Research Note

A Line Matching Method for Reliable Higher-Order Theme Identification

Michael B. Rose, *Furman University*

Suzanne Klonis, *Furman University*

Structured Abstract

- **Aim:** The purpose of this research note is to demonstrate a novel method for coding samples of written student work and then to assess the reliability of that coding method. Commonly, writing samples are coded using rubrics for items like style, organization, accuracy, and so on, with each item rated on a scale from 0 to 4, where the result is a summative score for the entire work. In our method, coders identify the line numbers *within* a paper that show evidence of some theme or element of interest. This method can be particularly useful for identifying higher-order themes such as irony or self-reflection. We demonstrate its use with reflection essays where we hope students demonstrate self-growth from a high-impact experience. This new type of data (a list of line numbers, per paper and per rater) requires us to test new formulations for inter-rater agreement.

  We will demonstrate that (1) the evaluation process is useful for rater coding of higher-order themes and (2) reliability of the ratings can be easily measured and compared.

- **Problem Formation:** We recognize a lack of simple methods that do not depend on costly software or intense training to reliably extract specific elements from an open form essay. We seek to address this gap by proposing a technique to identify elements from an essay in a manner that can be accurately measured for reliability.

- **Information Collection:** A corpus of 152 reflection essays detailing undergraduate summer internships and research experiences was collected. After essays were redacted and line-numbered, pairs of student raters read them for the presence of specific thematic elements. The line numbers containing those elements were recorded and subsequently analyzed for agreement compared to random chance agreement.
- **Conclusions:** We found that raters were able to identify and agree upon desired elements in open form essays at significantly higher rates than random. Specific, concrete language positively impacted the raters' ability to agree on the location of these essay elements.
- **Directions for Further Research:** After review of the collected data, we found that particular raters tended to mark more liberally (i.e., selecting long text excerpts) than others and that some writers were less clear in their written expression; thus, we recognized the need for more tools to disentangle the conflating effects of writer clarity from the severity of the rater. Suggestions are offered to increase the agreement between raters. We also discuss the possible application of this method to the evaluation of other forms of writing.

Keywords: coding, Cohen's Kappa, inter-rater agreement, reliability, rubric, text analysis, theme identification, writing analytics

1.0 Aim

The notion of "high impact practices" (HIPs), or student experiences thought to have especially high learning gains, is now widely accepted in higher education (Kuh, 2008). These so-called HIPs include internships, undergraduate research experiences, and studies abroad. How are we to assess the impact of these experiences? One attempt uses longitudinal standardized instruments (Kilgo, Sheets, & Pascarella, 2015), but the statistical summaries of the gains are so general as to be not very useful for understanding impact at the level of the person or for fine-tuning programmatic experiences.

Another approach is to use reflection essays that prompt students to reflect on their growth experiences through their HIPs. In our case, we would like to know if the student (1) had a personal growth experience with respect to their career goals, (2) applied classroom learning to a real-world problem (or "integrative learning"), or (3) gained a sense of life purpose. During one summer session of research and internship experiences, we collected 152 such essays.

A campus research group initially tried rating the essays with a rubric to produce an overall score for each of three outcomes per essay (sense of life purpose, career goals, integrative learning). Despite revising and testing the rubric several times, the group was unsatisfied with the results. The formal inter-rater statistics were very low, and it was clear from discussion that conceptual agreement was far off as well. So, we tried something new, inspired by Hathcoat's

(2018) question “What’s the meaning of a zero?” Hathcoat’s work attempts to disambiguate the case where a writing sample has *evidence of non-performance* versus a case where the writing sample *lacks evidence to make a decision*. This distinction is analogous to the difference between a rubric rating of 1 (the lowest scale choice) and NA (a blank in the data set).

In our “is it there?” attempt, instead of coding a paper along predetermined rubric dimensions (like organization and style), student raters were asked to read for evidence of three outcomes and record the line numbers of this evidence. This produced a spreadsheet with document identifiers running down the left column and each rater’s responses in the form of line ranges, like 31–42, for each of the three outcomes. A blank indicated that the outcome was not found within the paper, and multiple line ranges pointed to more than one occurrence.

It was immediately obvious that this method had much greater agreement than any of the rubric rating attempts, so we scaled up the rating process to employ students to help with the coding process. For our corpus of 152 documents, two students rated each paper, reading for evidence of three outcomes. We encourage others to try this method to investigate its efficacy on other types of writing.

2.0 Problem Formation

Our original research question was “How do we assess the efficacy of high-impact practices?” More generally, how can we identify excerpts of a text that speak to a complex outcome or higher-order theme? The latter task is commonly done, for example, in the context of grounded theory or text-mining, by marking and tagging passages. This process requires skill and sometimes custom software. How can we get similar benefits without intensive training or expensive software? And once we have data, how do we know that it is any good?

Our alternative method yielded high agreement between raters in the segments of texts that they selected as showing evidence of important kinds of impact. These text segments can then be easily extracted for further analysis. Although software packages are available to calculate inter-rater reliability in general, these packages typically rely on statistics discussed below that are not ideal for unbalanced data. As the portion of the essays which exhibit higher-order themes is much smaller than the portion devoid of those themes, unbalanced data is common in this project. We introduce a new type of analysis that is not currently available in software packages, is not as negatively impacted by unbalanced data, and yet is simple to implement. Students at our university who had just finished internships or research experiences were asked to write about how the experience involved “integrative learning” (applying knowledge gained through coursework to real-life settings) or how it influenced what they wanted to do with their lives. For example, one student wrote in a reflection essay after an internship (quoted with permission):

> This experience overall forced me to step back and think long-term. I realized that I need to work in an environment that does so by bettering others. I crave those connections with individuals in a way so that I can improve something about their existence.

This passage was selected by both raters to illustrate a developing sense of purpose with regard to career goals (in this case, finding a career that helps other people).

Confidence in the data requires some measure of reliability. It is unreasonable to trust either rater's assessment of the essays if the amount of agreement is not at least greater than rating at random. Rater agreement is often measured with a kappa statistic (Cohen, Fleiss, etc.) that adjusts actual rater agreement for chance to create a 0–1 index of "better than chance" agreement. For more context on rater agreement in writing assessment, see Eubanks (2017) and Ross and LeGrand (2017). A recurring problem with kappa-type statistics is the role of the underlying distribution of ratings. In our case, there are only yes/no marks on individual lines of essays (indicating "yes" there is evidence for one of the three outcomes or "no" there is no evidence). These are unbalanced in that there are many more "no" than "yes" marks. Such data sets are known to cause problems with kappa statistics, so we sought another way to assess reliability.

One challenge in calculating inter-rater reliability is to summarize agreement between text excerpts chosen by raters that might differ in length (two lines of text chosen as evidence by one rater versus ten lines chosen by another) and overlap (choosing excerpts as evidence that have some but not all lines in common). We experimented with different solutions to the "overlap" problem, but here only present a simpler exact-line matching agreement statistic. The line matching method compares the observed rates of *unanimous* line marking to the rates expected from two raters marking purely at random as described by a binomial distribution based on the overall frequencies of lines marked within the corpus for that outcome. This can be visualized by plotting the observed rate of unanimous marking (the match rate, in other words) with the expected rate if the raters were randomly selecting lines. This keeps the spirit of chance-corrected metrics like the kappas, but without incurring the problems of unbalanced sets. The comparison we show below can be formalized as a *t* test of differences in proportions between observed and expected (if random) rates.

3.0 Information Collection

Each summer, undergraduates have opportunities to engage in high-impact practices like research, study away, and internship experiences. Internship and research students were given the same reflection prompt at the end of their experience:

> Please take some time to write a short reflection on the Engaged Learning Experience you had this summer, focusing on how the experience contributed either to your Integrative Learning or your Sense of Purpose (see prompts for each below).
>
> **Integrative Learning**
>
> What are the connections between what you learned from this experience and what you learned in the classroom (e.g., academic concepts, theories, skills, etc.)? Please describe the connections between coursework and the experience in

depth, including identifying the course subjects you applied. Consider any instances during the experience where:

- you applied or integrated knowledge or skills from coursework,
- your experience was deepened by the application/integration of your coursework knowledge,
- you realized something new about your coursework or an academic discipline.

Sense of Purpose

How did the experience affect your understanding of yourself? Did it affect your sense of purpose, including your current educational/personal goals and/or possible post-Furman professional and career plans?

Prior to the essay rating step, lines in each reflection were numbered in order to allow raters to easily record the location of the evidence of gained insight within the essay. Names of the writers were also redacted. The "Sense of Purpose" category was subdivided into two separate themes: "Sense of Purpose - Life Goals" and "Sense of Purpose - Career Goals." Three raters, after being trained to identify evidence of the three themes, interchangeably worked in pairs on each essay, identifying text selections that explicitly mentioned the themes above.

To indicate evidence (or lack thereof) of "Sense of Purpose – Career Goals," "Sense of Purpose – Life Goals," and "Integrative Learning," raters indicated "Yes" or "No," and for those essays that were scored "Yes," they gave line numbers to indicate where evidence of the insight category could be found. An example of a rated essay is shown below in Table 1.

Table 1

Detail of Rating Data from Essay 1, Rated by AZ and HY

			Career Goals				**Life Purpose**				**Integrative Learning**			
Doc	Type	Rater	Present?	Lines			Present?	Lines			Present?	Lines		Nlines
Essay1	research	AZ	Yes	7-9	45-47		No				Yes	9-11	34-36	42
Essay1	research	HY	Yes	21-32	41-43	45-47	No				Yes	35		42

Note. Both raters found evidence of Career Goals and Integrative Learning, but not Life Purpose. There is overlap between some but not all marked selections. Nlines is the length of Essay 1.

For evidence in each category, each line of an essay has zero, one, or two marks, indicating the findings of the pair of raters. The following process was used in each of the three outcome categories: Out of the set of lines marked for evidence by at least one rater, the proportion of unanimously marked lines was calculated. The expected proportion according to a binomial distribution was also calculated for hypothetical random raters (the null hypothesis). The standard errors were found for both observed and expected proportions using $\sqrt{\frac{p(1-p)}{N}}$, where N

is the number of lines marked at least once in the outcome category (e.g., "Life Purpose"). The observed match proportion is the number of cases where both raters marked the same line, divided by *N*. Table 2 below lists the observed proportions.

Table 2

Proportion of Matched Lines

Outcome	Matched Lines	N Marked Lines	Observed Match Proportion
Life Purpose	33	146	0.226
Career Goals	340	1,051	0.324
Integrative Learning	632	1,264	0.500

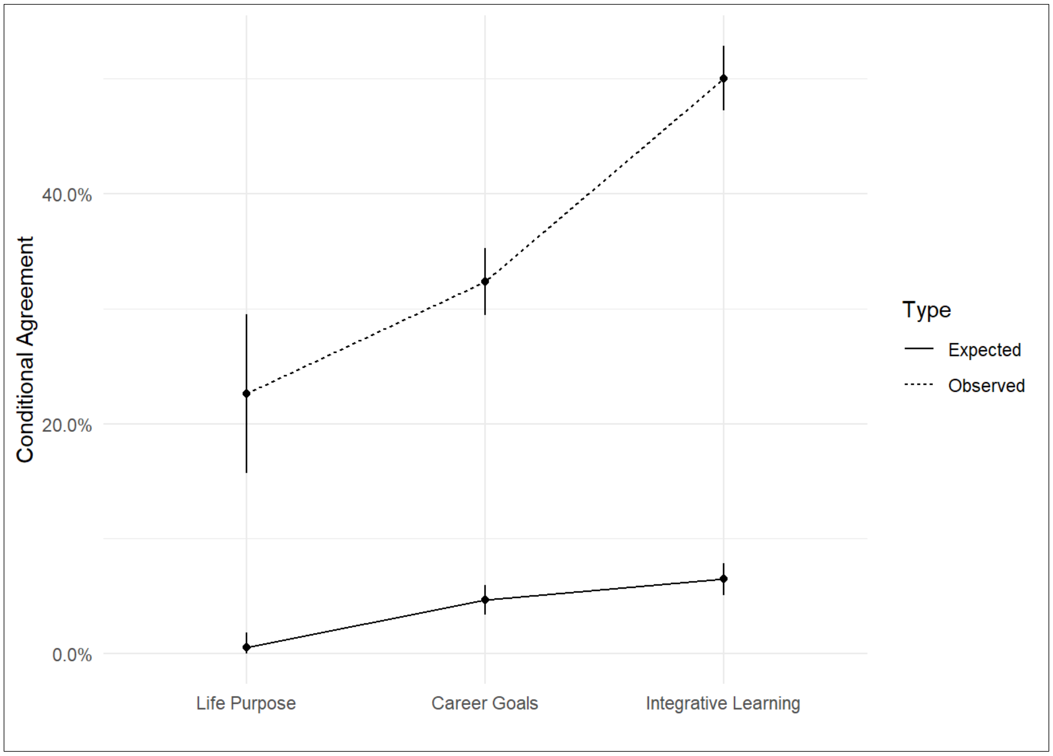

Figure 1

Proportion of Unanimously Marked Lines out of the Set of Lines with at Least One Marking

Note. *Expected proportions are calculated using the overall marking rate of raters but assume random marking. Vertical lines show 95% confidence intervals.*

Figure 1 takes into consideration only the set of lines that were marked at least once. Lines marked by both student raters are reported as a proportion of the lines marked at least once. As an example, approximately half of the lines marked as evidence of "Integrative Learning" were agreed upon by both student raters. These proportions of doubly marked lines ("observed" on the graph) are compared to random expected values and standard errors. The graphical comparisons in Figure 1 show large gaps between observed and expected match proportions for all three categories. In addition, the standard errors calculated show that observed and expected values are well separated, and formal comparisons have *p* values that are essentially zero. The line identification is not just non-random; it's useful: reading selections where student raters agree on a case makes it clear that the identified passages truly do reflect the outcome categories.

4.0 Conclusions

We developed a simple method to extract specific elements of undergraduate high-impact experiences from open form essays. The line-identification method demonstrates substantial reliability. Categories such as integrative learning prompt the use of specific keywords, such as class names or easily identifiable subject matter. We believe that concrete language and keywords increase the rate of inter-rater agreement by making the presence of themes easier to detect. This specific language is less likely to appear in excerpts describing a sense of life purpose. We believe a more abstract category such as "Life Purpose" results in fewer identified lines, as the language used to describe a meaningful experience in this regard is more ambiguous and less likely to be perceived as evidence. We observe in the "Life Purpose," "Career Goals," and "Integrative Learning" categories respectively that, as concepts increase in specific language, agreement also tends to increase. All rates of unanimous agreement above random were significant ($p < .001$). The reliability statistics combined with inspection of the selected passages give us confidence in the results.

This method can be utilized to evaluate multiple writing forms. A research team conducting a meta-analysis might find the method useful in developing reliable consensus during review of literature. In institutional assessment, instructors might be able to gauge the presence of specific writing abilities, such as use of irony.

5.0 Directions for Further Research

While successful in reliably identifying specific elements in an essay, this process revealed several obstacles and opportunities for improvement. In this project, we had no way to be certain that when raters indicated no evidence for impact (on Life Purpose, Career Goals, or Integrative Learning), this meant that evidence truly was absent, rather than just obscured by a lack of clear writing. We also took note of the different rates of evidence marking that specific raters exhibited. We have begun efforts to separate the effects of writer clarity from differing levels of severity inherent in specific raters using a general linear model with mixed effects. We will investigate how factors such as average word length and length of essay are associated with writing clarity and line-marking frequency of the raters. We intend to add a new rating of "richness" describing "how much" or "how profound" the evidence is within a marked line. This added dimension will accommodate measurements of presence, location, and degree of an essay element.

Rater training could be enhanced to improve reliability. While student raters were given training about what themes to read for, they were not told to limit the number of lines they cited as evidence. To improve reliability, we could instruct coders to only include lines with direct evidence, as opposed to including neighboring lines that give context for the evidence. In our own case, specific to itn, we found that one rater tended to mark lines detailing everything an author had learned instead of exclusively marking lines showing how the author applied the coursework to an internship. It seemed to us that raters differed only slightly in the amount of material surrounding the evidence they included, while they consistently targeted the same general locations of the evidence. A follow-up study could explore the sensitivity of the raters to training by varying scoring guidelines about how much evidence and context surrounding that evidence to include. We believe that rater agreement would not be significantly affected by this type of training. We assume that concrete keywords and clear writing contribute more to rater agreement than stringent rater training. In our specific writing context, it would be of future interest to request a sample of authors to mark their own lines of evidence. Raters' markings could then be compared to the original author's line markings, and agreement could be calculated. The author's line markings would help train raters to an acceptable threshold by having access to the exact lines the author intended to convey as evidence. The essay prompt and rating rubric could be manipulated and clarified to maximize inter-rater agreement by examining rate of unanimous agreement before and after change of rater guidelines.

The ability to accurately classify student experience essays is an important tool in understanding the impact of experiences. One can imagine that summer research and internship quality only scratch the surface of areas of student life that need to be communicated clearly to an institution. In order to effectively respond and adapt to the needs and experiences of students, more work must be done to differentiate the conflation of writing clarity, rater severity, and essay prompt alignment. We believe that this method is a reliable step in that direction.

Finally, the rater agreement statistics in Figure 1 understate the actual level of agreement, since we only count exact matches. In many cases, the start and end lines for a passage are debatable, and a more liberal matching algorithm could yield better results.

Note

Not all methods used in this evaluation are novel by themselves. We contend that their combined use and utilization of individually marked lines by student raters offer a new and inexpensive method amendable to reliability measures to extract complex learning outcomes from essays.

Acknowledgements

We are deeply grateful to David Eubanks for his insight and guidance in this project as well as his contributed R code. We would also like to thank Jimmy Chen, Ada Zhao, and Haining Yan for the many hours spent in the collection and coding of the data. Lastly, we are grateful for the contribution from the undergraduate authors of the essays.

Author Biographies

Michael B. Rose holds a Bachelor of Science degree in physics from Furman University. His interests include data science, scientific programming, and physics-based modeling. A biography is available at www.michaelbrose.com.

Suzanne Klonis is the Director of Assessment at Furman University. Her responsibilities in this role include: university survey design, administration, and analysis; assessment of high-impact engaged learning experiences; and coordination of academic assessment activities. She holds a PhD in Social Psychology from the University of Wisconsin-Madison.

References

Eubanks, D. (2017). (Re) Visualizing rater agreement: Beyond single-parameter measures. *The Journal of Writing Analytics*, *1*, 276–310. https://wac.colostate.edu/docs/jwa/vol1/eubanks.pdf

Hathcoat, J. D. (2018). The role of assignments in the multi-state collaborative: Lessons learned from a master chef. *Peer Review*, *20*(4). https://www.aacu.org/peerreview/2018/Fall/Hathcoat

Kilgo, C. A., Sheets, J. K. E., & Pascarella, E. T. (2015). The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education*, *69*(4), 509–525.

Kuh, G. D. (2008). *High-impact educational practices: What they are, who has access to them, and why they matter*. Association of American Colleges and Universities.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *The Journal of Writing Analytics*, *1*, 227–275. https://wac.colostate.edu/docs/jwa/vol1/ross.pdf