, admissions, placement, intervention, and grades. Increasingly, scores also impact institutional funding, rankings, assessment, and evaluation. Given the vast and expanding use of scores, significant consideration should be given to their construction and interpretation.

Our findings suggest that measuring learning may not be as straightforward as assumed. The reset suggests that the assumption of score increases as indicative of learning is complicated by other factors, but with contextual analysis, scores can produce invaluable information and feed innovative program evaluation. Stated simply, simple narratives of scoring slopes and student learning may unwittingly mask complex interactions.

Large-scale conversations on score valuation and validation require contextualization within and beyond the work of writing analytics. Many institutions use rubrics and scores to make determinations that impact students. Such considerations should also apply to VALUE rubrics and other academy-wide efforts.

Statistical modeling is needed to better address fairness and justice in education and educational measurement. Score use can limit access to education. We cannot hope to understand systematic barriers without considering the factors that predict those barriers within our calculations, including academic preparation of incoming students, race, gender, ethnicity, Pell eligibility, first generation and/or transfer status, and other demographic factors that can identify vulnerable populations.

Specifically, the presence of a reset complicates assumptions related to rubrics, scoring, and score interpretation and use; it may even complicate assumptions around teaching, learning, and transfer. Without doubt, it raises questions that deserve exploration. Factors such as rater mediation, extrarubric considerations, assignment complexity, scoring as grading, prerequisite models and transfer, and data collection contextualization can inform replication efforts, generate supplemental questions, and frame subsequent research.

Author Biographies

Alaina Tackitt serves as the Associate Director of First Year Composition at the University of South Florida.

David Eubanks serves as Assistant Vice President for Institutional Effectiveness at Furman University.

References

Eubanks, D., & Vanovac, S. (2020). Divergent writer development in college. *The Journal of Writing Analytics*, *4*, 15-53.

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing*. Southern Illinois University Press.

Haswell, R., & Elliot, N. (2019). *Early holistic scoring of writing: A theory, a history, a reflection.* Utah State University Press.

Kane, M. T. (2013) Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1-73.

Kellogg, R. T. (1993). Observations on the psychology of thinking and writing. *Composition Studies, 21,* 3-41.

MacArthur, C. A., & Graham, S. (2016). Writing research from a cognitive perspective. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 24-40). Guildford.

Morse, R.., & Brooks, E. (2020). How U.S. News calculated the 2021 best colleges rankings. *U.S. News & World Report*. Retrieved August 2, 2020, from https://www.usnews.com/education/best-colleges/articles/how-us-news-calculated-the-rankings

Moxley, J., & Eubanks, D. (2016). On keeping score: Instructors' vs. students' rubric ratings of 46,689 essays. *WPA: Journal of the Council of Writing Program Administrators, 39*(2), 53-80.

Pagano, N., Bernhardt, S. A., Reynolds, D., Williams, M., & McCurrie, M. A. K. (2008). An inter-institutional model for college writing assessment. *College Composition and Communication, 60,* 285-320.

Rourke, L., & Zhou, X. (2018). When scores do not increase: Notes on quantitative approaches to writing assessment. *The Journal of Writing Analytics, 3,* 264-285. https://wac.colostate.edu/docs/jwa/vol3/rourke.pdf

State University System of Florida. (2020). *Performance-based funding*. Retrieved August 2, 2020, from https://www.flbog.edu/finance/performance-based-funding/

Sullivan, D. F., & McConnell, K. D. (2018). It's the assignments—A ubiquitous and inexpensive strategy to significantly improve higher-order learning. *Change, 5,* 16-23.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Utah State University Press.

Wind, S., & Englehard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing, 18,* 278-299.

Xiao, Z., Higgins, S., & Kasim, A. (2017). An empirical unravelling of Lord's Paradox. *Journal of Experimental Education, 87,* 17-32.