12-14-2018

# Scaling Undergraduate Scientific Writing via Prominent Feature Analysis

Katarzyna Zaruska Gallo

# Scaling undergraduate scientific writing via Prominent Feature Analysis

By

Katarzyna Zaruska Gallo

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Educational Psychology
in the Department of Counseling, Educational Psychology, & Foundations

Mississippi State, Mississippi

December 2018

Scaling undergraduate scientific writing via Prominent Feature Analysis

By

Katarzyna Zaruska Gallo

Approved:

_____

David T. Morse
(Major Professor)

_____

Anastasia D. Elder
(Committee Member)

_____

Tianlan (Elaine) Wei
(Committee Member)

_____

Sherry S. Swain
(Committee Member)

_____

Carlen Henington
(Graduate Coordinator)

_____

Richard L. Blackbourn
Dean
College of Education

Name: Katarzyna Zaruska Gallo

Date of Degree: December 14, 2018

Institution: Mississippi State University

Major Field: Educational Psychology

Major Professor: David T. Morse

Title of Study: Scaling undergraduate scientific writing via Prominent Feature Analysis

Pages in Study: 138

Candidate for Degree of Doctor of Philosophy

Prominent Feature Analysis (PFA) is a reliable and valid writing assessment tool, derived from the writing it is used to assess. PFA, used to assess on-demand expository essays in Grades 3-12, uncovers positive and negative characteristics of a sample. To extend PFA to a new academic level and genre, I assessed scientific writing of 208 undergraduates, identifying 35 linguistic and 20 scientific prominent features. An essay could earn up to 28 positive (24 linguistic and four scientific), and up to 27 negative marks (11 linguistic and 16 scientific). The minimum prominent features number in a paper was 3, the maximum was 25 ($M$ = 12.45, $SD$ = 3.88). The highest positive and negative prominent features numbers noted were 17 ($M$ = 4.11, $SD$ = 3.96), and 16 ($M$ = 8.34, $SD$ = 3.25) respectively.

Rasch analysis revealed a good data-model fit, with item separation of 5.81 (.97 reliability). The estimated feature difficulty of items spanned over 10 logits; common errors were easier to avoid than “good writing” characteristics to exhibit. Significant correlations among linguistic, but not between linguistic and scientific features, suggest writing proficiency does not assure excellence in scientific writing in novices. Ten linguistic features significantly strongly and moderately inter-correlated with each other,

appearing to represent writing proficiency. Student GPA correlated significantly with the raw prominent features scores ($r = .37$; $p < .01$), and negatively with the sum of negative linguistic features ($r = -.40$, $p < .01$), providing support for scale's validity, and suggesting that good students are better at avoiding common writing errors than less able learners. Additionally, PFA scores positively significantly correlated with composite ACT scores.

To investigate PFA's ability to track change in writing over time, I compared 2 sets of prominent features scores of 25 students. In comparison with earlier essays, later (longer) essays exhibited significantly more positive, *and* more negative features. Prominent features scores did not correlate significantly between the sets. This suggests, that while PFA is a valid and appropriate tool for analysis of undergraduate scientific writing, it was not suitable for tracking change in writing ability in this small sample.

DEDICATION

For everybody who wants to know what makes good writing good.

"Poets say science takes away from the beauty of the stars—mere globs of gas atoms. I too can see the stars on a desert night, and feel them. But do I see less or more? The vastness of the heavens stretches my imagination—stuck on this carousel my little eye can catch one-million-year-old light. A vast pattern—of which I am a part… What is the pattern, or the meaning, or the *why*? It does not do harm to the mystery to know a little about it. For far more marvelous is the truth than any artists of the past imagined it. Why do the poets of the present not speak of it? What men are poets who can speak of Jupiter if he were a man, but if he is an immense spinning sphere of methane and ammonia must be silent?"

—Richard Feynman; drummer, theoretical physicist, Nobel Prize laureate

ACKNOWLEDGEMENTS

Numerous individuals supported me in completion of this project. Thank you, Dr. David Morse for your *time* and patience in explaining what I needed to do, why, and how. Your ability to clearly convey complex scientific content is unmatched by most. Thank you, Drs. Anastasia Elder and Elaine Wei for your helpful comments on the drafts of this manuscript. This document would be much harder to follow without your guidance. Thank you, Drs. Sherry Swain and Richard Graves for contributing your substantial expertise to the linguistic analysis of the project's writing samples. It has been an honor and a true pleasure to work with you, side by side. I have learned more about good writing and writing assessment from you, in such short time, than I could ever dream. Thank you, Dr. Gary Bradshaw, and your teaching assistants, Sarah Craig, Kara Nayfa, and Daniel Roberson, for allowing me to solicit permission of students in PSY 3314 to use their writing in my study, and for explaining to me how the assignments were structured. Thank you, Mary Emma Peavy and Belinda Strauss, for helping me with data management. Thank you, sister, parents, step-parents, and parents-in-law, for cheering me on for the past six years. Dziadku Mikołaju, I wish you could see me now! Thank you, my amazing-talented-hard-working-fabulous students, whose writing I have the pleasure to routinely read. You sure keep me on my toes! Last but first, thank you, Cory and Julian Gallo, for holding down the fort while I was "away," either in body or spirit. I love you, and I couldn't have done any of this without you.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

# CHAPTER I

# INTRODUCTION

The act of writing anchors students' educational experiences from the time they first learn to hold a pencil in preschool, to writing doctoral dissertations or professional licensing examinations, and beyond. Writing may serve as a mean of formative evaluation of learning, and, subsequently, inform instruction. For example, an instructor may require a quick, in-class writing task to check for understanding of the recently-presented material. If it becomes apparent that the material was misunderstood, a follow-up lecture may be in order. Writing may serve as a means for summative learning evaluation. Students compose term papers to demonstrate mastery of the course material, or showcase their writing and thinking abilities on high-stakes standardized tests, like the Scholastic Aptitude Test (SAT), or the Graduate Record Examinations (GRE). Lastly, writing assessment may be used for placement purposes, for example, a student may be able to skip a Composition I course in college if he or she scores high enough on a high-school Advanced Placement examination. Given writing's prominence in elucidating student thinking and learning, it is no wonder that writing assessment is often highly structured, standardized, intensely studied, and, occasionally, highly contentious.

## Writing Assessment Types

Writing assessment experts agree there is no single best way to evaluate a writing sample (e.g., Huot, 1990). Writing can be evaluated in its' entirety ("this is an 'A'" essay;

holistic assessment; see Appendix A for a holistic scale example), or in context of its different aspects to better understand its strengths and weaknesses (analytic assessment; see Appendix B for an analytic scale example). These two broad categories serve different purposes; while holistic scores are often used for placement and achievement assessment, the analytic models can be invaluable as "in progress" tools for identifying student struggles on individual and class levels, thus guiding instruction (Shohamy, Gordon, & Kraemer, 1992; Swain & LeMahieu, 2012). Some separate analytic from trait analyses, as the former focuses on the quality of the language use, while the latter notes different qualities of the content (stance, voice, etc.). Others prefer a simpler division of "sense of the whole"/"sense of parts" (Huot, 1990). One important point to note is this: a sum of points assigned on analytic scheme is commonly called a "holistic score." However, it is not the same thing as a holistic scoring scheme, which focuses on "taking in" the writing sample all at once, and assigning a single score to it. Both holistic and analytic scoring schemes have strengths and weaknesses; these are discussed below, along with an introduction to a relatively new scoring scheme, the Prominent Feature Analysis.

**Holistic Scoring**

Holistic scoring allows for a piece of writing to be evaluated in its entirety (along a few guiding parameters, for example, a quality of analysis, or organization), and its proponents argue that as writing can be viewed as an art form, it should be evaluated like one may evaluate a piece of art. Separating Michelangelo's David into individual body parts and evaluating them separately would likely not yield the same results as evaluating the intact sculpture (White, 2009). However, researchers have identified multiple

concerns with the holistic approach. Holistic scores may correlate with length and appearance, for example, longer or neatly-written essays may earn higher marks than their shorter or messier-looking counterparts, regardless of their content. The scores have poor transferability, for example, holistic scores assigned to a National Assessment of Educational Progress samples one year may not represent the same quality of writing as the scores assigned on a different year. Additionally, the process of reading to score holistically may alter the reader's thinking about the writing quality, as he or she tends to focus on the features identified in the rubric, and disregard other characteristics which may be present in a writing sample (Huot, 1990).

The holistically-scored GRE contains a writing component comprised of two analytical writing tasks; their scores are averaged. The scale ranges from 0 to 6; the points are assigned in 0.5 increments. A score of 0 signifies that the generated writing does not address the question; a score of 6 indicates the highest degree of writing proficiency. The evaluated dimensions include: idea analysis, development of persuasive arguments and examples, focus and organization, usage of correct sentence structure, sentence variety and precise vocabulary (Educational Testing Services, 2018). See Appendix A for a description of GRE writing scores.

**Analytic Scoring**

Analytic scoring relies on generating separate sub-scores for sub-skills. An advantage of this assessment scheme includes a potential to identify specific underlying weaknesses in writing, thus informing instruction. Additionally, having the means to identify levels of expertise along individual sub-categories may increase the validity of this scheme (Bang, 2013). However, while the sub-categories may provide additional

layers of information about a writing sample, creators of analytic schemes often disagree on what these sub-skills are. Results of past studies also demonstrate that it is hard to obtain high inter-rater reliability using analytic schemes. Lastly, as the schemes tend to be complex, rating tends to be time-consuming, and, therefore, expensive (Huot, 2009).

The Analytic Writing Continuum (AWC) is an example of an analytic rating scheme. The tool was developed by the National Writing Project, and is used by teachers in grades K-12. The developers were inspired by the Six + 1 Trait Writing Model; six distinct traits are evaluated and assigned separate scores. Additionally, an independent holistic score is assigned to each writing sample.

The AWC dimensions include: Content (addressing both quality and clarity of presented ideas); Structure (addressing the overall flow and organization of an essay); Stance (addressing the appropriateness of writing for the task and audience); Sentence Fluency (addressing the structure and flow of individual sentences); Diction (addressing the appropriate use of words and expressions); and Conventions (addressing the appropriate usage of punctuation, capitalization, spelling, and vocabulary). Each attribute is assigned a score between 1 and 6 (the higher the score, the better the writing; Swain & LeMahieu, 2012). See Appendix B for an example of an analytic scoring rubric.

**Prominent Feature Analysis**

Prominent Feature Analysis resembles an analytic scheme due to the multiple elements considered when appraising a writing sample. However, it is much more detailed than a typical analytic scoring rubric. If holistic scoring were a postcard shot of the Statue of Liberty on Ellis Island, then analytic scoring would be the close-ups of her head, torch, and base, and Prominent Feature Analysis would be a section drawing

exposing the specific structural elements that keep her upright and glorious. The tool is uniquely authentic, as it is derived from the specific writing samples it is used to assess. This is both its advantage and disadvantage. On one hand, it assures that a given sample's writing characteristics are fully noted and understood; on the other hand, it reduces the scale's generalizability. However, the aim of Prominent Feature Analysis is not to broadly generalize, but to fully and thoroughly understand the characteristics of a given writing sample. Assuming a large-enough participant group, both excellent and severely lacking writing examples will be present, thus allowing for extraction of specific characteristics collectively representing what a given group of writers can and cannot do well. This in-depth understanding of characteristic of a given sample can foster application of specific interventions aiming to increase the rate of occurrence of positive characteristics, and to decrease the rate of occurrence of negative features.

While the product of Prominent Feature Analysis (a list of sample-specific writing characteristics) may not apply to a writing sample penned by another group of students, the process of generating such a list—an analysis of sufficiently-large writing sample by trained experts—generalizes well. At the same time, the list of features created from one writing sample may serve as a relevant and appropriate starting point for an analysis of another, comparable writing sample (Morse, Swain, & Graves, 2007). See Appendix C for a Prominent Feature Analysis example.

In one study of seventh grade writing, 32 prominent features were identified, 22 positive and 10 negative. Positive features examples included: transition words, sentence variety, metaphor, effective repetition, effective organization, and coherence/cohesion.

Negative features examples included: redundancy, usage problems, faulty spelling, weak structural core, and garbles (Swain, Graves, & Morse, 2010).

## Statement of the Problem

Prominent Feature Analysis is a writing assessment tool capable of elucidating student writing characteristics to an unusually detailed degree. It is powerful, because it is derived from the very writing it is used to assess, instead of exemplifying a compilation of goals reflecting an ideal criterion. This degree of relevant detail affords an opportunity for uncommonly systematic understanding of writing characteristics of a given sample. Behavioral scientific writing is a complex genre. Writers are expected to present information in a particular format, and in an expected order. Scientific manuscripts are organized into specific sections, with specific headings. For example, an empirical study report will be usually divided into a review of past literature on the topic of interest ending with an identification of some unknown, a description of a method to investigate the identified unknown, a summary of the results, and a discussion on the results' meaning.

In scientific writing, strict and consistent measures are taken to properly credit ideas and words of others. For example, a writer is expected to attribute all direct quotes and paraphrased ideas of others with the authors' last name, and the year in which the source paper was published. Attributions of direct quotes also contain page numbers indicating a specific location where the quote appears in the original paper. Subsequently, more specific information on the source will be found in the references section at the back of the manuscript; each reference will be included in a specific and consistent format.

The language used by scientists contains many uncommon words or common words used in new, specific ways. For example, the word "significant" means "important" in common language, but in science, it denotes an occurrence at a rate different than chance.

Given the complexity of the genre, novice writers and writing instructors would greatly benefit from systematic understanding of the novice scientific writing characteristic. It would be helpful to quantify and understand which of the genre's requirements are met by the novices with ease, and which give them trouble. However, it is not clear whether Prominent Feature Analysis would prove useful for assessing novice scientific writing, or for tracking growth as scientific writers.

## Justification of the Study

Because of its attention to grammatical, structural, and stylistic aspects of writing, Prominent Feature Analysis provides an unusually precise method for understanding of a writing sample. The tool has demonstrated evidence for construct and criterion validity, and inter-rater reliability, for on-demand expository writing of seventh-graders; its usefulness and psychometric properties have been subsequently confirmed with elementary and high school students (Morse, Swain, & Graves, unpublished; Swain, Graves, Morse, & Patterson, 2012).

Despite many years of writing instruction, post-secondary students greatly struggle with writing. This is demonstrated by the existence of writing centers within universities, and university-sponsored training programs for faculty from all domains to improve their students' written communication. In addition to basic writing, many fields require mastering area-specific scientific writing skills. Each branch of sciences has its

own set of rules; researchers in the formal sciences (i.e., mathematics and logic) write differently than those in natural sciences (i.e., physics or biology) and social sciences (i.e., sociology or psychology).

The present study focused on scientific writing used in behavioral sciences (a sub-category of social sciences). For novices, the complex content, often-rigid structure, and domain-specific jargon of the genre adds an additional layer of difficulty and cognitive strain to an already-challenging task of committing one's ideas to paper. It therefore appears that elucidating a pattern of positive and negative writing features of novice behavioral scientific writing has merit, and may have a potential to inform and impact writing instruction and improvement—on individual and class-wide level.

## Purpose of the Study

The purpose of this study was to extend the Prominent Feature Analysis scale into a new genre (behavioral scientific writing) and academic level (undergraduate-age adults), and to explore the new scale's psychometric properties. The present research consisted of identifying genre-specific prominent features in a representative sample of novice behavioral scientific writing, and exploring the relationships between the features. To assure the rigor of the writing sample analysis process, I collaborated with two experts in linguistic aspects of writing, and one expert well-versed in behavioral scientific writing requirements.

I investigated the validity of the scale by correlating the prominent feature scores with student college GPA and ACT scores (composite as well as language, math, reading, and science sub-scores). Additionally, to check the scale's sensitivity to changes in

writing over time, I compared the prominent feature scores between two small writing samples penned by the same students during two consecutive semesters.

## Theoretical Background

All major proposed models of the writing process include revisions as one of their components (e.g., Bereiter & Scardamalia, 1987; Galbraith, 1999; Hayes, 1996). Revisions are commonly informed by feedback. In the broadest of strokes, feedback is most helpful (in terms of its potential for improving subsequent writing) when it is copious, timely, legible, and specific (e.g., Agius & Wilkinson, 2014; Sommers, 2006).

The ability to significantly improve writing through rewriting and revising differentiates novices from experts; experts revise their drafts to a greater extent, in both breath and depth (Galbraith & Torrance, 2004). The only way to acquire expertise in writing and revision is through writing and revising, ideally, with help from a more-knowledgeable other. In the context of academic writing, revisions are commonly required as a part of the assignment, and are fueled by instructor (and/or peer) feedback; the more insightful and voluminous the comments, the better off the writer. Additionally, these comments are often the only writing instruction a student gets in a non-writing focused course (Lyon, 2016).

Prominent Feature Analysis for novice behavioral scientific writing has potential to provide an instructor and, subsequently, the writer with a clear, specific, and extensive set of directions for improvement. It would do so because of its detailed structure and attention to linguistic, structural, and genre-specific components. Additionally, the assessment results can inspire brief, class-wide writing instruction exercises (in any

course type), which have been shown to result in significant gains in writing quality, despite very short duration (Lucas, 2010; Lyon, 2016).

## Research Questions

The present study aims to extend an existing writing assessment scheme to a new genre (novice behavioral science), and new demographic segment (undergraduate students), and to investigate its validity and reliability. The study is guided by the following four questions:

1. What writing characteristics comprise the Prominent Feature Analysis scale for novice behavioral scientific writing?
2. What are the relationships among the identified features?
3. Do students' prominent features scores relate to their college GPA, or ACT scores, including composite score, and language, math, science, and reading sub-scores?
4. Can the Prominent Feature Analysis scale be used to assess change in student writing when two samples from the same students are compared across time?

## Hypothesized Outcomes

Regarding Research Question 1 (RQ1), I hypothesized that the scale will keep many of its original linguistic features, identified in writing of seventh-grade students (Morse et al., 2007), as they pertain to writing in general. One was irrelevant ("illegible handwriting"); the frequency of some would be much lower than in previous applications, but still present ("voice" or "metaphors"). Additional, genre-specific

features would be added (“references errors”; “analysis rigor”). Due to strict genre characteristics, the new additions would likely be negative features, for example, not meeting an American Psychological Association (APA) manuscript format requirement. However, uncommon levels of scientific thinking for a novice, identified as positive features, may appear also. For example, I anticipated identifying features demonstrating student struggles with in-text source attributions and references, as well as difficulties with professional jargon use.

Regarding Research Question 2 (RQ2), I suspected that struggles with writing in general would correlate with the struggles in scientific aspects of writing. Therefore, I expected to see direct relationships between positive linguistic and scientific features, and negative linguistic and scientific features.

Regarding Research Question 3 (RQ3), I expected to see positive correlations between the Prominent Feature Analysis composite score and students’ ACT composite scores; I was unsure whether any of the ACT sub-scores would correlate with any of the prominent feature sub-scores, based on lack of previous literature. Additionally, I expected to see positive correlations between students’ college GPA and the Prominent Feature Analysis composite score. I supposed this to be the case based on a Vygotskian belief that good thinking, required for success in academia, correlates with linguistic excellence (Vygotsky, 1986). While some classes may not emphasize writing, being an overall good student (as indicated by one’s GPA) suggests some combination of an intellectual potential, good writing skills, and good study habits.

Regarding Research Question 4 (RQ4), I was unsure whether I would see differences in prominent features present in two samples of writing generated by one

student: a shorter sample (about three pages) and a longer sample (about 10-12 pages). Correlation between the two sets of scores would indicate that student writing characteristics are stable between writing samples (meaning not prone to task effect). Additionally, I was curious whether, if present, the differences between the scores may be able to elucidate student growth as writers. Two scenarios were possible. First, both writing samples could exhibit the same positive and negative prominent features. Second, the samples could exhibit different positive and negative features, based on the level of difficulty of the prompt for the writer, increased writing expertise between the two samples, or chance circumstances like demanding class schedule of the writer during a given semester. Growth as writers could be suspected if the first set of positive scores (signifying desirable writing characteristics) is significantly lower than the second set of scores, and/or if the first set of negative scores (signifying undesirable writing characteristics) is significantly higher than the second set of scores.

## Summary

Writing assessment is used for both formative and summative purposes. While holistic and trait/analytic rating schemes have advantages in some instances, they are not ideal for providing extensive and specific feedback, so helpful in a classroom setting for both instructors and novice writers. Scientific writing is a peculiar writing genre, which —when done well—combines mastery of expository writing, demonstration of domain knowledge, excellence in communication to diverse audiences, and adherence to strict formatting and stylistic rules. Because of the required cognitive load, it is particularly hard for novices. Students need practice in revising and rewriting to master scientific writing. Extensive and specific feedback facilitates the process.

Prominent Feature Analysis is a unique writing assessment scheme, derived from the very writing it is used to assess. In a large-enough sample, one is bound to find examples of utmost writing excellence, as well as failed attempts at conveying meaning through written language. This range, reflected in the range of specific positive and negative features, assures that the Prominent Feature Analysis accurately portrays the abilities of writers whose writing it scales.

Prominent Feature Analysis has been demonstrated to be a reliable and valid tool for assessing elementary and high school-level writing. Extending Prominent Feature Analysis to novice scientific writing will allow for a methodical identification of present (and lacking) characteristics of the genre, and will be immediately useful for informing instruction, student revision, and rewriting processes.

# CHAPTER II

# LITERATURE REVIEW

In 1912 and 1913, Daniel Starch, a renowned educational researcher and administrator, and Edward C. Elliott, a renowned psychologist, published an interesting trio of studies on writing assessment (Starch & Elliott, 1912, 1913a, 1913b). The authors sent out four papers, two sample students' high school examination essays in English, one in mathematics, and one in history, to about 200 high schools each, with a request that the main teacher of the given subject review and grade the work. The resulting responses, in all three subjects, varied significantly. On a 0-100 scale, English essay grades spanned from 50 to 98; history essay grades ranged from 43 to 92; and geometry proof paper grades ranged from 25 to 92. The raters clearly took great care in reviewing the work, and explained their grading decisions at length. Some deducted points based on hand-writing legibility, spelling errors, or the paper's overall aesthetic in addition to content, others based their grading on content only. Despite clear and detailed explanations of the review logistics by the raters, the specific grades in Elliott and Starch studies appeared to be assigned nearly at random. Despite over a century passing, the worry regarding the accuracy and consistency of writing assessment remains.

## Writing Assessment Background

Assessing writing is a tricky task, even for the most experienced raters. Numerous factors affect how a writing sample is judged: what criteria are used, the rater's

experience in using the scheme, the rater's level of expertise in writing assessment, difficulty of the assignment for the writer, perceptions of the topic/content by the rater, and more (i.e., Engelhard, 1992; Huot, 1990).

Depending on context, different ways to evaluate a writing sample may be appropriate. Writing can be evaluated holistically (e.g., this is exemplary work; an "A" paper), or different aspects of writing may be evaluated separately (excellent ideas; poor spelling and punctuation). Additionally, the reviewer may or may not generate more detailed feedback to the writer. When discussing writing assessment, it is important to understand the overall purpose of writing assessment, the importance of feedback during the writing and assessment process, and the advantages and disadvantages of different assessment types.

**Overall Purpose of Writing Assessment**

Writing assessment is used for a variety of purposes within an educational context. Broadly speaking, writing assessment in educational settings falls into three categories: administrative, instructional, and evaluation and research (Cooper & Lee, 1977).

**Administrative uses.** The administrative realm focuses on achievement rating, and includes assessing writing for assigning course grades, and summative assessment in form of standardized high-stakes tests (Cooper & Lee, 1977). These high-stakes tests' results can be used for course placement (or course exemption), for example, a student who earns a sufficiently high score on an Academic Placement (AP) test in a course in high school may be exempt from taking that class in college. Some high-stakes tests with

writing components, like American College Testing (ACT) and GRE, are used, in conjunction with other materials, for admissions into institutions of higher learning at the undergraduate or graduate level, respectively.

**Instructional uses.** The instructional realm involves identifying student writing difficulties, tracking student or class progress, and adjusting instruction and feedback as needed to maximize learning (Cooper & Lee, 1977). These formative assessments can take many forms, from "typical" instructor feedback on course assignments and tests, peer-review of course papers, or small, focused assignments like one-minute papers (Lucas, 2010).

**Evaluation and research.** Writing assessment may be used for measuring students' growth in a course. Conversely, it could be used for evaluating effectiveness of a writing program or a writing instructor. Lastly, writing assessment results can be correlated with other measures of student achievement to understand student learning (Cooper & Lee, 1977).

## Summary

Writing assessment is one of the primary modes of assessing student learning. It takes many forms, and is used for different purposes, including administrative, instructional, and evaluation and research-related. As some of the applications of writing assessment are high-stakes, it should be rigorous, thorough, reliable and valid. Depending on context, feedback to the author may be desired, or required.

## Importance of Feedback

Feedback is essential for students to evaluate their progress while completing an assignment and to alter the product as needed to best meet the assignment's requirements (Nicol & Macfarlane-Dick, 2006). Every major model of the writing process includes revisions as one of its elements (e.g., Bereiter & Scardamalia, 1987; Galbraith, 1999; Hayes, 1996). While some revisions are self-generated by the writer, in an educational setting, revisions are typically based on feedback from instructors, peers, or other reviewers. A recent research synthesis on undergraduate students' and teachers' views on written feedback revealed four themes: quality of feedback; quantity and location of feedback; feed-forward; and timeliness (Agius & Wilkinson, 2014). Each of these is discussed below.

### Quality of Feedback

Students have clear preferences regarding what constitutes helpful feedback. They claim to benefit from clear, focused, critical comments, and specific explanations of mistakes they make (Agius & Wilkinson, 2014). Desire for feedback specificity has been a constant in the literature on student writing. It appears in studies on high-school students (Bardine, 1999; Bardine, Bardine, & Deegan, 2000), as well as in studies on post-secondary level students. A study of 400 Harvard students' experiences and perceptions of college writing over their undergraduate careers ("six hundred pounds of student writing, five hundred hours of recorded interviews and countless megabytes of survey data"; Sommers, 2006, p. 249) confirmed the sentiment. When asked what suggestions they have for faculty to improve student writing outcomes, 90% declared the specificity of feedback to be of utmost importance, above all else. Ample and specific

feedback was, in students' eyes, the key factor responsible for engaging with the faculty member, and for facilitating their ability to both think about their content, and express their thoughts with increasing precision through writing.

Students consider ambiguous praise superfluous, but appreciate specific notes regarding things they do well. They also feel frustrated and demotivated by negative-only feedback. Additionally, students get frustrated by the discrepancy between a low grade and a lack of explanation justifying it. Lastly, students strongly feel they need to be able to understand the feedback, both in terms of legibility, and clarity of content (Bardine et al., 2000). Illegible handwriting and jargon-ridden comments obscure the meaning and may prevent students from benefitting from instructor feedback (Robinson, Pope, & Holyoak, 2013). While instructors claim to understand the need for positive feedback, they do not always provide it, and often focus on justifying the grade by highlighting only the work's shortcomings (Agius & Wilkinson, 2014; Sommers, 2006).

Psychological perspective supports the importance of specificity and encouragement. Specific comments that give directions for improvement, if worded correctly, may promote intrinsic motivation and growth mindset among novice writers (Willingham, 1990). Intrinsic motivation is a desire to engage with an activity for "love of the game," and correlates with multiple academic benefits. Comments implying instructor's genuine engagement with the ideas presented by the student may offer a boost of confidence to a novice writer, and a desire to continue the "discussion" through reworking of the writing to address the instructor's feedback. Growth mindset is cultivated by carefully-worded comments implying instructor's belief that the writer is capable of improving the manuscript by revising it further. Such feedback fosters the

belief that effort, not some unchangeable/inborn trait, is key to growing as a writer. Willingham (1990) proposed two specific strategies for successful feedback. First, he suggested offering feedback on writing in form of a brief summary of the current paper's main ideas, and letting the writer decide whether that is indeed what he or she intended to convey. Alternately, he suggested offering feedback in form of leading questions. These constitute an invitation to a dialogue, and are designed to keep the students engaged with critical assessment of the content, instead of mindlessly picking up edits.

Additionally, he emphasized the importance of a transparent and overt hierarchical structure to writing feedback. Specific comments on ideas and overall structure should clearly be most important, and comments on mechanics (spelling, grammar) should be secondary, though still present. This helps students understand that thinking is prized most, while mechanics are less important (though still need addressing).

**Quantity and Location of Feedback**

In general, according to literature, the more feedback the better; the "deeper" the feedback, the better. The "deep" feedback involves clear cues for the students, which help them to understand the expected standards, current deviation from the expected standards, and suggestions for how to bridge the gap between the present effort, and the excellence in writing (Pokorny & Pickford, 2010). Student-preferred comments are detailed and include correct examples or unambiguous directions for improvement rather than just highlight grammatical, stylistic, or content shortcomings (Agius & Wilkinson, 2014). Additionally, students claim to benefit more from comments in margins, located near the place to which the comment pertains, over feedback grouped on one "comments"

page. In a focus-group-based study of business school student perceptions of feedback on writing, students explained their preferences for feedback location. The participants stated that feedback near the location to which it pertains leaves no doubt as to what needs attention (Pokorny & Pickford, 2010).

Interestingly, students value voluminous feedback regardless of the grade received on the assignment. A sample of 166 first-year undergraduate psychology students responded to a survey investigating their perception of value of feedback on their writing completed as homework assignments. Results of a 2x2 analysis of variance (low/high grade; sparse/extensive feedback) revealed a significant effect of quantity of feedback on students rating of the comments as helpful, regardless of the grade received on the assignment (Robinson et al., 2013).

A puzzling discrepancy was noted regarding student and instructor views on the relationship between feedback and grades. While students claim they appreciate and benefit from voluminous comments on their work, some instructors believe that students are mostly interested in grades, and disregard the accompanying feedback. Others go as far as to say receiving grades *prevents* students from engaging with and addressing instructor feedback (Agius & Wilkinson, 2014). However, this discrepancy may be partially explained by the timing of instructor's feedback. Having no chance to revise an assignment, or to complete another one in a class, students may indeed be more interested in the grade than comments on their writing. Such seemed to be the case in a study of writing assessment behavior of 16 seasoned faculty in teacher education and nursing education programs. Use of both think-aloud protocol and analysis of marked student

work allowed the researcher to note this mismatch of expectation and feedback timing on part of faculty (Orrell, 2006).

**"Feed-Forward"**

Feedback has potential for guiding improvement on future assignments. Feed-forward describes feedback that specifically aims at improving future drafts or revisions, or performance on subsequent assignments (Agius & Wilkinson, 2014). Motivated students view feed-forward as a mechanism to improve short-term outcomes (i.e., grades on an assignment) as well as long-term outcomes (i.e., professional skills after graduation). Instructors tend to view feed-forward as needed only in case of weak performance, students however claim to value and expect advice for future improvement as a part of all their feedback.

Additionally, students claim they are more likely to note and heed feed-forward suggestions phrased as options rather than orders. In a survey study of 142 college freshmen perceptions of instructor's comments on a student essay, corrections, criticisms, and commands were viewed as most controlling of students writing, rather than facilitating growth and fostering a dialogue between a student and an instructor. Corrections included physical changes to student's text, criticisms consisted of negative evaluations without additional directions for change, and commands included direct and specific demands for change rather than invitation to rethinking one's writing (or thinking) choices (Straub, 1997).

### Timeliness of Feedback

Both instructors and students acknowledge the importance of timely feedback on student work. However, differences were noted between instructors and students regarding the importance of timely feedback on different types of assignments. In terms of formative feedback, including feedback on assignment drafts, or assignments followed by other assignments within a span of a semester, both parties recognize the need for immediacy (e.g., Bevan, Badge, Cann, Willmott, & Scott, 2008). In terms of summative feedback, or feedback on end-of-semester assignments with no chance for revisions or transfer of skills to subsequent class projects, teachers felt less strongly than students that timeliness was important (Agius & Wilkinson, 2014).

### Summary

Written feedback is an important tool for enhancing student writing quality and promoting learning. Overall, both instructors and students agree that the more feedback, the better; the more specific the feedback, the better. In general, students favor timely, clear, focused, and specific advice for improvement, legibly written in margins, near the location it addresses. They appreciate specific positive feedback in addition to mistake corrections. Instructors recognize the power of specific, formative feedback, though do not consistently perceive students using the feedback to improve their subsequent work. Some faculty may be less inclined to offer extensive and timely comments on summative assignments. Feedback focusing on ideas rather than mechanics of writing may foster most growth in novice writers, and may offer additional benefits in terms of intrinsic motivation and increases of growth mindset.

## Writing Assessment Types

Numerous advances took place in the field of writing assessment since Starch and Elliott's (1912, 1913a, 1913b) dramatic demonstration of poor interrater agreement in domains of English, history, and mathematics. In the 1960s, the precursors of two main rating schemes, holistic and analytic scoring, were proposed; about ten years later, primary trait scoring method gained momentum, then lost its popularity. In most general terms, holistic scoring involves assessing a writing sample in its entirety, along a few guiding parameters (i.e., organization, quality of analysis). Analytic scoring involves assessing a writing sample along a set of predetermined sub-skills (i.e., mechanics, ideas), and assigning a separate score to each (Huot & O'Neill, 2009). Primary trait analysis involves assessment of sub-domains, and focuses on categories specific to the writing tasks. For example, assessing expository writing (writing that describes a product, or explains a process), may involve assigning sub-scores for clarity and depth of understanding. These categories may be different in analytical writing (Fredriksen & Collins, 1989). Due to its narrow, task-specific focus, primary trait analysis is not broadly used, and will not be discussed further in this document. Prominent Feature Analysis resembles an analytic scheme, as it assesses a writing sample along multiple dimensions, but it is much more detailed, and therefore informative, than other common analytic measures (Swain et al., 2010). Detailed descriptions of the rating scheme types follow.

### Holistic Assessment

ETS pioneered the creation of the early "General Impression Marking"/holistic scoring schemes, and heavily researched them. ETS' primary goal was to design a reliable, valid, and affordable way to conduct large-scale, standardized assessments that

included generating writing samples in addition to completing multiple choice questions. It is important to differentiate "first impression" scoring from the contemporary holistic scoring. The former involves assigning scores quickly, without much guidance, and relying on rater's expertise to differentiate between "A" and "B" papers. The latter involves assigning a single score to a piece of writing, based on a precise prompt, while keeping in mind a few predetermined and clearly defined key criteria, like the rigor of analysis, or essay organization. (Charney, 1984; White, 2009). For example, the evaluated dimensions in the GRE advanced writing subtest include: idea analysis, development of persuasive arguments and examples, focus and organization, usage of correct sentence structure, sentence variety and precise vocabulary (ETS, 2018).

To increase the inter-rater reliability of holistic scores, the following six practices were proposed to standardize the grading process. These include: a) standardizing essay reading (all scorers are in one place, at one time, and follow the same schedule of rating and breaks); b) developing a scoring rubric (the rubric initially reflects the goals and expectations of the test designers, but is adjusted based on the qualities of the sample); c) extracting anchor papers (these exemplify a given score point, for example, one for a 3.0 and one for a 4.0 paper); d) regular checks during the rating process (performed by "table leaders" whose job is to assure consistency of the raters grouped at a given table, typically of 6 to 7 raters); e) using multiple readers for each paper (typically, two independent readers blinded to the other's rating score each paper); and f) session evaluation (to verify that raters remained consistent throughout the session; White, 2009; Yancey, 1999). It is important to note, that scores are considered equal if they are no more than one scale unit apart. The scorers are considered to be in agreement when one

rater assigns a 3.0 rating to a paper while another one assigns a score of 4.0; the paper's final rating will be 3.5. When the difference is larger than one point, a third rater reviews the writing sample (White, 2009).

Assuming participation of trained, expert reviewers, the strengths of holistic assessment include reliability, speed, and low cost of evaluation per paper (Huot, 1990). Alas, concerns abound. The commonly cited concerns with holistic writing include: a) limited information about the quality of each writing sample; b) limited utility to inform instruction; c) questionable reliability; and d) unclear decision-making process by raters. Some of these concerns may be rooted in the origins of holistic scoring; it was initially created by the measurement community, not by writing teachers (Charney, 1984; Huot, 1990; White, 2009).

First and foremost, holistic scoring allows the rater to broadly rank the papers only; the scoring does not inherently dictate a passing/failing cutoff. Each time papers are graded holistically (the term coined by Fred Godshalk of ETS), the sample needs to be evaluated in terms of what score is deemed "good enough." Additionally, as the point category descriptions are succinct, holistic scores do not provide adequate information for an instructor to facilitate improvement, or elaborate feedback to the student that may catalyze change on the next assignment (Huot, 2003; White, 2009).

Despite all the precautions, a holistic rating process can result in reliability problems. In a study of 699 California State University English Equivalency Examination papers (each containing two essays), two raters assessed each entry on a six-point scale, generating scores from 0-24. The papers were rescored a year later; scores matched perfectly with the previous ones for only 20.7% of participants. Forty-two percent of

papers were assigned scores that were more than two points off from the scores assigned the previous year by the same scoring team (White, 2009). While two points may not sound as much, any discrepancy resulting in a lower score for a given student may mean the difference between passing and failing the examination.

Lastly, research suggests that holistic scores may correlate with numerous features not related to the rubric. These include: length of the essay; uncommon or mature vocabulary; spelling errors; paper's appearance or organization (for example, neat handwriting, or obvious five-paragraph structure); presence of final free modifiers (modifiers placed after the main clause, for example: students worked on their paintings, *their cheeks flushed with excitement*); or the content of the essay (Charney, 1984; Huot, 1990). Interestingly, it has also been suggested that holistic scoring may reflect the readers' exclusive attention to the specific features listed in the rubric (representing the "looking for Waldo" cognitive processing of the written work), and disregard for whatever other qualities may be present in the paper (Stock & Robinson, 1987).

**Analytic Assessment**

Analytic scoring involves defining a series of sub-domains or performance elements for an assignment, and evaluating each separately within a paper. For example, the AWC is an analytic rating scheme developed by the National Writing Project for use in K-12 classrooms for writing assessment and research (Swain & LeMahieu, 2012). The National Writing Project is a network of faculty development centers across America, focused on supporting educators in improving writing in students of all ages (Swain & LeMahieu, 2012). The AWC scale is based on the Six + 1 Trait Writing Model (Culham, 2003). In the Six + 1 model six distinct sub-skills are evaluated and assigned separate

sub-scores, and an additional score ("+1") represents the sum of these sub-scores. The six AWC sub-domains include: Content (quality and clarity of presented ideas); Structure (the overall flow and organization); Stance (appropriateness of writing for the task and audience); Sentence Fluency (the structure and flow of individual sentences); Diction (the appropriate use of words and expressions); and Conventions (the correct use of punctuation, capitalization, spelling, and vocabulary). Each sub-domain is assigned a score between 1 and 6 (the higher the score, the better the writing; Swain & LeMahieu, 2012). AWC's holistic score represents an independent holistic evaluation, *not* a sum of the other 6 sub-scores.

Unlike holistic evaluation, analytic scoring provides information regarding individual's or class' specific writing strengths or weaknesses, and thus includes possible direction for improvement. Additionally, the ability to differentiate student ability within each sub-domain may increase validity of the scores (Bang, 2013). For example, each of the six AWC subskills is scored on a six-point scale, thus allowing for a finer differentiation than a sample holistic scheme spanning 1 to 6.

White (1994) identified three potential drawbacks to analytic assessment: a) the writing community disagrees regarding what the key sub-domains of writing are; b) increased scale scope (in comparison with holistic assessment) reduces scoring accuracy; as more judgments are made, there is a greater possibility of disagreement between raters; and c) due to the rating scheme's complexity, training the raters and using the scale is time-consuming, therefore not economic for large testing events.

**Prominent Feature Analysis**

Prominent Feature Analysis is a relatively new rating instrument; the first study using the scheme was presented at a conference in 2007 (Morse et al., 2007). It is somewhat like an analytic rating scheme, as it independently evaluates multiple elements within a writing sample. Unlike the analytic scheme, the sub-domains are evaluated in a binary fashion (either present or absent) rather than on a continuum. However, an expert eye is needed to distinguish prominence in a given writing type, penned by particular group of writers. For example, a single metaphor noted in an expository essay of a seventh-grader would rise to prominence; a few minor spelling errors would not. Also, unlike the traditional analytic schemes, which can be used on many samples and often on many types of writing, Prominent Feature Analysis is genre and writing sample-specific. As any seasoned writing instructor or assessment expert would attest, given a large-enough set of writing samples, both excellent and poor examples of writing within a given genre will be present. The requirements of the statistical tool best suited for making sense of prominent feature data (Rasch analysis) set the minimum number of participants to 200, which amply assures meeting the somewhat vague "large-enough sample" criterion. This number of independent writing samples allows for extracting numerous features that a given group of writers can generate. While some characteristics are universal (i.e., poor punctuation, or voice), others will be very specific to the writing type.

When creating the scale, the raters review the writing samples looking for characteristics that stand out, either positively or negatively. As they note the characteristics that stand out, the raters compile a list of sample-specific prominent

features. As mentioned above, this process hinges on the rater's ability to differentiate between ordinary and extraordinary characteristics of a given writing type, therefore extensive training and/or background in the relevant genre is a prerequisite for serving as a scorer. It is important to highlight that prominent feature scores are *derived* from student writing, while holistic and analytic scores are *assigned* using a priori-generated criteria (Swain et al., 2012). To date, all Prominent Feature Analysis studies focused on expository, on-demand writing, generated during state-mandated testing or during the National Writing Project-led interventions to improve the teaching of writing; participants included students in elementary, middle, and high school.

Prominent Feature Analysis raters did not set out to find examples of an a priori-generated list of characteristics based on their expectations; rather, they read the essays to see what stands out, what is prominent in each essay. However, they did not search blindly. Based on a long-standing theoretical knowledge of writing, and their professional experiences, the raters sought a few particular characteristics, for example: "cumulative sentences containing final free modifiers, voice, flawed sentences, and certain intersentential connections" (Swain, Graves, & Morse, unpublished, p. 8). The importance of final free modifiers, or modifiers that follow the main clause, has been highlighted by Christensen (1963) and other linguists as a constant feature in high-quality written communication. Voice, or "the presence of an original, personal or authentic conception of the subject" (Morse et al., 2007, p. 14) has been researched extensively for the last four decades. While the definition may be vague, voice is readily noticeable in high-quality writing (Swain, Graves, & Morse, 2015), especially by writing assessment experts. Lastly, a family of flawed sentences has been studied and described by Krishna

(1975) as sentences with a "weak structural core" (p. 45). These are typically comprehensible, but grammatically incorrect, and require a more complicated revision than, for example, fixing a subject-verb agreement flaw. The following sentence has a weak structure core: "By paying directly, it is assured we get better service" (Krishna, 1975, p. 48).

In a Prominent Feature Analysis of seventh-grade expository writing sample generated following state-mandated testing in 2004, 32 prominent features were identified, 22 positive and 10 negative. While the topic was not included with the writing samples, Prominent Feature Analysis authors inferred the students were asked to write about activities they enjoy doing outside of school. It also appears they were asked to write at least three paragraphs, and to plan their writing.

Positive feature examples derived from this sample of essays included: transition words, sentence variety, metaphor, effective repetition, effective organization, and coherence/cohesion. Negative features examples included: redundancy, usage problems, faulty spelling, weak structural core, and garbles (Swain et al., 2010; see Appendix C for the full list of positive and negative features identified in this study).

In addition to seventh grade writing, to date Prominent Feature Analysis has been used to assess the writing of students in Grades 3-5 (Morse et al., 2007), and of students in Grades 9-10 (Morse et al., unpublished). The scale grew to 40 features for student writing in Grades 3-5, signifying that the scale is subject to change with the prompt, and/or particular participant sample. Twenty-six features remained constant between the seventh grade and Grades 3-5 grade studies, and, importantly, largely kept their estimated Rasch difficulty levels, meaning features that were occurring often (i.e., faulty

punctuation), and seldom (i.e., transition words) remained "easy" and "hard," respectively, between the samples (Morse et al., 2007). This suggests that the scale consistently reflects the underlying construct, quality of writing.

**Theoretical grounding of Prominent Feature Analysis.** Both holistic and analytic schemes share a following characteristic: once created (and until revised), they become static measuring sticks against which many participant samples can be measured. This requires, on one hand, a general conception of key criteria (for holistic scoring) and of sub-domains (for analytic scoring), and, on the other hand, a predefined list of writing prompts which allow the writers to demonstrate their abilities along the expected characteristics. These scoring schemes inherently imply that student writing ability is fixed and independent of the circumstances in which the writing is generated, and that the raters (if properly trained) are objective, constant in their appraisals, and interchangeable (Huot, 1996). This set-up may be adequate for summative evaluations, including standardized, state-mandated testing in K-12 schools, or high-stakes university admissions tests like ACT or GRE. However, holistic and analytic rating schemes are less helpful in formative testing or as vehicles for feedback.

Noting the shortcomings of these two scoring schemes, Huot (1996) identified a need for a different assessment type. He advocated creating a measurement process rather than a measurement tool, that would conceptualize writing not as a one-time showcase of skill, but as a "communication event" (p. 559), an attempt of a student to convey information/meaning, in a particular context, to a specific reader/audience. He also strongly advocated for assessment to be a communal activity, in which the educators collaborate to understand, interpret, and assess student communication efforts. Such

collaboration should involve generating a measure appropriate for the task at hand, based on the writing samples to be reviewed, and not on abstract, external standards. Huot proposed five "principles for a new theory and practice of writing assessment" (p. 562); these include assertions that assessment should be: a) site-based; b) locally-controlled; c) context-sensitive; d) rhetorically-based; and e) accessible.

Site-based refers to the fact that an assessment should be dictated by the need of a particular site, for example, an institution, agency, or a department. Local control pertains to the need for the specific site institution (like a department or perhaps even a single classroom or course) to define, manage, and update the relevant procedures as needed. Context sensitivity refers to the need of the assessment to reflect the instructional goals and objectives, along with the idiosyncratic reality of a given institution or department. For example, scientific writing required of educational psychology students is very different than literary critique or creative writing required of students in an English department. Rhetorical base is required to assure that the prompts, scoring requirements, and the review process follow best practices in using language for effective, persuasive, and thoughtful expression. Lastly, by "accessible," Huot means that the entire process, including prompt creation, assessment criteria, review protocol, and samples of work with judgement explanations, should be transparent, and open for review by individuals whose writing is being assessed. While this could apply to large testing efforts, it may not be relevant in a context of a single university course.

A noteworthy and unique aspect of Huot's (1996) proposal is the fact, that it renders the traditional interrater reliability, the "sacred cow of writing assessment," (p. 563) irrelevant. Huot suggests a radical change of procedure, from raters working

individually and then comparing scores, to well-trained raters collaborating as needed as they score, with a goal of building common, deep understanding of the assessed writing samples, and achieving 100% agreement regarding the observed characteristics.

Prominent Feature Analysis is a response to Huot's (1996) call. The rating scheme, each time it has been applied so far, has been created or adjusted to respond to a particular context, used by local experienced teachers, invested in the process, as it was their collective students who generated the writing samples. The raters, familiar with the sociocultural reality of the writers and the schools, collaborated on creating the scale, and collaborated on rating the essays as a group. As all Prominent Feature Analysis applications so far have been published or presented at relevant conferences, the process has been disseminated in the spirit of transparency and Huot's accessibility.

**Validity of Prominent Feature Analysis.** In the study of seventh grade writing generated as a part of state-mandated assessment ($N = 464$, from three schools, from two school districts in Mississippi), 32 prominent features were identified (22 positive and 10 negative). The holistic scores assigned to the writing samples by the district raters significantly positively correlated with the summed feature score ($r = .54$), positive features ($r = .48$), and significantly negatively correlated with negative features ($r = -.48$). In the study of 551 students in grades 3-5 from two schools from two districts in Mississippi, 40 prominent features were identified. These correlated significantly with National Writing Project-assigned holistic scores ($r = .58$). These correlations clearly provide support for criterion-related validity of Prominent Feature Analysis (Morse et al., 2007).

The features are derived from the writing samples through an open-ended process. Subsequently, that scale is then applied, in a communal setting, by experts in writing instruction and assessment, while continuously seeking consensus on all papers (Swain et al., 2010). This process supports the scale's high content validity.

**Reliability of Prominent Feature Analysis.** Prominent Feature Analysis creators proposed several techniques to ensure consistency of the rating process; many of these echo White's (2009) suggestions for increasing the reliability of holistic scoring. Swain et al. (2010) suggested that evaluators gather in one place at one time to conduct the scoring; others have noted the benefits of teacher teams collaborating on assessment (e.g., LeMahieu & Friedrich, 2007). Additional recommendations include rating papers blind to authorship, and selecting a group of essays to serve as anchor papers/training materials. The raters should together decide what prominent features appear in these training essays, versus what features constitute ordinary writing. Upon completion of the training phase, the raters should work independently, albeit side by side, and double-read the papers to assure consistency. Lastly, the researchers suggest reflection on the created feature list, and a discussion regarding its content (Swain et al.).

Following the above recommendations, to provide data-based support for reliability of the scheme, Swain and colleagues (2010) investigated classification consistency of identifying a feature as present or absent in the seventh-grade sample—generated by multiple readings of multiple raters (each paper was read by two raters during the rating process, followed by the reading by two researchers). In the set of 464 essays, and with 32 features considered, 14,848 changes (classification inconsistencies) could have occurred, meaning features could have been misidentified as present or

absent, then changed. Instead, four hundred eighty-four changes were made, indicating a 97% agreement among the raters (Morse et al., 2007).

### Summary

Holistic rating, analytic rating, and Prominent Feature Analysis serve different functions, and are appropriate in different circumstances. When reliable and valid, holistic scoring is sufficient for judgement of a paper vs. a single threshold for mastery/proficiency, and analytic assessment provides a moderate amount of information which may subsequently guide instruction. Among the three kinds, Prominent Feature Analysis distinguishes itself by its high potential to elucidate numerous characteristics of student writing. As previous research demonstrated, the more voluminous and specific the feedback, the higher the chance of growth in writing ability (Sommers, 2006). Therefore, Prominent Feature Analysis may prove useful in informing instruction and fueling self and peer review process in context of complex writing genres.

## Importance of Mastering Scientific Writing for Novice Scientists

As students progress in their academic careers into post-secondary levels, domain-specific writing becomes a new challenge that students must master. Students not only must be able to express their thoughts with clarity and precision; they also must do so in a highly-prescribed and rigorous manner. One example of such domain-specific writing types is scientific writing. Good scientific writing follows a long and strict set of rules concerning manuscript format, appropriate and expected grammar and style, usage of field-specific jargon, and correct source attribution, among others. Different disciplines use different manuscript formatting styles; social and behavioral sciences use

the American Psychological Association's manuscript formatting guidelines (VandenBos, 2010). It is important for students in these domains to master the expected rules of written scientific communication, as such proficiency is tied to their academic and post-graduation success. Not following these guidelines impedes students' or novice practitioners' academic or professional progress by, for example, thwarting their professional publication efforts.

Editors of *Research in the Schools* sought to understand the impact of careless scientific writing and formatting on acceptance for publication decisions of the editorial staff (Onwuegbuzie, Combs, Slate, & Frels, 2010). They learned that overall poorly written manuscripts were 12 times more likely to be rejected than well-written ones, while poorly-structured submissions were 5 times more likely to be rejected than their well-organized counterparts. Manuscripts with subpar literature review sections were six times more likely to be rejected than submissions with adequate literature reviews, and three or more incorrect citations resulted in four times the likelihood of rejection over citation error-free documents. Manuscripts that had nine or more violations of the APA publication guidelines were three times as likely to be rejected as their less-incorrect counterparts; manuscripts that contained errors in eight or more different categories were four times more likely to be rejected than their less-incorrect counterparts.

Behavioral science university faculty echo Onwuegbuzie and colleagues' (2010) sentiments regarding the prevalence of formatting mistakes in APA-formatted writing efforts. Greenberg (2015) reported on her recent effort to use rubrics to improve the quality of "APA-formatting style-compliant" novice scientific writing enrolled in research methods course in psychology. The rubric used in the study helped to guide the

students through the writing process, and provided a checklist to make sure all the important information and formatting elements are included. It was divided into 3 sub-areas, Content, Expression, and Formatting. Content focused on introduction and literature review. It guided the students through an introduction of the topic, summarizing only the relevant past literature, defining the purpose of the present study, and clearly stating relevant hypothesis(ses). Expression highlighted the need for organization, correct mechanics, tone, and appearance; Formatting addressed the specifics of in-text attributions.

Students enrolled in six sections of the course penned empirical study reports on a "true" experimental study. Students enrolled in three of the sections used the rubric as they worked on their writing ($n$ = 78), while students enrolled in the other three sections ($n$ = 68) did not. Students who used the rubric ($M$ = 79.50, $SD$ = 14.40) significantly outscored the students who did not ($M$ = 73.70, $SD$ = 17.50), $t(144) = 2.20$, $p = .03$, $d = .36$.

While these results are promising, the rubric used in the described study was rather broad and did not address many of the nuances of the APA manuscript formatting style. It is likely that a more-detailed rubric would be even more helpful to students and instructors. It may be possible to generate such rubric using the Prominent Feature Analysis.

**Summary**

Onwuegbuzie et al. (2010) clearly demonstrated why scientific writers must pay attention to multiple features of writing, including language use, organization, citation conventions, and other genre-specific requirements. Mastering the peculiarities of the

genre requires attention to multiple aspects of writing, many of which are not intuitive (i.e., the format of references section or in-text citations). It seems that copious and specific feedback is necessary for mastering scientific written communication requirements. It is important that faculty are well-equipped to help students master this genre. Prominent Feature Analysis of novice behavioral scientific writing may prove helpful in setting direction for instruction and self and peer review efforts.

## Assessment Type and Feedback

Writing is an iterative process. Writing, revising, and editing are separate steps of composing a written work, and best not confused. They require different focus: when writing, the author engages in "top-down" cognitive processes, or processes that are directed by the writer. While editing, one often seeks clues from the text ("bottom-up" processing) to guide his or her attention. However, an ability to spot the mistakes in the text and correct them often hinges on experience, and thus requires the help of a more knowledgeable other. The richer and more nuanced the analysis of the writing sample, the better the feedback available to both teacher and student.

### Holistic Assessment and Feedback

Holistic assessment has been repeatedly criticized for its inability to direct subsequent instruction due to lack of relevant feedback. The GRE writing subtest is an example of a holistically scored assessment. The test assesses students' critical thinking ability to "reason, assemble evidence to develop a position and communicate complex ideas" (ETS, 2018), as well as the command of syntax, semantics, and spelling. A student who receives a score of 3 and 2.5:

> Displays some competence in analytical writing, although the writing is flawed in at least one of the following ways: limited analysis or development; weak organization; weak control of sentence structure or language usage, with errors that often result in vagueness or lack of clarity. (ETS, 2018, para. 6).

This description is helpful and meaningful in a context of large-scale summative writing assessment; it may be less helpful in context of a classroom. An experienced writing instructor may be able to identify which of the possible mistakes are evident in the text, and may be able to suggest ways to overcome them. However, a student reading this description may not be able to identify the shortcomings at all, much less figure out how to correct them.

**Analytic Assessment and Feedback**

Analytic assessment provides more direction for improvement than holistic. The National Writing Project's AWC is an example of an analytic assessment tool. The AWC rating process generates 6 analytic scores reflecting the following qualities of writing: Content, Structure, Stance, Sentence Fluency, Diction, and Conventions (Swain & LeMahieu, 2012). Each of the six sub-skills is evaluated on a six-point scale. For example, the Diction attribute at score point 3:

> Contains words and expressions that are sometimes clear and precise; contains words that are primarily simple and general, yet adequate, contains mostly bland verbs or commonplace nouns and inappropriate modifiers; may include imagery or figurative language; when present, it is simple, and generally not effective. (Swain & LeMahieu, 2012, p. 51).

The scheme has been hailed the only analytic scheme to feature a combination of high reliability *and* a focus on features that are "authentic and central to student writing" (National Writing Project, nd).

The six sub-skills are thoroughly described, address a wide range of student writing characteristics, and provide direction for teaching to remedy the shortcomings. However, by design, the scheme attempts to be general enough to address many different writing types. To master a particular writing genre, a set of genre-specific best writing practices is needed.

While the above statement is true in primary and secondary education, it gets even more important in higher education. College-level writing requirements get very specific. An excellent laboratory experiment report calls for a very different writing style than a short science-fiction story. In such instances, a much more specific assessment than a typical analytic scheme would be helpful to both instructor and student as a catalyst for generating meaningful feedback.

**Prominent Feature Analysis and Feedback**

Prominent Feature Analysis has three powerful characteristics that relate to feedback to instructors and students. First, numerous specific features are identified, addressing syntax, semantics, style, and mechanics of writing. Second, both positive and negative features are noted. Last, the scale is authentic, relevant, and specific to the writing sample it is used to assess.

In a study of seventh-grade expository writing, 32 prominent features were identified, 22 positive features and 10 negative ones. In a subsequent study of third-to-fifth grade expository writing, the scale grew to 40 features, 27 positive and 13 negative

ones (Morse at al., 2007). Prominent Feature Analysis of a high-school sample of students in ninth and ten grades yielded 35 features, 24 positive and 11 negative ones (Morse et al., unpublished). These numbers provide a stark contrast to a single score of a holistic evaluation, or even a six-score result of an analytic evaluation, and provide a detailed picture of the writing of each student.

Additionally, Prominent Feature Analysis identifies both positive and negative aspects of student writing. Therefore, while not-yet investigated, a Prominent Feature Analysis scale may be sensitive enough to track progress of writers. Ideally, once identified, the negative features are addressed in the classroom. In time, classroom or one-on-one interventions may result in an observable increase of positive features and concomitant decrease of negative features in individual student's writing as well as class-wide.

Last, as the scale is derived from the writing sample it is used to assess, it has the power to elucidate what sophisticated means of expression writers at a given level are capable of, and what common and uncommon problems they encounter. In other words, no matter how rare, no linguistic tools (like metaphors in scientific writing) will be lost when writing is scaled with Prominent Feature Analysis. And if a small group of students is capable of using sophisticated and mature means of expression, perhaps these skills can be taught to others through careful scaffolding and intentional instruction.

**Summary**

Holistic and analytic rating schemes have been used to assess writing for many decades; research on their strengths and weaknesses abounds. While each type is appropriate for certain circumstances, they both lack one important characteristic: neither

provides feedback that is in-depth enough to significantly inform instruction, and to assist students with improving their writing. To inform instruction, a more detailed assessment tool is needed. Prominent Feature Analysis, due to its detailed and authentic structure, holds a promise to remedy this shortcoming. Results of past studies on Prominent Feature Analysis provide support for the scheme's reliability and validity for evaluating writing of students in elementary, middle and high school.

**Rationale for Present Study**

Prominent Feature Analysis allows for an uncommonly detailed picture of a writing sample to which it is applied. The features are derived by writing assessment and domain experts, from a sufficiently-large writing sample, assuring scale's authenticity, scope, and usefulness for feedback and instruction.

Scientific writing is a demanding and complex genre to master, rich in linguistic, stylistic, and genre-specific requirements. Genre-specific characteristics include manuscript organization, conventions for citing work of others, conventions for displaying figural information, proper jargon usage, and more. Mastering the rules and idiosyncrasies of the genre poses multiple difficulties for novice writers.

As it contains rich information about the writing it assesses, Prominent Feature Analysis results in much feedback for the student and the instructor. The scale's utility has been demonstrated for students in Grades K-12, and evidence has been gathered towards demonstrating its reliability and validity. It is not clear whether the scale retains its reliability and validity at post-secondary level. It is also not clear whether it can be used for understanding the characteristics of more-complex and demanding student-penned writing.

Generating and understanding the prevalence of positive and negative features of novice scientific writing may be an invaluable tool for both writers and instructors in improving novice scientific writing ability, and extends the utility of the Prominent Feature Analysis tool. The present study investigated the usefulness of Prominent Feature Analysis in assessing undergraduate level scientific writing.

# CHAPTER III

# METHODOLOGY

The purpose of this study was to investigate the usefulness of the Prominent Feature Analysis as a means for assessing and understanding novice undergraduate behavioral scientific writing. Scientific writing poses a notable challenge to novices; an authentic and detailed rating scale is needed to evaluate it and to guide instruction. Having a reliable and valid tool to systematically evaluate student scientific writing, while simultaneously providing detailed and structured feedback, will benefit both instructors and writers. The more voluminous and specific feedback students have, the richer the direction for writing improvement.

In this chapter I present the method used for the current investigation. The chapter includes a description of study's participants, a description of the writing sample and statistical tools used for its analyses, and a description of the procedure followed to execute the study.

## Research Questions

The study goal was to extend Prominent Feature Analysis to a new genre and a new demographic segment and to investigate its psychometric properties. The following four questions guided this study:

1. What writing characteristics comprise the Prominent Feature Analysis scale for novice behavioral scientific writing?

2. What are the relationships among the identified features?
3. Do students' prominent features scores relate to their college GPA, or ACT scores, including composite score, and language, math, science, and reading sub-scores?
4. Can the Prominent Feature Analysis scale be used to assess change in student writing when two samples from the same students are compared across time?

## Study Participants

An ideal student participant for the present study is familiar with the publication manual of the APA (VandenBos, 2010) and is expected to apply its manuscript formatting requirements to the best of his or her abilities when generating scientific writing. In many classes students are told to write their papers following the APA-required manuscript format, yet often these directions imply following APA's format of in-text citations and references only. To investigate undergraduate students' ability to generate scientific writing that follows professional guidelines in behavioral sciences, participants in the current study were expected to write their entire class papers in the APA-required professional manuscript format, based on stated class objectives.

The study's purposive sample consisted of novice scientific writers from Mississippi State University, previously enrolled in EPY 3513 (Writing in the Behavioral Sciences), EPY 4033 (Application of Learning Theories in Educational and Related Settings), and PSY 3314 (Experimental Psychology). Per their Mississippi State University records, a clear majority of the 208 participants self-identified as females ($n$ = 192; 92%). A majority of participants self-identified as Caucasian ($n$ = 154; 74%), or

African American (*n* = 45; 22 %). These gender and race statistics are representative of the majors which most participants were pursuing, including Educational Psychology (*n* = 169; 81.3%) and Psychology (*n* = 30; 14%). Participants' ages spanned from 19 to 48 years old (*M* = 21.95; *SD* = 2.3) A majority of participants were 20-23 years old (*n* = 185; 88.5%). This was expected; the papers originated in classes most commonly taken by undergraduate students in their junior and senior year. See Table 1 for more information regarding participants' demographics. The required sample size was determined based on previous research on Rasch models, suggesting that, for one parameter logistic Rasch model and dichotomous items, satisfactory item estimates can be obtained using samples of a minimum of 200 participants (e.g., Lai, Teresi, & Gershon, 2005). Only participants' writing samples were analyzed; no further actions were required of students.

Table 1

*Participants' Demographic Information, N = 208*

| Major: | | Sex: | |
|---|---|---|---|
| Edu. Psychology (EPY) | 169 (81%) | Male | 16 (8%) |
| Psychology (PSY) | 27 (13%) | Female | 192 (92%) |
| Interdisciplinary Studies | 4 (2%) | | |
| PSY/EPY | 2 (1%) | Race: | |
| PSY/English | 1 (0.5%) | Caucasian | 154 (74%) |
| Biological Science | 1 (0.5%) | African American | 45 (22%) |
| Accounting | 1 (0.5%) | Hispanic/Latino | 3 (1%) |
| Human Sciences | 1 (0.5%) | American Indian | 1 (0.5%) |
| Kinesiology | 1 (0.5%) | Multiracial | 3 (1%) |
| Secondary Education | 1 (0.5%) | Unknown | 2 (0.5%) |

*Notes*: Gender and racial make-up is representative of the primary majors included in this sample. Percent amounts for each category may not add to 100 due to rounding.

## Materials

A total of 233 student papers were reviewed in this study. To answer RQs 1, 2, and 3, 208 independent writing samples were reviewed. RQ 4 was answered by comparing 25 of the literature reviews penned by students in EPY 3513 (Scientific Writing), and reviewed to answer RQs 1-3, to same-student writing efforts completed during a following semester, in EPY 4033 (Learning Theories). To maintain independence of writing samples, these 25 additional repeat writing efforts from a subgroup of participants were not used in scale calibration through Rasch analysis.

The 208 writing samples used to answer RQs 1, 2, and 3 consisted of 29 empirical study reports (by PSY 3314 students), 154 short literature review papers (by EPY 3513 students), and 25 long literature review papers (by EPY 4033 students). The remaining 25 samples used for answering RQ 4 consisted of long literature reviews completed in EPY 4033. All papers generated in these three classes represent the behavioral scientific writing genre, and the students were required to follow the APA publication manual's guidelines (VandenBos, 2010).

### PSY 3314 (Experimental Psychology)

PSY 3314 is a junior-level class for psychology majors. The course has a lecture component taught by the instructor of record, and multiple laboratories, taught by graduate-level teaching assistants. During a semester, students participate in several short scientific experiments in their respective laboratories, and learn how to write experiment reports. Twenty-nine experiment reports written by PSY 3314 were analyzed in this study.

All final reports in PSY 3314 are to adhere to APA manuscript formatting guidelines, meaning they are to contain the following sections: an abstract, an introduction, literature review, method, results, discussion, and references. As students choose the experiment they conduct for their final project, report content varies. Sample final experiments include an investigation of influence of background music type on maze completion time, or an investigation whether bold-colored words shown on a computer screen were more likely to be recalled than pastel-colored words. Completion of this course satisfies a university requirement for students to take a junior or senior level course in which writing is emphasized.

**EPY 3513 (Writing in the Behavioral Sciences)**

EPY 3513 is a junior-level class for educational psychology majors. During the semester, students learn how to evaluate and summarize published scientific literature, and write a short literature review paper (500 to 650 words, excluding references) based on five peer-reviewed empirical sources on a psychology-related topic of their choice. All literature review papers completed in EPY 3513 are to adhere to APA manuscript formatting guidelines; completed papers include introduction, literature review, discussion, and references. Sample paper topics include investigation of effectiveness of antibullying programs in schools, or the cognitive benefits of bilingualism. In total, 154 literature review papers written by EPY 3513 students were analyzed in this study. Completion of this course satisfies a university requirement for students to take a junior or senior level course in which writing is emphasized.

### EPY 4033 (Application of Learning Theories in Educational and Related Settings)

EPY 4033 is a senior-level, capstone class for educational psychology majors. During the semester, students individually choose a topic related to learning, and write a 2,500 to 3,750-word literature review paper based on at least 15 empirical, peer-reviewed sources; this word count does not include an abstract and references. Sample topics include the impact of illicit drugs on memory, or usage of music therapy by speech-language pathologists. Fifty literature review papers written by EPY 4033 students were analyzed in this study. Twenty-five were used to RQs 1, 2, and 3; 25 were used to answer RQ4.

## Procedure

The study protocol was reviewed and approved by Mississippi State's Internal Review Board (see Appendix D). Based on the characteristics of the study, it was exempt from Internal Review Board's oversight. Writing samples collected prior to fall 2017 were deemed "existing data" by Internal Review Board, and, based on steps taken to preserve participants' anonymity, consent was not required. However, participant consent was obtained for samples that originated in courses offered in fall 2017 and spring 2018.

### RQ1: What Writing Characteristics Comprise the Prominent Feature Analysis Scale for Novice Behavioral Scientific Writing?

The writing analysis team consisted of four individuals: two writing instruction/assessment experts, and two scientific writing experts (me and another researcher). We used the Prominent Feature Analysis scale previously developed for seventh-grade expository writing as our starting point. Previous research results have

provided support for reliability and validity of the scale (Swain et al., 2010), and demonstrated stability of about two dozen features across samples (Morse et al., 2007). We deleted one feature from the original list (illegible handwriting); it was irrelevant, as all currently-reviewed samples were typed. In addition to verifying the relevance of the features present in the original scale to the current sample, we identified new features, specific to novice scientific writing (i.e., "excessive use of passive voice", or "design rigor").

Prior to analysis, each student writing sample was de-identified and assigned a unique identification number. A plain sheet of paper was stapled to the front of each sample; on it, we recorded the observed writing characteristic. We started with a blank sheet of paper instead of a checklist of features to ensure that we only record what is present in each paper, instead of looking for presence of all possible listed features. This process was employed in all previous Prominent Feature Analysis sessions conducted by the scale's authors.

During the initial group writing review session, all team members (two writing instruction/assessment experts, and two scientific writing experts) read 56 papers representing the two general sample types (literature review and empirical report), and noted the features that stood out. The team collectively discussed the prominent features present in each paper until 100% rater agreement was reached. Individual prominent features are treated as dichotomous scores; we noted each feature present the individual papers; others were implicitly regarded as absent.

I (a behavioral scientific writing expert) and one linguistic assessment expert together reviewed the remaining samples and consulted with the other two team members

as needed. This collaboration was necessary for two reasons. First, each of the two of us had different areas of expertise. I identified elements relating specifically to scientific writing requirements (i.e., misuse of scientific jargon, or in-text attribution errors), while the language-use expert identified specific grammatical and stylistic constructs (i.e., adverbial leads, or weak structural core sentences). Neither one of us could perform the analysis in the other's domain of expertise. Second, reading these papers side-by-side allowed us to discuss what we saw present in them, and minimize the chances of prominent features going unnoticed.

To investigate interrater agreement regarding papers not reviewed simultaneously by the entire team, I randomly selected 17 papers (14%) not reviewed by everyone, and provided each team member with his or her own copy. Each member assessed the 17 papers independently. I calculated the interrater agreement between the two writing assessment experts, and between myself and the other scientific writing expert separately. For each of the two reviewer pairs, I noted what percentage of features was identified as present/absent in each of the 17 papers (of possible 35 linguistic and 20 scientific, respectively) by both experts. For example, if I noted 8 features as present (implying 12 as absent), and the other scientific writing expert noted 10 features as present (implying 10 as absent), and 15 (of 20) of these judgments overlapped (meaning the same features were marked/not marked by both of us), then our interrater agreement was 75%. Lastly, I calculated an average of interrater agreements over the 17 papers within each pair.

Upon completion of Prominent Feature Analysis, the scale was calibrated using Rasch analysis of the 208 independent participants' results. Rasch analysis is a special case of an application of Item Response Theory. Item Response Theory is a paradigm for

creation and analysis of tests or scales; a model describing a relationship between one's latent trait/ability, and a probability of selecting a particular response. The basic logistic form of the IRT model considers three parameters of variation in responses. These parameters include: difficulty (location on a scale such that there is a probability of at least 0.5 of answering the question correctly by a participant whose latent ability equals the given difficulty parameter), discrimination (how well a given item distinguishes higher ability from lower ability respondents), and pseudo-guessing (a likelihood of getting credit for an answer without the requisite ability) (Crocker & Algina, 1986).

Rasch analysis is an example of a one-parameter model; all items are assumed to have equal discriminability, and pseudo-guessing is not considered. Based on previous research on Prominent Feature Analysis (e.g., Morse et al., 2007), I assumed the discriminability to be equal for all items. Therefore, along the test-taker's (in present study, writer's) ability, the only parameter required for items (in present study, features) was the item difficulty. Therefore, the probability of "success" on a dichotomously-scored item/feature is:

$$p(\text{success} \mid Bn) = \exp(Bn - Di) / [1 + \exp(Bn - Di)$$

where:

Bn is the ability of a person *n*

Di is the difficulty measure of the item/feature *i*

To be appropriate for Rasch analysis, three assumptions about the measure must be met: the scores must be generated by independent participants, a unidimensional latent trait must underlie the items, and the items must be independent in a given sample (Crocker & Algina, 1986). For convenience in calibration, features were placed on a scale

having a mean of zero. Each logit (log unit, similar to standard deviation) above zero implies an equal interval of difference in challenge level, such that higher scores (e.g., +2.3) imply features that are "harder" (less frequently observed), and lower scores (e.g., -1.6) imply features that are "easier" (more frequently observed).

Per Linacre (2017), a good model fit includes a comparison of predicted measure outcomes vs. the actual measure outcomes. Two indices helped me in making the determination regarding model fit: infit and outfit. Infit weighs more heavily results from items that closely match the participant's estimated ability. It focuses on information pertaining to the overall performance of an item or participant, and is based upon a standardized relationship between the expected and observed performance. Outfit assumes all person-item outcomes are equally weighted, and focuses on instances where predicted values do not match the observed. Both infit and outfit values are standardized to an expected value of 1.0. When data are too unpredictable (underfit the model; the amount of observed noise in data exceeds the predicted amount of noise), the fit statistics exceed 1.0. When data overfit the model, the amount of observed noise is less than expected, and the fit statistics fall below 1.0. Values of fit that fall between 0.5 and 1.5 are deemed acceptable, and suggest that items (here, features) are useful for measurement. Fit values outside of 0.5 and 1.5 suggest a need for additional inspection (Linacre, 2017). Additional information regarding Rasch scaling and analysis can be found in Rasch (1960; reprinted in 1980), or Bond and Fox (2015).

For item calibration purposes, positive prominent features were scored as "1" if present, and "0" if absent. Negative prominent features were scored "1" if absent, and "0" if present. Therefore the "desired state" for each feature was always noted as 1 (presence

of a positive feature, and absence of a negative one). I used the WINSTEPS Rasch measurement software (Version 4.0.1; Linacre, 2018) to calibrate the scale.

To confirm the scale's unidimensionality, I investigated the structure of the Prominent Feature Analysis scale in two ways. First, a principal component analysis of the Rasch modeled residuals was completed. This allows for detections of presence of additional factors vs. just random noise. The WINSTEPS software (Linacre, 2018) optionally executes this analysis. Second, an exploratory factor analysis of the feature scores was conducted using the FACTOR software (Lorenzo-Seva & Ferrando, 2018). I speculated that one factor (scientific writing skill) is responsible for a large portion of the score variance.

**Summary of analyses utilized to answer RQ1**. To create a prominent feature scale based on novice behavioral scientific writing, I and three other writing assessment experts conducted an analysis of student writing samples, and identified the prominent features which comprise the present prominent features scale. A portion of the papers were reviewed by all four team members; I reviewed the remaining papers as a scientific writing expert, working alongside a linguistic assessment expert. To calibrate the scale, I conducted Rasch analysis. To confirm unidimensionality of the scale, I conducted a principal component analysis of Rasch analysis residuals, and an exploratory factor analysis of feature scores.

### RQ 2: What Are the Relationships Among the Identified Features?

In addition to Rasch analysis, I conducted Pearson correlations to investigate positive and negative relationships among individual features. I used the IBM SPSS Statistics software (Version 24; IBM Corp., 2016). While Rasch analysis identified which

items are “easier” (present more often) or “harder” (present less often), Pearson correlations highlighted the bivariate relationships between features.

**Summary of analyses utilized to answer RQ2**. I conducted Pearson correlations to understand the relationships among the identified features comprising the present prominent features scale.

**RQ 3: Do Students’ Prominent Features Scores Relate to Their College GPA, or ACT scores, Including Composite Score, and Language, Math, Science, and Reading Sub-scores?**

I used the IBM SPSS Statistics software (IBM Corp., 2016) to conduct Pearson correlations to uncover relationships among the Prominent Feature Analysis scores and student achievement (represented by their university GPA and ACT scores). I correlated college GPA and ACT scores (including composite score and language, math, science, and reading sub-scores) with positive and negative linguistic features scores, positive and negative scientific features scores, and calibrated Prominent Feature Analysis scale score (meaning a score combining both the presence/absence of positive and negative features). These provided evidence for Prominent Feature Analysis score validity.

**Summary of analyses utilized to answer RQ3**. Pearson correlations were conducted to understand the relationships between the prominent feature scores, college GPA, and student ACT composite score and language, math, science, and reading sub-scores.

**RQ 4: Can Prominent Feature Analysis Scale Be Used to Assess Change in Student Writing When Two Samples from the Same Students Are Compared Across Time?**

I assessed a small number of short and long literature review assignments written by the same individuals (25 of each kind). I did so to investigate a potential change of prominent feature scores between two samples penned by the same group of students. Increases in positive features scores and/or decreases in negative feature scores may indicate increases in writing ability, or may be related to having a greater opportunity to demonstrate one's writing skills, based on task characteristic.

This was *not* a main line of the current investigation. However, if Prominent Feature Analysis were to be sensitive to change in writing skill, its potential utility would be greatly enhanced, from the perspective of instructors. To investigate potential changes in the writing skill, I compared the scores between the two sets of literature review papers, written during two consecutive semesters. The first set of papers was penned in EPY 3513; the second set was written while the students were enrolled in EPY 4033. I used the IBM SPSS Statistics software (IBM Corp., 2016) to separately investigate Pearson correlations between the positive prominent features scores, the negative prominent features scores, and the summed raw prominent features scores (the arithmetic difference between the positive and negative prominent features scores for each writing sample) between the two score sets. Using the same software, I also investigated whether the prominent features scores (positive, negative, and summed) were significantly different between the two sets using dependent *t*-test for paired samples. Significant differences between the scores of shorter and longer samples may demonstrate evidence

of growth as writers (in case of positive features' numbers increase, and/or negative features' numbers decrease).

**Summary of analyses utilized to answer RQ4.** Pearson correlations were conducted to understand the relationships between the prominent feature scores of two literature review paper sets, written by the same students, across time. Paired samples *t*-test was used to investigate differences between the score sets.

# CHAPTER IV

# RESULTS

This chapter contains the results of the analyses conducted during the present study. The following four research questions guided the current investigation:

1. What writing characteristics comprise the Prominent Feature Analysis scale for novice behavioral scientific writing?
2. What are the relationships among the identified features?
3. Do students' prominent features scores relate to their college GPA, or ACT scores, including composite score, and language, math, science, and reading sub-scores?
4. Can the Prominent Feature Analysis scale be used to assess change in student writing when two samples from the same students are compared across time?

## RQ1: What Writing Characteristics Comprise the Prominent Feature Analysis Scale for Novice Behavioral Scientific Writing?

As not all samples were simultaneously reviewed and discussed by all four team members (which would have resulted in 100% classification consistency), I investigated the interrater agreement among the expert pairs involved in sample analysis. Each of the four raters individually assessed 17 randomly selected student papers not reviewed

collectively. I compared the classification consistency of the two linguistic assessment experts, and two scientific writing experts. Writing assessment experts similarly identified linguistic features as present or absent 82% of time; scientific writing experts agreed on scientific writing features' presence or absence in 86% of cases. These values are commonly acceptable for writing assessment research (e.g., Englehard, 1992; Shohamy et al., 1992). See Table 2 for a list of prominent features identified in the present study; see Appendix E for definitions and examples of the present features set.

Table 2

*Prominent Features and Fit Indices*

| Feature | Type | Frequency | Difficulty | SE | Infit | Outfit |
|---|---|---|---|---|---|---|
| Hyperbole | L+ | 0 | ** | n/a | n/a | n/a |
| Aside to reader | L+ | 0 | ** | n/a | n/a | n/a |
| DV task exhibit | S+ | 1 | 5.28 | 1.01 | .97 | .19* |
| Design rigor | S+ | 2 | 4.58 | .72 | .98 | .47* |
| Alliteration | L+ | 3 | 4.16 | .59 | .95 | .38* |
| Metaphor | L+ | 6 | 3.42 | .42 | .98 | 1.23 |
| Sensory language | L+ | 7 | 3.26 | .39 | .89 | .72 |
| Noun cluster | L+ | 7 | 3.26 | .39 | .98 | 1.20 |
| Absolute | L+ | 7 | 3.26 | .39 | 1.03 | 1.34 |
| Narrative storytelling | L+ | 7 | 3.26 | .39 | .93 | .75 |
| Subordinate sequence | L+ | 8 | 3.11 | .37 | 1.02 | .81 |
| Coordinate sequence | L+ | 9 | 2.98 | .35 | 1.07 | .78 |
| Effective repetition | L+ | 12 | 2.66 | .31 | .98 | .66 |
| Striking words | L+ | 13 | 2.56 | .30 | .97 | .58 |
| Analysis rigor | L+ | 19 | 2.11 | .25 | 1.04 | 1.25 |
| Verb cluster | L+ | 32 | 1.44 | .21 | 1.09 | 1.30 |
| Cumulative sentence | L+ | 33 | 1.40 | .20 | 1.04 | 1.16 |
| Diction | L+ | 38 | 1.20 | .19 | .88 | .66 |
| Well-blended sources | S+ | 42 | 1.05 | .19 | .94 | .86 |
| Vivid verbs/nouns | L+ | 43 | 1.02 | .19 | .81 | .76 |
| Voice | L+ | 48 | .85 | .18 | .83 | .81 |
| Balance/parallelism | L+ | 49 | .82 | .18 | .91 | .83 |
| Attribution errors | S- | 49 | .82 | .18 | 1.13 | 1.07 |
| References errors | S- | 53 | .70 | .17 | 1.36 | 1.37 |
| Usage problems | L- | 60 | .49 | .17 | 1.03 | .99 |
| Coherence/cohesion | L+ | 70 | .23 | .16 | .76 | .69 |

Table 2 (Continued)

| | | | | | | |
|---|---|---|---|---|---|---|
| Effective organization | L+ | 72 | .17 | .16 | .79 | .72 |
| Transitions | L+ | 75 | .10 | .16 | .84 | .79 |
| Adverbial leads | L+ | 76 | .07 | .16 | .92 | .92 |
| Sentence variety | L+ | 87 | -.19 | .15 | .77 | .71 |
| Elaborated details | L+ | 88 | -.22 | .15 | .86 | .81 |
| Underdeveloped | L- | 93 | -.33 | .15 | 1.02 | 1.01 |
| Weak structural core | L- | 95 | -.38 | .15 | .93 | .89 |
| Procedural ambiguities | S- | 96 | -.40 | .15 | 1.20 | 1.24 |
| Inappropriate personification | S- | 102 | -.54 | .15 | 1.29 | 1.45 |
| Undefined terms/abbreviations | S- | 108 | -.68 | .15 | 1.15 | 1.25 |
| Required scientific elements missing | S- | 114 | -.82 | .15 | 1.18 | 1.19 |
| Lack of examples | S- | 128 | -1.14 | .15 | 1.06 | 1.06 |
| Misuse of terms/jargon | S- | 130 | -1.19 | .15 | 1.05 | 1.06 |
| Faulty punctuation | L- | 140 | -1.44 | .16 | 1.05 | 1.18 |
| Excessive passive voice | S- | 140 | -1.44 | .16 | 1.14 | 1.21 |
| Extrapolating beyond data/faulty logic | S- | 153 | -1.78 | .17 | 1.06 | 1.25 |
| List technique | L- | 161 | -2.01 | .18 | .99 | .86 |
| Faulty spelling | L- | 162 | -2.04 | .18 | .95 | .92 |
| Weak organization | L- | 167 | -2.20 | .18 | .98 | .83 |
| Redundancy | L- | 171 | -2.34 | .19 | .99 | .92 |
| Garbles | L- | 175 | -2.49 | .20 | .93 | .74 |
| Analysis/statistics misinterpretation | L- | 188 | -3.10 | .24 | 1.09 | 1.46 |
| Shifting point of view | L- | 193 | -3.43 | .27 | 1.02 | .80 |
| Misplaced modifier | L- | 194 | -3.51 | .28 | .99 | .99 |
| Statistical reporting error | S- | 196 | -3.68 | .30 | 1.07 | 1.40 |
| Wrong placement of scientific information | S- | 201 | -4.24 | .38 | .95 | .98 |
| Design flaw | S- | 204 | -4.75 | .45 | .80 | 1.24 |
| Hypothesis incongruent w/ presented literature | S- | 204 | -4.75 | .45 | .79 | 1.27 |
| Wrong analysis | S- | 206 | -5.17 | .46 | .44 | .82 |

*Notes:* L+/- signifies linguistic features, positive or negative; S+/- signifies scientific features, positive or negative. Lower Rasch difficulty values indicate positive features which higher number of students exhibited (or negative ones avoided). Conversely, higher Rasch difficulty values indicate fewer instances of positive, and of avoiding negative features, among student samples. Infit and outfit values are expressed as mean square. "*" indicates a possible overfit. "**" indicate features for which a true calibration value could not be obtained.

**Prominent Features Frequencies**

Combining linguistic and scientific features, a writing sample could earn up to 28 positive marks (24 linguistic and four scientific), and up to 27 negative marks (11 linguistic and 16 scientific). The minimum number of features, positive and negative, noted in a paper was 3, the maximum number was 25. The mean number of features in a paper was 12.45 (*SD* = 3.88). The highest numbers of features noted in a paper was 17 positive (*M* = 4.11, *SD* = 3.96), and 16 negative (*M* = 8.34, *SD* = 3.25). The mean values demonstrate that, on average, papers exhibited more negative than positive features.

Not surprisingly, more positive linguistic features were noted in papers on average than scientific features; only four positive scientific features were identified in the current sample/scale, in comparison with 24 positive linguistic features. Similarly, on average, more negative scientific features were noted than linguistic, which corresponds to a higher number of possible identified scientific writing errors. Additionally, as some features were opposites (i.e., effective organization and weak organization), they were unlikely to both rise to prominence in one paper. If they were to be both marked (for example, by two different readers), one would be removed, or replaced with another feature, upon a group discussion. This reduced the number of features that could be simultaneously noted in single paper. See Table 3 for overall counts of prominent features present in the analyzed writing sample.

Table 3

*Descriptive Statistics of Prominent Features Categories; N = 208*

| Features | Min. Observed | Max. Possible | Max. Observed | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Positive PF Sum | 0 | 28 | 17 | 4.11 | 3.96 |
| Negative PF Sum | 0 | 27 | 16 | 8.34 | 3.25 |
| Positive Ling. PF Sum | 0 | 24 | 15 | 3.80 | 3.67 |
| Negative Ling. PF Sum | 0 | 11 | 8 | 3.25 | 1.94 |
| Positive Sci. PF Sum | 0 | 4 | 2 | 0.31 | 0.57 |
| Negative Sci. PF Sum | 0 | 16 | 10 | 5.09 | 1.95 |

*Note*: Minimum number of possible features (0) = minimum number of observed features

**Rasch Analysis**

Rasch analysis revealed the presence of items separated into about 6 levels of difficulty (item separation of 5.81 with .97 reliability). This means that, in terms of frequency with which they were noted, some items were of very low difficulty (appeared often), some were of very high difficulty (appeared seldom or never), with four other distinct levels of difficulty in between. Having at least three levels of item separation with item reliability greater than .9 is necessary to confirm the item (here, feature) difficulty (and construct validity) of the measure (Linacre, 2017). The estimated feature difficulty of the 53 features exhibited in student writing spanned over 10 logits (from 5.28 to -5.17; logits are log units, similar to standard deviation). Both results are encouraging, as they demonstrate the scale contains items of varied difficulty, meaning it reflects the diverse student writing ability levels of the sample.

Review of infit mean square values revealed that all 53 prominent features noted in the writing sample displayed model-data fit between 0.5 and 1.5 (thus they are deemed useful for measurement). Review of outfit mean square values revealed three features exhibited overfit: exhibition of DV task, design rigor, and alliteration. All three of these

items were of very high difficulty: presence of the exhibition of the DV task feature was marked only once in the dataset; design rigor was marked twice (one of these instances was generated by the same student who presented the exhibition of the DV task feature), and alliteration was noted three times. As these instances were observed only in case of six students, they are unlikely to distort the measurement utility of the model.

In Rasch analysis, the difficulty of items is arranged on a scale where the higher difficulty values the fewer participants had incorporated the features in into their writing, or, in case of negative features, the fewer students avoided them. For the most part, negative features were easier to avoid than positive features to earn in present analysis. This means the study participants had less difficulty exhibiting error-free writing, than showcasing complex and nuanced means of expression. The negative features exhibited difficulty values from -5.17 (analysis error) to -0.13 (underdeveloped). Three negative features proved hardest to avoid: attribution errors (159 cases), references errors (157 cases), and usage problems (148 cases), with Rasch difficulty values 0.82, 0.70, and 0.49, respectively. This implies that these features were likely to be present in writing of students of various ability levels. The positive features were harder to earn, with -0.22 Rasch difficulty value of elaboration to 5.28 value of exhibition of DV task discussed earlier. Two features were not present in the current sample: aside to the reader, and hyperbole. See Table 2 for listing of features' difficulty levels, and their fit indices.

**Rasch analysis assumptions**. Three assumptions should be met for Rasch analysis to be appropriate: a unidimensional construct must underlie the scale, the items must be independent, and the scores must be generated by independent participants (Crocker & Algina, 1986). The unidimensionality of the underlying construct

(presumably, writing ability) was confirmed by the residual component analysis, as well as by an exploratory factor analysis.

In the current model, about 52% of raw variance is explained by the measure (a combination of items and people). If too much pattern/order is present in the raw variance not explained by the measure, there is a possibility of other constructs/factors underlying the model. Inspection of the residuals revealed 3.6% of total variance in the first (and largest) contrast. That is less than the variance explained by the prominent features (39.6%). As well, because the first contrast accounted for less variance than two features if all variance were common (2/53 = 3.8%) I conclude that this evidence is supportive of a unidimensional scale.

The next analysis was an exploratory factor analysis of the feature scores. I conducted an exploratory factor analysis with FACTOR software (Lorenzo-Seva & Ferrando, 2018), using polychoric correlations as basis for the dispersion matrix, and parallel analysis as the method for determining the number of factors to extract. The use of polychoric correlations is preferred, as the individual features may represent latent traits that are not dichotomous, but continuous. While one student may use a single cumulative sentence in his or her writing, another individual may use multiple ones, yet both will get a"1" as a score. The results suggest the presence of one factor/dimension underlying the answers, with only one eigenvalue exceeding the value of one; with goodness of model-data fit index value of .914 for a one-factor model. I therefore conclude that the Prominent Feature Analysis scale is unidimensional.

The independence of items was confirmed with absence of overly high correlations between the features; the highest noted Pearson correlation was .71 (see

discussion on correlations below). Lastly, no participant contributed more than one writing sample among the 208 which comprised the present Rasch model, assuring the independence of observations.

### RQ 1 Results Summary

The 53 prominent features derived from the analysis of 208 writing samples (partially based on previously-created scale) constitute a unidimensional scale of novice behavioral scientific writing ability. As hypothesized, numerous new, genre-specific features were added to the original scale. Data fit the model well. Based on item difficulty values, it is easier for students to avoid common errors in this genre than to exhibit more complex "good writing" characteristics.

## RQ 2: What Are the Relationships Among the Identified Features?

Given the presence of 53 features in the present sample, the number of bivariate correlations between features is 1,378. Of these, 340 were statistically significant. See Appendix F for the complete correlation matrix among the prominent features identified in the present writing sample. As the features appear to share one underlying construct, writing ability, multiple significant correlations are expected. In the analysis below, I am using values of .1 to .3 to signify weak correlations, .3 to .5 to signify moderate correlations, and .5 to 1 to signify strong correlations (Cohen, 1988).

Five observed significant correlations were strong, while 48 were moderate. For the most part, linguistic features correlated with other linguistic features, and scientific features correlated with other scientific features. However, while there were multiple strong and moderate correlations between linguistic features, there were very few

correlations between scientific features, and only one scientific feature correlated moderately with three linguistic ones. Broadly speaking, this suggests that overall writing proficiency does not automatically assure the ease of mastering the rules of behavioral scientific writing in novices. The following information helps to elucidate some of the identified relationships.

**Linguistic Features Correlation Cluster**

An interesting set of relationships emerged among a group of 10 linguistic features exhibiting strong and moderate Pearson correlations with each other. Of the observed five strong and 48 moderate correlations in the entire prominent features set derived from analysis of present sample, 38 are included in this group (three strong and 35 moderate). This feature cluster appears to be strongly interrelated; see Table 4 for a list of correlation cluster features. Factor analysis of the group (using maximum likelihood extraction) revealed one underlying factor, which appears to be writing proficiency. That result was expected, however, since the analyses reported for RQ1 support a claim for unidimensionality.

Table 4

*Linguistic Features Correlation Cluster*

| Feature | Freq. | Correlates with the following features within the cluster |
|---|---|---|
| **Elaborated details** | 88 | vivid verbs/nouns; sentence variety; coherence/cohesion, voice, **effective organization** |
| Sentence variety | 87 | elaborate details, vivid verbs/nouns, diction, adverbial leads, balance/parallelism, effective organization, transitions, coherence/cohesion, voice |
| Adverbial leads | 76 | sentence variety, transitions, voice, narrative storytelling, vivid verbs/nouns, diction |
| Transitions | 75 | vivid verbs/nouns, diction, adverbial leads, sentence variety, effective organization, coherence/cohesion, voice |
| Effective organization | 72 | vivid verbs/nouns, diction, balance/parallelism, sentence variety, transitions, voice, elaborate details, coherence/cohesion |
| **Coherence/cohesion** | 70 | elaborated details, vivid verbs/nouns, diction, transitions, balance/parallelism, sentence variety, **voice**, **effective organization** |
| Balance/parallelism | 49 | vivid verbs/nouns, effective repetition, sentence variety, effective organization, coherence/cohesion, voice |
| Voice | 48 | elaborated details, vivid verbs/nouns, diction, adverbial leads, balance/parallelism, sentence variety, effective organization, transitions, coherence/cohesion |
| Vivid verbs/nouns | 43 | elaborated details, diction, adverbial leads, transition, balance/parallel, sentence variety, effective organization, coherence/cohesion, voice |
| Diction | 38 | vivid verbs/nouns, adverbial leads, sentence variety, effective organization, transitions, coherence/cohesion, voice |

*Notes:* Frequency denotes the number of times this feature was marked in the present sample. Strong correlations are bolded.

**Strong Pearson Correlations**

Five strong correlations were noted in the dataset, one between two positive scientific features, and four among the positive linguistic features. The only strong

correlation among scientific features was exhibit of DV task and design rigor ($r = .70$, $p < .01$). While it was the strongest correlation observed among *all* features, its practical significance is hard to assess; exhibit of DV task was noted only once in the entire sample of 208 papers, while design rigor was noted twice.

The highest strong correlation among linguistic features was observed between cumulative sentence and verb clusters ($r = .62$, $p < .01$). Cumulative sentences were noted in 33 writing samples, verb clusters were noted in 32 student papers, indicating both are relatively hard to master. This relationship is logical, as verb clusters are types of free modifiers, and cumulative sentences, by definition, are comprised of a base clause, and free modifiers. While a verb cluster cannot exist on its own, other constructs may comprise a cumulative sentence, which is why the correlation between these two features is less than one. As expected, this correlation also appeared in the previous prominent feature literature (Swain et al., 2010), and was the only moderate or strong correlation that co-occurred in both studies.

Effective organization, noted 72 times, strongly correlated with coherence/cohesion, noted 70 times ($r = .51$, $p < .01$), as well as with elaborated details, noted 88 times ($r = .50$, $p < .01$). In turn, coherence/cohesion, noted 70 times, strongly correlated with voice, noted 48 times ($r = .50$, $p < .01$). These relationships are not surprising, as linguistic cohesive devices often include elaborated details and examples, and commonly enhance a sense of internal organization.

**Moderate Pearson Correlations**

Forty-eight moderate correlations were present among features. Thirty-five of these occurred among 10 linguistic features (described in the linguistic correlations

cluster). Only three positive correlations were noted between scientific features, and only one positive correlation was noted between a scientific feature and four linguistic features.

**Moderate scientific features correlations.** Misuse of terms/jargon (a scientific negative feature), noted 78 times, correlated with undefined terms (a scientific negative feature), noted 100 times ($r = .31$, $p < .01$). This relationship suggests an overall struggle with discipline-specific language use; knowing how to use the jargon terms correctly, as well as knowing which terms are specialized enough to need defining, comes with experience. Surprisingly, design rigor positively correlated with hypothesis/thesis incongruent with reviewed literature ($r = .34$, $p < .01$). This relationship is unexpected, because design rigor is a positive feature and hypothesis/thesis incongruent with reviewed literature is a negative feature of scientific writing. However, each occurred in only a few instances in the dataset: design rigor was noted twice, while hypothesis/thesis incongruent with reviewed literature was noted four times. This suggests an idiosyncrasy of this particular dataset which may not appear in other samples. Lastly, wrong placement of scientific information, noted seven times, significantly correlated with statistics reporting error, noted 12 times ($r = .30$, $p < .01$). Both features broadly imply a "novice scientist" mindset.

Well-blended sources, a positive scientific writing feature, moderately correlated with three positive linguistic features: vivid verbs/nouns ($r = .48$, $p < .01$), voice ($r = .32$, $p < .01$), elaborate details ($r = .30$, $p < .01$) and sentence variety ($r = .30$, $p < .01$). This is an interesting find, because it denotes a relationship between an aspect of critical thinking

(using multiple sources to support a point), and a linguistic ability represented by the other three features.

***Moderate negative correlations.*** Two negative correlations were noted among moderate values. Effective organization negatively correlated with underdeveloped writing ($r = -.36$, $p < .01$) and with weak organization ($r = -.33$, $p < .01$). While weak organization is expected to negatively correlate with effective organization, the other correlation is more interesting. It suggests that, in the current writing sample, part of developing the paper involves organizing it well.

**RQ2 Results Summary**

As expected, numerous statistically significant correlations between features were noted; positive linguistic features correlated among themselves, and two negative linguistic features inversely correlated with positive features. I identified a cluster of 10 linguistic features which commonly co-occurred in papers (or were concurrently absent), and statistically significantly correlated among each other; they appear to represent writing proficiency. Contrary to my hypothesis, positive linguistic features for the most part did not correlate with positive scientific features, and positive scientific features did not correlate with positive scientific features. This suggests that overall writing competency does not necessarily assure an ease of acquiring command of scientific jargon.

## RQ3: Do Students' Prominent Features Scores Relate to Their College GPA, or ACT scores, Including Composite Score, and Language, Math, Science, and Reading Sub-scores?

To provide evidence towards scale validity, I investigated Pearson correlations between student GPA and prominent feature scores, as well as ACT (language, math, reading, science, and composite scores) and prominent features scores. Based on pervious Prominent Feature Analysis research (e.g., Swain et al., 2010), I expected the prominent features scores to correlate with college GPA, as good writers tend to do well in undergraduate level classes which commonly require writing.

### GPA and Prominent Features Scores

Student GPA correlated moderately and significantly with the scaled prominent features scores, meaning the sum of positive and negative scores assigned to student writing samples ($r = .37$; $p < .01$). Student GPA weakly though significantly correlated with the sum of positive linguistic and scientific prominent features scores ($r = .27$; $p < .01$), and inversely moderately correlated with the sum of negative linguistic and scientific prominent features scores ($r = -.36$; $p < .01$). However, a more interesting picture emerged when the correlations between GPA and linguistic and scientific features were considered separately. The strongest inverse correlation exists between the sum of negative linguistic features and GPA ($r = -.40$, $p < .01$; moderate strength). This suggests, that error-free writing is more common in "good" students than in those with lower grades. GPA was weakly positively correlated with sum of the positive linguistic features ($r = .27$; $p < .01$); GPA was also weakly inversely correlated with negative scientific

features ($r$ = -.21; $p$ < .01), and weakly correlated with positive scientific features ($r$ = .15; $p$ = .05). This may suggest that "good" students are better able to employ scientific thinking and use the scientific language more ably than students with lower GPA. However, only four positive scientific writing features were identified in the sample, and two of these (exhibit of the DV task and design rigor) were noted only one and two times, respectively. Therefore, the relationship of GPA with positive scientific features should be regarded with caution. See Table 5 for additional information regarding the relationships between prominent features scores and ACT and GPA.

**ACT and Prominent Features Scores**

Positive scientific prominent features scores did not significantly correlate with any ACT sub-scores, nor the ACT composite score. This was not surprising, due to the low overall number and few instances noted of positive scientific features.

Two unexpected sets of relationships emerged during the analysis of correlations between ACT sub-scores and prominent features sub-scores (excluding positive scientific prominent features scores discussed above). First, ACT math sub-scores exhibited the strongest and significant correlations with all prominent features sub-scores other than positive scientific features scores. Second, of all prominent features sub-scores, negative linguistic prominent features scores exhibited the strongest and significant inverse correlations with ACT language, math, reading, and composite scores. ACT science scores were an exception; these correlated most strongly with the raw prominent features scores (see Table 5 for strengths of relationships), though the correlation with negative linguistic prominent features scores was the next highest recorded value. This may mean

that students who avoid common grammatical and stylistic errors in their writing tend to do better on ACT than the students whose writing is more mistake-ridden.

Table 5

*GPA, ACT Descriptive Values; Pearson Correlations Between GPA, ACT and Prominent Features Scores*

| | GPA | ACT Lang. | ACT Math | ACT Reading | ACT Science | ACT Comp. |
|---|---|---|---|---|---|---|
| Minimum noted | 1.79 | 11 | 15 | 14 | 13 | 14 |
| Maximum noted | 4.00 | 36 | 34 | 36 | 36 | 35 |
| Mean | 3.27 | 23.98 | 21.37 | 24.19 | 22.31 | 22.53 |
| Standard Deviation | .53 | 5.43 | 4.34 | 5.15 | 3.72 | 4.23 |
| | | | | | | |
| Positive Sci. PF Sum | .15* | .10 | .13 | .02 | .08 | .11 |
| Positive Ling. PF Sum | .27** | .33** | .36** | .23** | .32** | .36** |
| Negative Ling. PF Sum | -.40** | -.44** | -.47** | -.34** | -.38** | -.45** |
| Negative Sci. PF Sum | -.21** | -.24** | -.26** | -.18* | -.24** | -.25** |

*Notes:* * indicates $p$ values < .05; ** indicates $p$ values < .01
Scaled prominent features score is derived by adding the number of positive and negative feature scores for each participant noted in a paper.

**RQ3 Results Summary**

As hypothesized**,** Prominent Feature Analysis scores positively significantly correlated with student college GPA scores, and their composite ACT scores. This provides support for validity of the scale. However, these correlations were not the strongest noted in the set. Interestingly, negative linguistic prominent features scores

exhibited strongest significant inverse correlations with GPA. This suggests that "good" students tend to generate error-free writing.

Additionally, ACT math sub-scores exhibited the strongest significant correlations with most of the positive and, strongest inverse correlation, most of the negative prominent features score categories. This may suggest that students who are strong in math tend to be overall good students; such students tend to be good writers.

## RQ4: Can Prominent Feature Analysis Scale Be Used to Assess Change in Student Writing When Two Samples from the Same Students Are Compared?

To explore the differences between two sets of writing penned by repeat students, I compared the prominent features scores assigned to short literature review papers by 25 students to the scores assigned to their long literature reviews completed in the following semester. The scores assigned to the 25 long literature reviews were not a part of the original dataset used for Rasch analysis.

Students in both classes were explicitly directed to generate writing that adhered to all relevant APA manuscript formatting guidelines. In both classes, students received instructor feedback on the drafts, and had opportunities to revise their writing prior to the final submission. Papers reviewed for this analysis were written during two consecutive semesters to minimize students' growth as writers due to uncontrolled-factors, like enrollment in other writing-heavy classes, or other classes that require the use of APA formatting style.

Based on results from RQ 3, I separately investigated correlations between the positive prominent features scores, the negative prominent features scores, and the

summed raw prominent features scores between the two score sets (the arithmetic difference between the positive and negative prominent features scores for each writing sample). I also investigated whether the prominent features scores were significantly different between the two sets using dependent *t*-test for paired samples. Again, I compared the positive prominent features scores, the negative prominent features scores, and the summed raw scores. See Table 6 for the overview of descriptive statistics of the writing sample.

Table 6

*Descriptive Statistics and Correlations of Prominent Features in the Two Essay Sets*

| | Set 1 | | Set 2 | | |
|---|---|---|---|---|---|
| Score | Mean | SD | Mean | SD | *r* (Set1, Set2) |
| Positive PF | 2.96 | 2.78 | 5.56 | 4.07 | .13; $p = .53$ |
| Negative PF | 8.24 | 2.49 | 9.60 | 2.90 | .35; $p = .09$ |
| Raw PF sum | -5.28 | 4.57 | -4.04 | 6.46 | .34; $p = .10$ |

*Notes*: Both sets of papers were penned by the same 25 students. Raw prominent features sum is derived by subtracting the number of negative features noted in a paper from the number of positive ones.

The bivariate correlations between positive, negative, and raw summed prominent features scores between paper sets 1 and 2 were not statistically significantly different from zero. Significantly more positive prominent features were noted in Set 2 ($M = 5.56$, $SD = 4.07$) than Set 1 ($M = 2.96$, $SD = 2.78$); $t(24) = -2.81$, $p < .01$. Significantly more negative prominent features were noted in Set 2 ($M = 9.60$, $SD = 2.90$) than Set 1 ($M = 8.24$, $SD = 2.49$); $t = -2.20$; $p = .04$. No statistically significant differences were noted between the summed prominent features (obtained by subtracting the negative prominent

features scores from the positive prominent features scores) between Set 1 and Set 2. See Table 7.

These results suggest that prominent features scores may be prone to task effect, and may depend on passage/text length. While the significantly higher number of positive features between the sets is promising (though, on average, small), the number of negative features significantly increased as well. This may reflect the overall higher word/page counts between the assignments, meaning longer essays allowed students to exhibit more positive writing skills, but also provided more opportunity for making mistakes.

Table 7

*Paired Samples t-test of Two Sets of 25 Writing Samples*

| | | | 95% Confidence Interval of the Diff. | | | | |
|---|---|---|---|---|---|---|---|
| | Mean | Std. Dev. | Std. Error Mean | Lower CI | Upper CI | *t* (2 - tailed) | Sig. |
| Set 1 & 2 (Pos. PF) | -2.6 | 4.62 | 0.92 | -4.51 | -0.69 | -2.81 | .01 |
| Set 1 & 2 (Neg. PF) | -1.36 | 3.09 | 0.62 | -2.64 | -0.08 | -2.20 | .04 |
| Set 1 & 2 (Raw PF sum) | -1.24 | 6.54 | 1.31 | -3.94 | 1.46 | -0.95 | .35 |

*Notes:* Degrees of freedom were constant in all comparisons (*df* = 24). Raw prominent features sum is derived by subtracting the number of negative features noted in a paper from the number of positive ones.

**RQ4 Results Summary**

In comparison with shorter literature reviews, later-written, longer literature reviews exhibited significantly more positive prominent features as well as significantly more negative prominent features. This suggests that for novice writers in the present participant sample increase in passage length was both a blessing and a curse. It afforded

more opportunities for exhibiting complex means of expression, as well as making more mistakes. The prominent features scores did not correlate significantly between the sets, suggesting that, for this small participant sample and these two prompts, Prominent Feature Analysis was not suitable for tracking change in writing ability over time and over differing prompts.

# CHAPTER V

# DISCUSSION

Scientific writing is a challenging genre to master. It combines the elements of expository writing, demonstration of domain knowledge and sound scientific thinking, excellence in communication to diverse audiences, and adherence to strict formatting and stylistic rules. Students need practice in revising and editing to master scientific writing; extensive and specific feedback facilitates the process.

Novice scientific writers face multiple challenges when completing their manuscripts. Typical struggles include problems with writing overall, distinguishing between the required standard structure and content of the paper ("what goes where"), and an inability to consistently extract and describe the relevant information from literature of others (Shah, Shah, & Pietrobon, 2009). Additionally, novices differ from experts in their approach to using or citing the ideas of others. While experts tend to extract meaning from relevant literature, and use citations as attributions, novices place citations at the beginning of the point they are making. The former lends itself to deeper analysis and synthesis, while the latter results in a list of relevant but separate studies (Mansourizadeh & Ahmad, 2011). While this exact comparison has not been reported on, it is likely that the references and citations errors noted by Onwuegbuzie et al. (2010) were more common in the writing of novices than experts.

Prominent Feature Analysis allows for an in-depth understanding of the writing it is used to assess. During analysis, a rich and detailed picture of the writing sample is generated, allowing for ample and informative feedback. Prominent Feature Analysis has been demonstrated to be a reliable and valid tool for assessing elementary and high school-level writing. Extending Prominent Feature Analysis to novice scientific writing allowed for a methodical identification of present (and lacking) characteristics of the genre, and will be immediately useful for informing instruction and student revision processes.

The following four research questions guided the study:

1. What writing characteristics comprise the Prominent Feature Analysis scale for novice behavioral scientific writing?
2. What are the relationships among the identified features?
3. Do students' prominent features scores relate to their college GPA, or ACT scores, including composite score, and language, math, science, and reading sub-scores?
4. Can the Prominent Feature Analysis scale be used to assess change in student writing when two samples from the same students are compared across time?

To answer these questions, a prominent features scale was derived from 208 independent novice behavioral science student writing samples. Subsequently, the scale was calibrated using Rasch model. Additional relationships between the derived features were further explored through bivariate correlations. Evidence for the scale's validity was gathered by correlating the assessment results with student college GPA and ACT scores.

Lastly, relationships between features found in two writing samples penned across time by same group of 25 students were investigated.

## RQ1 Results Discussion: What Writing Characteristics Comprise the Prominent Feature Analysis Scale for Novice Behavioral Scientific Writing?

The present prominent feature scale is comprised of 55 characteristics. A writing sample could exhibit up to 28 positive features (24 linguistic and 4 scientific) and up to 27 negative features (11 linguistic and 16 scientific).

### Linguistic Prominent Features

Three new linguistic features rose to prominence, in addition to the ones used in the study of seventh grade writing; these included: coordinate construction, diction, and misplaced modifier (see Appendix C for definitions and examples of prominent features in present study). Coordinate construction occurs when two elaborate elements are linked with "and" or "but." It presently rose to prominence in descriptions of past studies included in students writing (for example, "Problems with social interaction and peer acceptance could lead to depression and other mental health issues throughout adolescence for individuals with Asperger syndrome (Elst, et al., 2013)."). Diction denoted a strikingly appropriate choice of words to meet the requirement of the scientific writing genre (for example, "Mrug et al. (2014) interviewed the girls and their parents to reveal characteristics of delinquency, best friend's deviant behavior, age of menarche, relational, physical, and non-physical aggression, and ethnicity."). Misplaced modifier occurred often when novice writers attempted to describe research procedure by others. The resulted usage of "they" confounded the actions of multiple subjects in the sentence

(for example, “These four therapy sessions were conducted to focus on how couples should behave for the betterment of their baby and their development.”).

Past literature on prominent features contains rich and nuanced discussion of previously-noted linguistic features; writing assessment experts are best suited to elucidate these. See Swain et al. (2010; 2012; and 2015) for discussion on recurring linguistic features.

**Scientific Prominent Features**

The strict formatting, stylistic, and usage requirements of scientific writing resulted in 16 negative, genre-specific features rising to prominence in present writing sample. These were accompanied by only four new positive scientific writing features. See Appendix E for definitions and examples of identified scientific writing features.

**Positive scientific prominent features**. Four positive scientific features emerged from the present sample: DV task exhibit, design rigor, analysis rigor, and well-blended sources. Of these, design rigor is applicable only to experiment reports. It addresses an explicit mention of actions undertaken to reduce the limitations of a study, for example, counterbalancing stimuli between trials, or randomly assigning participants to groups. Out of 29 experiment reports, two exhibited this feature.

DV task exhibit pertains to inclusion of a graphic representation (in addition to verbal description) of the task employed in a study. It was noted once in present sample; an experiment report included examples of mazes participants were completing during the study. Its rarity in the present sample may relate to limited exposure of students to certain types of scientific literature. While cognitive science experiment reports often contain images that foster visualization of complex stimuli presentations in blocks and

trials, literature pertaining to counseling methods or educational interventions is less likely to contain graphics. Most writing samples in the present study originated in educational psychology department, and only a handful of papers focused on cognitive science-based research.

Analysis rigor, noted 19 times, pertains to astute displays of understating of study limitations, overstated results, or other ambiguities and inconsistencies presented in literature by others. A relatively low count of this feature may reflect students' lack of faith in their common sense and critical thinking skills. Novices often implicitly trust scientific literature simply because it has been published.

Well-blended sources were noted 42 times in present sample. This feature pertains to skillful use of multiple sources to support an idea, and it embodies both critical thinking and writing skills. It was noted in more papers that the other three positive scientific features, suggesting it is the easiest of the four to learn and use. The relative rarity of this feature in novice scientific writing has been highlighted in past literature on novice scientific writing (Mansourizadeh & Ahmad, 2011). Novices often use a single source, and begin describing it by starting with author's name, for example "Smith (2018) investigated…." Using multiple sources to make a point, in particular when a concept (not a person) is the subject of a sentence, often comes with time and practice.

**Negative scientific prominent features.** Among 16 identified, two negative prominent features pertain specifically to experiment reports: design flaw (noted four times), and wrong analysis (noted twice). Design flaw pertains to study design which is not appropriate for answering the research question, for example attempting to demonstrate differences in effectiveness between two interventions without including a

pretest. Wrong analysis pertains to choosing an incorrect analysis to calculate study results, for example, an independent sample *t*-test in a study that involves one group of participants, tested twice. Both problems occurred rarely, and signify not so much struggles with writing, but lack of basic statistical knowledge. Requiring drafts prior to final project submissions may allow the instructor to identify these problems and address them before the final report is completed.

Another two statistics-based features emerged in both experiment reports and literature reviews, analysis/statistics misinterpretation (noted 20 times) and statistical reporting error (noted 12 times). Analysis/statistics misinterpretation, for example, took the form of reporting that a result was significant, while including a *p* value greater than .05. In contrast, statistical reporting error denoted presenting correct information in general, but failing to follow statistical reporting conventions, for example, not including mean and standard deviation information with statistical test results. These two problems, again, signify lack of statistical experience more so than inadequacies in writing ability.

Two negative scientific features which seemed harder to avoid include attribution errors (crediting outside sources in text; noted 49 times), and references errors (noted 53 times). Multiple attribution error types were present in the sample, including, among others, failing to acknowledge a source altogether, failing to include all authors' names the first time a paper is discussed, failing to use first author only and "et al." in subsequent mentions of a paper authored by three or more authors, failing to include a year of a publication, or failing to include page numbers with direct quotes. References errors included incorrect capitalization, incorrect italics, incorrect order of elements within a reference, lack of page numbers, or lack of doi number. Others agree that

references and attribution errors seem to pose significant difficulties for scientific writers (i.e., Onwuegbuzie et al., 2010), not only for novices. These problems most likely signify lack of experience with scientific writing, as well as simple carelessness. While familiarity with writing comes with experience and time, carelessness can be perhaps addressed in a classroom through targeted interventions sensitizing students to potential attributions and references traps.

Two features that rose to prominence in the current writing sample addressed shortcomings in reporting on methodology of past research: procedural ambiguities (noted in 112 papers), and lack of examples (noted in 80 cases). Procedural ambiguities feature was marked in papers in which the participants, order of tasks, and/or the tasks themselves, were not adequately described. For example, a description of a survey study would lack information regarding the number of participants or other essential participant characteristics (like gender or age), omit method of survey delivery, or not include any characteristics of a survey (like survey length or question type). The lack of examples feature pertained to missing information, vital to comprehend a study. For example, a student may describe an intervention aiming at reducing disruptive classroom behaviors in a child with autism, yet never actually specify what constituted disruptive behaviors at the heart of the study. While these features may signify shortcomings in conventions of scientific writing, they may also represent a lack of student understanding of what the described study entailed.

Six identified features pertained to more "mechanical" use of language in scientific writing, or APA format requirements. These included required scientific elements missing (noted 94 times), wrong placement of scientific information (noted

seven times) undefined terms/abbreviations (noted 100 times), misuse of terms/jargon (noted 78 times) inappropriate personification (noted 106 times), and excessive passive voice (noted 68 times). Scientific elements missing from student papers included APA-required section headings, or entire sections, like abstract, discussion, or references. Alternately, all the required elements may be present in a student's paper, but some may be misplaced, for example, results may be stated in the method section. This appears to be a common struggle for novice writers, previously noted by others (Shah et al., 2009).

Undefined terms/abbreviations signified lack of definitions of key concepts in the paper (sometimes the focus of investigation). In some cases, it may have been assumed that a term is common knowledge in the domain of behavioral sciences, for example, a paper on applied behavioral analysis did not contain a definition of the technique, or a paper on attention deficit hyperactivity disorder referred to the disorder by its abbreviation only, never spelling out what the abbreviation stands for. Misuse of jargon may include a mistaken use of the term "experiment" instead of the more general term "study," or references to study results as proven facts. Inappropriate personification involved referring to studies as performing human actions, for example, "the study researched." Lastly, while passive voice is commonly used in scientific writing (especially in a results section), its excessive use rose to prominence in the sample, and many instances were awkwardly worded, for example "literature was found, where authors…." Some of these mistakes likely result from lack of exposure to scientific literature, or lack of explicit guidance while reading scientific literature. Once a "more knowledgeable other" points out explanations of the jargon or uncommon scientific terms in writing of others, students may be able to apply this behavior in their writing.

Additionally, some of these mistakes may signify lack of experience with science as much as with scientific writing. A seasoned researcher is not likely to use the term "experiment" when describing a survey study.

Two negative scientific prominent features that addressed flawed scientific thinking included hypothesis incongruent w/present literature (noted four times), and extrapolating beyond data/faulty logic (noted 55 times). The first contained a few instances where the presented literature would fail to make case for a hypothesis (in case of an experiment report), or the thesis sentence/topic of the paper would not reflect the focus of the summarized research by others (in case of a literature review paper). These may result from flawed thinking; however, it is also possible they reflect inadequate final revision process, where the final version of the paper does not "smooth out" all the additions and subtractions of information which took place along the writing process. Faulty logic took many forms in student papers, often representing attempts in making meaning of science by generalizing the results too broadly. For example, a novice may write "everybody can benefit from counseling."

The high number of identified negative scientific features, and the low numbers of positive ones, confirm that the genre, overall, poses difficulty for novices. However, a more nuanced picture emerges from considering the Rasch analysis results. While, overall, plenty of mistakes are possible, avoiding them happens more commonly than featuring the "hard" positive characteristics.

To summarize, linguistic prominent features noted in the sample differed very little from the previous prominent feature analysis of seventh grade writing that served as a starting point for the current effort. Additionally, 20 features (four positive and 16

negative) illuminating the genre-specific characteristics of novice behavioral scientific writing were identified. Broadly speaking, these features highlighted the difficulties in thinking like a scientist as much as writing like one. The features also underscored the difficulties novices face when attempting to format their work to follow the requirements of the APA manuscript formatting guidelines (VandenBos, 2010). These format difficulties are also common among more seasoned scientific writers (Onwuegbuzie et al., 2010).

## RQ2 Results Discussion: What Are the Relationships Among the Identified Features?

Statistically significant correlations were noted among 340 identified prominent features (see Appendix F for the complete correlation matrix). Five observed significant correlations were strong, while 48 were moderate. For the most part, linguistic features correlated with other linguistic features, and scientific features correlated with other scientific features.

I identified a cluster of 10 linguistic features which commonly co-occurred in papers (or were concurrently absent), and statistically significantly correlated among each other; they appear to represent writing proficiency. An argument can be made that identifying features present in this group reflects a bias or heightened sensitivity of the reviewers to certain characteristics of writing. However, many of them represent unambiguous grammatical constructs, like balance/parallelism or adverbial leads. They either were objectively present in student writing or they were not. At the same time, other features in this group may appear more subjective (i.e., coherence/cohesion), and

their identification may reflect a sensitivity of the reviewer. Minimizing such risks is achieved by conducting the analysis as a group, and by engaging seasoned writing analysis experts.

While there were multiple strong and moderate correlations between linguistic features, there were very few correlations between scientific features, and only one scientific feature correlated moderately with three linguistic ones. This may imply that overall writing proficiency does not automatically assure the ease of mastering the rules of behavioral scientific writing in novices.

The uncovered correlations between the features may hold practical writing instruction implications. As previously mentioned, some of the identified features represent easily-recognizable linguistic constructs, like adverbial leads, or verb clusters. These may correlate with more abstract features, like voice. While it is important to note that a correlation may or may not signify a functional relationship, these ties may prove useful in teaching scientific writing. Instructing students in use of adverbial leads is much more straightforward than asking them to employ a consistent voice. Yet, based on the identified correlation, it may be possible to increase one's voice in scientific writing by increasing one's skill in writing sentences that contain adverbial leads. Similarly, if two features, one positive and one negative, demonstrate to be strongly inversely correlated, then instead of focusing on eradicating the negative one, an instructor may choose to try to foster consistent usage of the positive one instead. It is hard to direct novice writers to avoid under-developing their writing. It is easier to work with them to utilize a wide variety of sentence types, a feature negatively correlating with underdeveloped writing in present sample. Such approach may (assuming a functional relationship binds the two)

result in mastering a positive trait concurrently with eradicating a negative one. This is important, because error-free writing is not the same as great writing.

## RQ3 Results Discussion: Do Students' Prominent Features Scores Relate to Their College GPA, or ACT scores, Including Composite Score, and Language, Math, Science, and Reading Sub-scores?

As hypothesized**,** Prominent Feature Analysis scores positively significantly correlated with student college GPA scores, and their ACT scores. This provides support for validity of the scale, and reflects past research results on scale validity at elementary and middle school (Morse et al., 2007; Swain et al, 2010) as well as high school levels (Morse et al., unpublished).

Interestingly, negative linguistic prominent features scores exhibited strongest significant inverse correlations with GPA, suggesting that "good" college students tend to generate error-free writing. Additionally, ACT math sub-scores (not language or reading scores) exhibited the strongest significant correlations with most of the positive and, strongest inverse correlation, with most of the negative prominent features score categories.

To further understand why language-based ACT sub-scores may not have been the strongest to correlate with prominent features scores, I investigated the nature of ACT language and reading tasks. The language section contains five passages and 75 questions to be answered in 45 minutes. Each question contains four answer options. This section tests one's knowledge of usage and mechanics (punctuation, usage, grammar, sentence structure) and rhetorical skills (strategy, organization, style) (Edwards, 2015).

The ACT reading section contains five passages and 40 multiple-choice questions based on these passages, to be answered in 35 minutes; each question contains four answer options. Passages always cover domains of humanities, social studies, natural sciences and literary fiction. This task requires vocabulary knowledge, knowledge about the world at large, familiarity with English syntax and semantics to make sense of the presented ideas, understanding language conventions necessary for comprehending concepts like humor or sarcasm, reasoning ability to extract ideas presented implicitly rather than explicitly (Safier, 2015).

In case of both language-related ACT sections, the emphasis is on comprehending and manipulating the text, instead of generating it. While both sub-section scores correlated with prominent feature scores, these relationships may have been stronger when compared with scores on standardized task that required generating original text.

When comparing the correlation values of ACT scores (in language, math, reading, science, and composite) and prominent features scores (summed positive, negative, raw prominent features scores, and summed scientific and linguistic), I learned that ACT math correlated most strongly with most of prominent features scores. This relationship is unexpected, and may be explained by the overall proficiency of good students (good in math; good writers). ACT Math scores significantly correlated with ACT Language scores ($r = .78$, $p \leq .01$) in the present sample. This may suggest possible characteristics external to math and writing, yet helpful with both, like overall intelligence, or high capacity of working memory.

## RQ4 Results Discussion: Can Prominent Feature Analysis Scale Be Used to Assess Change in Student Writing When Two Samples from the Same Students Are Compared Across Time?

To investigate whether Prominent Feature Analysis can be used to track change in writing over prompts and/or over time, I analyzed prominent features scores generated by 25 students during two writing efforts, a short literature review and a long literature review, during two consecutive semesters. In comparison with shorter literature reviews, longer literature reviews exhibited significantly more positive prominent features, however, the number of negative prominent features increased significantly as well. This suggests that for novice writers in the present sample increase in passage length was both a blessing and a curse. It afforded more opportunities for exhibiting complex means of expression, as well as making more mistakes. The prominent features scores did not correlate significantly between the sets, suggesting that, for this small participant sample and these two prompts, Prominent Feature Analysis was not suitable for tracking change in writing ability over time and prompts.

This lack of correlation of prominent features between two writing samples penned by same group of students over time confirms results of a previous research study conducted on 222 students, and analyzing their papers written in ninth and tenth grade (Morse et al., unpublished). While the interrater agreement was very high (about 98% for both sets of papers), and the relative challenge level of features remained stable across the two samples, the median correlation of prominent features scores across time was only $r = .06$.

However, both studies (high-school and current one) featured confounds which may have affected these results. The prompts were not counterbalanced in the high-school Prominent Feature Analysis, introducing the possibility of the task effect. Similarly, in present study, the compared essays differed in length, as the first set was comprised of writing that was about three pages long, the second set contained papers that averaged 8-12 pages. Neither effort included targeted instruction to specifically remedy the problems identified in the first set, thus possibly reducing opportunities for growth in writers. Lastly, in the present study, students picked their own topics. This may have been to their benefit, if they were excited about the topic and capable of writing on it. However, it also may have resulted in students having to review and describe scientific literature which was beyond their understanding, thus lowering the quality of their writing.

## Limitations

Several limitations of this study are important to consider. While a clear majority of writing samples evaluated in the present study were generated in classes taught by the same Educational Psychology instructor, a small group of samples (29 empirical reports) was completed in three laboratories taught by Psychology teaching assistants. Despite a potential instructor effect, these samples were included in analysis for two reasons: 1) they were penned by students from a different department, thus increasing an overall number of independent writing samples needed to conduct Rasch analysis, and 2) they represent a second scientific writing type, different than literature reviews.

While the presence of experiment reports was overall beneficial to the study, their small number (29 of 208) may have affected Rasch item difficulty ranking of a few

features exclusively pertaining to this writing type (i.e., design rigor). Experiment report-specific features had a smaller chance of appearing due to overall lower number of experiment reports reviewed.

Past research on writing assessment (including research on prominent feature analysis) repeatedly indicates task effect as a confound (Van den Bergh, Rijlaarsdam, & Van Steendam, 2016). In the present study, task effect existed on two levels. First, students selected their own topic for all reviewed papers. In most cased, it probably helped to assure that the domain content was within their realm of interest and understanding. However, if a student picked a topic that proved too difficult for him or her, the quality of both writing mechanics and scientific thinking may have suffered. Second, paper requirements differed between classes. Students in EPY 3513 were expected to write short write literature reviews which did not exceed three pages (excluding references); students in EPY 4033 wrote longer literature reviews, often ranging between 10 and 12 pages (excluding references and abstract); students in PSY 3314 wrote experiment reports commonly ranging from 4-6 pages (excluding references). Paper lengths may relate to the quality of student writing; a comparison of mean numbers of prominent features noted in a small sample of short and long literature reviews penned by repeat students (RQ4) appears to confirm the presence of this limitation. Shorter literature reviews contained on average fewer positive features, and fewer negative features than their longer counterparts.

Gender imbalance within the present participant sample was significant; a clear majority of the students whose work I have reviewed are females, according to their Mississippi State University records (192 of 208). This is representative of the make-up

of the students in the departments from which the majority of the writing samples originated, Educational Psychology, and Psychology. However, while gender differences may reduce generalizability in scientific studies in general, my concern was minimal. Research on verbal skills does not support a claim for large and significant gender differences. A meta-analysis of 165 studies on gender similarities and differences highlighted a lack of differences between males and females (aged between 3-64 years old) in vocabulary, reading comprehension, or essay writing (Hyde, 2014).

Prominent Feature Analysis, by design, reflects the characteristics of a given writing sample only. The goal of the process is *not* a creation of a generalizable scale, but a thorough understanding of the writing at hand. Even so, a combination of factors suggest that the present set of prominent features may reflect novice scientific writing outside of junior and senior years of psychology and educational psychology. First, many graduate students in behavioral sciences take multiple research methods courses and other advanced classes in which they complete scientific manuscripts, but not a class devoted solely to APA-formatted writing. Depending on their past scientific writing experiences, graduate students may exhibit similar writing characteristics as undergraduates. Second, previous research on common mistakes in APA-formatted scientific writing submitted for publication to a professional journal suggests that many of the problems identified in current analyses persist past undergraduate and graduate education. Therefore, some aspects of the scale may reflect the struggles of behavioral scientific writers at large. However, it is advisable that anyone seeking to use the present scale in another study does so only as a starting point. While the process is valid and generalizable, the product may be less so.

## Implications for Instruction

As past research on writing feedback demonstrates, ample and specific feedback is invaluable for the writer during revision process. More specifically to present effort, short units of writing instruction targeting APA-established standards for scientific writing in behavioral sciences resulted in significant improvements in student writing performance in a general psychology course (Fallahi, Wood, Austad, & Fallahi, 2006).

Present analysis of novice behavioral scientific writing yielded identification of a total of 55 prominent characteristic, positive and negative. An instructor tasked with improving behavioral science novice writing may benefit from reviewing this list to see whether his or her students also seem to struggle with the negative features, and from noting how successfully the students incorporate the positive features. While, ideally, a full analysis is performed to fully understand any other sample of novice scientific writing, the present list can certainly be used as a guide and a starting point for designing targeted instruction.

Writing instructors can assist behavioral science faculty with designing interventions to eradicate the common mistakes, as well as, to increase the presence of the more sophisticated tools of expression. Ample research exists on strategies for teaching individual writing characteristics. For example, cumulative sentences, which embody rich and detailed means of sophisticated expression, so helpful for conveying complex scientific ideas, have been explicitly taught to elementary through high school students (Graves, Swain, & Morse, 2011). In the current writing sample, cumulative sentences significantly correlated (at moderate values) with both summed prominent features scores and positive prominent features scores, which may justify such

intervention. To teach multiple features at once, one can employ contrasting cases instruction, which focuses on a comparison of poorly-written material with well-written material, and applying the lessons learned to one's own writing. Past research results suggest that this approach is more beneficial than focusing on well-written examples only (Lin-Siegler, Shaenfield, & Elder, 2015).

To make instructor's feedback easier to understand for students, a rubric can be created to track presence and absence of the individual features, as well as to highlight examples of each. Such rubric may also prove to be an invaluable tool for self or peer assessment. Past research on peer evaluations suggests that this form of feedback is powerful and valuable for the author *and* the reviewer, as both benefit from deeply engaging with text. Peer review process may be even more powerful when the rubric is specifically designed for improving the APA-formatted novice scientific writing in psychology (Greenberg, 2015). As peer feedback in a writing classroom involves interactions of two novices working towards understanding and improving a writing sample, a fair amount of structure and facilitating of the process is needed; prominent feature rubric can catalyze the process.

## Recommendations for Future Research

While Prominent Feature Analysis results in a rich and detailed picture of characteristics of a given writing sample, its' ability to track writers' growth is still unknown. Past prominent feature research on high school students' writing resulted in virtually no correlation between two samples generated by repeat students over time. Current, small-scale investigation into novice scientific writing (attempted to answer RQ4) also resulted in no statistically significant correlation of prominent features scores

between two writing samples written by repeat students. However, neither of these research efforts was specifically designed to answer this question. Optimal study design would involve using multiple writing prompts, similar in scope and difficulty, counterbalanced (or randomized) among participants. To explicitly foster writing ability growth in participants, the study could include an intervention to improve writing delivered between the prompts, based on analysis of the participants' early writing effort.

It would also be informative to explore novice scientific writers' opinions about the quality and quantity of feedback generated with the current scale. Per previous literature, maximizing the impact of feedback would require clearly marking the presence of each feature in student papers, and adequately explaining to students what each feature name stands for. Lastly the feedback should contain examples of ways to avoid the negative, and to incorporate the positive features (Agius & Wilkinson, 2014).

Another potential line of investigation centers on evaluation of stability of features, and of feature difficulty level, across the samples investigated so far. This would be an interesting question to pursue, as it would illuminate which negative features tend to persist over time despite more and more schooling, and which positive features tend to be hard to manifest, despite continuous writing instruction throughout one's academic career.

Additionally, it would be interesting to compare the characteristic of novice scientific writers and scientific writing experts; an analysis of expert scientific writing may yield additional ideas for improving the novice efforts. Such a comparison could be accomplished by analyzing an adequate sample of recent, published scientific literature

from respected peer-reviewed publications, and comparing the emergent features with the present set. I hypothesize both additional positive and negative features would be noted.

## Conclusions

This study sought to extend the Prominent Feature Analysis scale into a new genre and academic level of writing, and to explore the new scale's psychometric properties. Assisted by three other writing assessment experts, I identified genre-specific prominent features in a representative sample of novice scientific writing, and analyzed the relationships between the features. I confirmed the validity of the scale by correlating the scores with student GPA and ACT scores. Lastly, I conducted an exploratory investigation of the task environment effect, or the potential of students displaying different writing features in two different writing samples (Van den Bergh et al., 2016).

The results suggest that Prominent Feature Analysis is a valid tool for illuminating the characteristics of novice behavioral scientific writing, extending the usefulness of the measure into new genre and academic level. As mastering behavioral scientific writing poses a significant challenge to novices, this tool may be very helpful to both students and instructors. As a behavioral scientific writing instructor, I will immediately put the results of present research to use by designing interventions to address the common mistakes and to train the students in usage of the more sophisticated linguistic tools and scientific thinking patterns.

While the present study answered some questions, others are still unanswered. It is not clear whether the tool can be used for tracking writing progress, how stable is the presence and order of features when compared between different writing samples, and

how well does the present set of features align with characteristics of scientific writing penned by experts.

REFERENCES

Agius, N. M., & Wilkinson, A. (2014). Students' and teachers' views of written feedback at undergraduate level: A literature review. *Nurse Education Today*, *34*, 552-559. doi:10.1016/j.nedt.2013.07.005

Bang, H. J. (2013). Reliability of National Writing Project's Analytic Writing Continuum assessment system. *Journal of Writing Assessment*, *6*, 13-24.

Bardine, B. A. (1999). Students' perceptions of written teacher comments: What do they say about how we respond to them? *The High School Journal*, *82*, 239-247.

Bardine, B. A., Bardine, M. S., & Deegan, E. F. (2000). Beyond the red pen: Clarifying our role in the response process. *The English Journal*, *90*(1), 94-101. doi:10.2307/821738

Bevan, R., Badge, J., Cann, A., Willmott, C., & Scott, J. (2008). Seeing eye-to-eye? Staff and student views on feedback. *Bioscience Education*, *12*(1), 1-15. doi:10.3108/beej.12.1

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*, 65-81.

Christensen, F. (1963). A generative rhetoric of the sentence. *College Composition and Communication*, *14*, 155-161. doi:10.2307/355051

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, C. & Lee, O. (1977). *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt, Brace, Jovanovich College Publishers.

Culham, R. (2003). *6 + 1 traits of writing: The complete guide*. New York, NY: Scholastic Inc.

Educational Testing Service. (2018). *Score level descriptions for the analytical writing measure.* Retrieved from https://www.ets.org/gre/revised_general/prepare/analytical_writing/score_level_descriptions/

Edwards, M. (2015, May 13). What's actually tested on the ACT English section? [Web log post]. Retrieved June 13, from https://blog.prepscholar.com

Englehard, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education, 5*, 171-191. doi:10.1207/s15324818ame0503_1

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32. doi:10.3102/0013189x01800902

Galbraith, D. (1999). Writing as a knowledge-constituting process. In G. Rijlaarsdam & E. Esperet (Eds.) *Studies in writing (Vol. 4) Knowing what to write: Conceptual processes in text production* (pp. 139-164). Amsterdam, Netherlands: Amsterdam University Press.

Graves, R., Swain, S., & Morse, D. (2011). The final free modifier–Once more. *Journal of Teaching Writing*, *26*(1), 85-105.

Greenberg, K. P. (2015). Rubric use in formative assessment: A detailed behavioral rubric helps students improve their scientific writing skills. *Teaching of Psychology*, *42*, 211-217. doi:10.1177/0098628315587618

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. Levy & S. Ransdell (Eds.) *The science of writing: Theories, methods, individual differences, and applications* (pp. 6-44). Mahwah, NJ: Lawrence Erlbaum Associates.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology*, *65*, 373-398. doi:10.1146/annurev-psych-010213-115057

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*, 201-213. doi:10.2307/358160

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549-566. doi:10.2307/358601

Huot, B., & O'Neill, P. (2009). An introduction to writing assessment theory and practice. In B. Huot & P. O'Neill (Eds.) *Assessing writing* (pp.1-9). Boston, MA: Bedford/St. Martin's.

Huot, B. (2003). *Rearticulating writing assessment for teaching and learning*. Boulder, CO: University Press of Colorado.

IBM Corp. (2016). IBM SPSS Statistics for Windows, Version 24.0 [computer software]. Armonk, NY: IBM Corp.

Krishna, V. (1975). The syntax of error. *Journal of Basic Writing*, *1*, 43-49.

LeMahieu, P. G., & Friedrich, L. (2007). Looking at student work to build an evaluative framework: Why and more important, how. In Davies, A., & Busick, K. U. (Eds.). *Classroom assessment: What's working in high schools*. *Book one.* Courtenay, BC: Building Connections Publishing.

Linacre, J. M. (2017). *A user's guide to WINSTEPS, MINISTEP Rasch-model computer programs*. *Program manual 4.0.0*. Retrieved from https://www.winsteps.com/winsteps.htm

Linacre, J. M. (2018). WINSTEPS [computer software]. Retrieved from https://www.winsteps.com/winsteps.htm

Lin-Siegler, X., Shaenfield, D., & Elder, A. D. (2015). Contrasting case instruction can improve self-assessment of writing. *Educational Technology Research and Development*, *63*, 517-537. doi:10.1007/s11423-015-9390-9

Lorenzo-Seva, U. & Ferrando, P. M. (2018) FACTOR [computer software]. Retrieved from http://psico.fcep.urv.es/utilitats/factor/Download.html

Lucas, G. M. (2010). Initiating student-teacher contact via personalized responses to one-minute papers. *College Teaching*, *58*, 39-42. doi:10.1080/87567550903245631

Lyon, C. (2016). Using audio feedback to facilitate student revising. *Journal of Teaching Writing*, *31*(2), 49-67.

Mansourizadeh, K., & Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *Journal of English for Academic Purposes*, *10*, 152-161. doi:10.1016/j.jeap.2011.03.004

Morse, D. T., Swain, S. S, & Graves, R. L. (November, 2007). *Scaling writing proficiency via prominent features of essays.* Paper presented at the annual meeting of Mid-South Educational Research Association, Hot Springs, AR.

Morse, D.T., Swain, S.S, & Graves, R. L. (unpublished). Additional research on scaling writing proficiency via prominent features of essays. 1-14.

National Writing Project. (nd). *National Writing Project offers high-quality writing assessment services.* Retrieved from https://www.nwp.org/cs/public/print/resource/3776

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*, 199-218. doi:10.1080/03075070600572090

Onwuegbuzie, A. J., Frels, R. K., & Slate, J. R. (2010). Evidence-based guidelines for avoiding the most prevalent and serious APA error in journal article submissions-the citation error. *Research in the Schools*, *17*(2), ix-xxxvi.

Orrell, J. (2006). Feedback on learning achievement: Rhetoric and reality. *Teaching in Higher Education*, *11*, 441-456. doi:10.1080/13562510600874235

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quart*erly, *14*, 61-69. doi:10.2307/3586809

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research (Expanded edition: 1980. Chicago: University of Chicago Press).

Robinson, S., Pope, D., & Holyoak, L. (2013). Can we meet their expectations? Experiences and perceptions of feedback in first year undergraduate students. *Assessment & Evaluation in Higher Education*, *38*, 260-272. doi:10.1080/02602938.2011.629291

Safier, R. (2015, May 2). What's actually tested on the ACT Reading section? Skills you need [Web log post]. Retrieved June 13, 2018, from https://blog.prepscholar.com

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.). *Advances in applied psycholinguistics: Disorders of first-language development, reading, writing and language learning (Vol. 2)* (pp. 142-175). Cambridge, UK: Cambridge University Press.

Shah, J., Shah, A., & Pietrobon, R. (2009). Scientific writing of novice researchers: What difficulties and encouragements do they encounter? *Academic Medicine*, *84*, 511-516. doi:10.1097/acm.0b013e31819a8c3c

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of writing tests. *The Modern Language Journal, 76*, 27-33. doi:10.1111/j.1540-4781.1992.tb02574.x

Sommers, N. (2006). Across the drafts. *College Composition and Communication*, *58*, 248-257. doi:10.2307/357362

Starch, D., & Elliott, E. C. (1912). Reliability of the grading of high-school work in English. *The School Review*, *20*, 442-457. doi:10.1086/435971

Starch, D., & Elliott, E. C. (1913a). Reliability of grading work in history. *The School Review*, *21*, 676-681. doi:10.1086/436185

Starch, D., & Elliott, E. C. (1913b). Reliability of grading work in mathematics. *The School Review*, *21*, 254-259. doi:10.1086/436086

Stephens, P. (2008, November). Rubrics: The heart of assessment. *SchoolArtsOnline.com, 108.* Retrieved from https://www.davisart.com/Promotions/SchoolArts/PDF/11_08Rubrics.pdf

Stock, P. L., & Robinson, J. L. (1987). Taking on testing: Teachers as tester-researchers. *English Education*, *19*, 93-121.

Swain, S. S., Graves, R. L., & Morse, D. (2006). *The effect of Mississippi Writing/Thinking Institute professional development on the writing achievement of ninth-graders*. Berkeley, CA: National Writing Project. Retrieved from National Writing Project website: http://www.nwp.org/cs/public/download/nwp_file/10563/Mississippi_Writing_Thinking_Institute.Pdf

Swain, S. S., Graves, R. L., & Morse, D. T. (2010). Prominent Feature Analysis: What it means for the classroom. *English Journal*, *99*, 84-89.

Swain, S. S., Graves, R. L., Morse, D. T., & Patterson, K. J. (2012) Prominent feature analysis: Linking assessment and instruction. In C. Bazerman et al. (Eds.), *International advance in writing research: Cultures, places, measures* (pp. 152-166). Anderson, SC: Parlor Press.

Swain, S. S., Graves, R. L., & Morse, D. T. (2015). The emerging shape of voice. *English Journal*, *104*(5), 30-36.

Swain, S. S., Graves, R. L., & Morse, D. T. (unpublished). A syntactic and prominent feature analysis of seventh-grade writing. 1-74.

Swain, S. S., & LeMahieu, P. (2012). Assessment in a culture of inquiry: The story of the National Writing Project's analytic writing continuum. In N. Elliot & L. Perelman (Eds.) *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 45-66). New York, NY: Hampton Press.

Van den Bergh, H., Rijlaarsdam, G., & Van Steendam, E. (2016). Writing process theory: A functional dynamic approach. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed.) (pp. 57-71). New York, NY: The Guilford Press.

VandenBos, G. R. (Ed). (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Vygotsky, L. (1986) *Thought and language*. Cambridge, MA: MIT Press.

White, E. M. (1994). Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance. Revised and expanded. San Francisco, CA: Jossey-Bass Publishers.

White, E. M. (2009). Holisticism. In B. Huot & P. O'Neill (Eds.) Assessing writing (pp. 19-28). Boston, MA: Bedford/St. Martin's.

Willingham, D. B. (1990). Effective feedback on written assignments. Teaching of Psychology, 17, 10-13. doi:10.1207/s15328023top1701_2

Wright, B. D. (1977). Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-116. doi:10.1111/j.1745-3984.1977.tb00031.x

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. College Composition and Communication, 50, 483-503. doi:10.2307/358862

# APPENDIX A

# HOLISTIC RATING SCHEME EXAMPLE

The following text presents holistic rating levels assigned to the Graduate Record Examinations (GRE) writing samples. This information has been retrieved, and is directly cited from:

https://www.ets.org/gre/revised_general/prepare/analytical_writing/score_level_descriptions/

**Score Level Descriptions for the Analytical Writing Measure**

Although the *GRE*® Analytical Writing measure contains two discrete analytical writing tasks, a single combined score is reported because it is more reliable than a score for either task alone. The reported score ranges from 0 to 6, in half-point increments

The statements below describe, for each score level, the overall quality of analytical writing demonstrated across both the Issue and Argument tasks. The test assesses "analytical writing," so critical thinking skills (the ability to reason, assemble evidence to develop a position and communicate complex ideas) are assessed along with the writer's control of grammar and the mechanics of writing (e.g., spelling)

**Scores 6 and 5.5**

Sustains insightful, in-depth analysis of complex ideas; develops and supports main points with logically compelling reasons and/or highly persuasive examples; is well focused and well organized; skillfully uses sentence variety and precise vocabulary to convey meaning effectively; demonstrates superior facility with sentence structure and language usage, but may have minor errors that do not interfere with meaning.

**Scores 5 and 4.5**

Provides generally thoughtful analysis of complex ideas; develops and supports main points with logically sound reasons and/or well-chosen examples; is generally

focused and well organized; uses sentence variety and vocabulary to convey meaning clearly; demonstrates good control of sentence structure and language usage, but may have minor errors that do not interfere with meaning.

**Scores 4 and 3.5**

Provides competent analysis of ideas; develops and supports main points with relevant reasons and/or examples; is adequately organized; conveys meaning with reasonable clarity; demonstrates satisfactory control of sentence structure and language usage, but may have some errors that affect clarity.

**Scores 3 and 2.5**

Displays some competence in analytical writing, although the writing is flawed in at least one of the following ways: limited analysis or development; weak organization; weak control of sentence structure or language usage, with errors that often result in vagueness or lack of clarity.

**Scores 2 and 1.5**

Displays serious weaknesses in analytical writing. The writing is seriously flawed in at least one of the following ways: serious lack of analysis or development; lack of organization; serious and frequent problems in sentence structure or language usage, with errors that obscure meaning.

**Scores 1 and 0.5**

Displays fundamental deficiencies in analytical writing. The writing is fundamentally flawed in at least one of the following ways: content that is extremely confusing or mostly irrelevant to the assigned tasks; little or no development; severe and pervasive errors that result in incoherence.

**Score Level 0**

The examinee's analytical writing skills cannot be evaluated because the responses do not address any part of the assigned tasks, are merely attempts to copy the assignments, are in a foreign language or display only indecipherable text.

**Score NS**

The examinee produced no text whatsoever.

# APPENDIX B

# ANALYTIC RATING SCHEME EXAMPLE

# Analytic Rubric Sample

From "Rubrics: The Heart of Assessment"

***Pam Stephens***

**Objective**

After defining the concept of art criticism, each student will write a 500-word essay with a beginning, middle, and end that accurately describes, analyzes, interprets, and judges a selected work of art.

| | Value 3 | Value 2 | Value 1 | Score |
|---|---|---|---|---|
| **Describe** | Provides a complete and accurate description of the key subject matter and elements seen in the artwork. | Provides a partial but mostly accurate description of the subject matter and/or elements seen in the artwork; some key components overlooked. | Provides an incomplete, unclear, or inaccurate description of subject matter and/or elements seen in the artwork; many key components overlooked. | |
| **Analyze** | Accurately relates how the structures of art function together to make a complete composition. | Relates with limited proficiency how the structures of art function together to make a complete composition; overlooks some important components. | Has trouble relating how the structures of art function together to make a complete composition. | |
| **Interpret** | Suggests a logical and/or symbolic meaning expressed in a work of art; supports idea with multiple points of visual evidence found in the piece. | Suggests a literal meaning expressed in a work of art; supports idea with limited points of visual evidence found in the piece. | Finds it difficult to interpret the meaning of the work; guesses meaning without visual support. | |
| **Evaluate** | Uses multiple criteria to judge the quality of a finished work of art; avoids personal opinion. | Uses a limited range of criteria to judge the quality of a work of art; personal opinion shown. | Uses personal opinion to judge the quality of a finished work of art. | |
| **Technical** | Finished paper follows rules of grammar and essay writing; is in publishable form. | Finished paper contains minor flaws in grammar and essay writing; needs editing. | Finished paper has numerous flaws in grammar and does not follow conventions of essay writing; needs rewriting. | |
| **Notes to student** | | | | **Total score** |

Rubric included with author's permission; downloaded from:

https://www.davisart.com/Promotions/SchoolArts/PDF/11_08RubricSample.pdf

# APPENDIX C

## PREVIOUS PROMINENT FEATURE ANALYSIS EXAMPLE

The following features were identified in a seventh-grade writing sample. See Swain, Graves, & Morse (2010) for more information regarding this study.

Table C1

*Prominent Features Identified in a Seventh-Grade Writing Sample; N = 464*

| Feature |
|---|
| Hyperbole |
| Aside to reader |
| Alliteration |
| Metaphor |
| Sensory language |
| Noun cluster |
| Absolute |
| Narrative storytelling |
| Subordinate clause |
| Effective repetition |
| Striking words |
| Verb cluster |
| Cumulative sentence |
| Vivid verbs/nouns |
| Voice |
| Balance/parallelism |
| Coherence/cohesion |
| Effective organization |
| Transitions |
| Adverbial leads |
| Sentence variety |
| Elaborated details |
| **Usage problems (-)** |
| **Weak structural core (-)** |
| **List technique (-)** |
| **Faulty spelling (-)** |
| **Faulty punctuation (-)** |
| **Weak organization (-)** |
| **Redundancy (-)** |
| **Shifting point of view (-)** |
| **Garbles (-)** |
| **Illegible handwriting (-)** |

*Notes:* Negative features are bolded and marked with "(-)".

# APPENDIX D

# MISSISSIPPI STATE UNIVERSITY INTERNAL REVIEW BOARD'S NOTICE OF APPROVAL FOR HUMAN RESEARCH

## NOTICE OF APPROVAL FOR HUMAN RESEARCH

| | |
|---|---|
| **DATE:** | September 07, 2017 |
| **TO:** | David Morse, PhD, Counseling Ed Psyc & Foundations |
| **FROM:** | Nicole Cobb, HRPP Officer, MSU HRPP |
| **PROTOCOL TITLE:** | Scaling Undergraduate Scientific Writing Via Prominent Feature Analysis |
| **FUNDING SOURCE:** | NONE |
| **PROTOCOL NUMBER:** | IRB-17-393 |

EXEMPTION DETERMINATION

This letter is your record of the Human Research Protection Program (HRPP) approval of this study as exempt.

On September 07, 2017, the Mississippi State University Human Research Protection Program approved this study as exempt from federal regulations pertaining to the protection of human research participants. The application qualified for exempt review under CFR 46.101(b)(1, 4).

Exempt studies are subject to the ethical principles articulated in the Belmont Report, found at www.hhs.gov/ohrp/regulations-and-policy/belmont-report/#

If you propose to modify your study, you must receive approval from the HRPP prior to implementing any changes. The HRPP may review the exempt status at that time and request an amendment to your application as non-exempt research.

In order to protect the confidentiality of research participants, we encourage you to destroy private information which can be linked to the identities of individuals as soon as it is reasonable to do so.

The MSU IRB approval for this project will expire on September 06, 2022. If you expect your project to continue beyond this date, you must submit an application for renewal of this HRPP approval. HRPP approval must be maintained for the entire term of your project. Please notify the HRPP when your study is complete. Upon notification, we will close our files pertaining to your study.

If you have any questions relating to the protection of human research participants, please contact the HRPP by phone at 325.3994 or email irb@research.msstate.edu. We wish you success in carrying out your research project.


Jodilyn Roberts

---

| | |
|---|---|
| **Review Type:** | EXEMPT |
| **IRB Number:** | IORG0000467 |

# APPENDIX E

# DEFINITIONS AND EXAMPLES OF PROMINENT FEATURES IDENTIFIED IN CURRENT STUDY

The following prominent features emerged from the present analysis of 208 undergraduate novice scientific writing samples. Some of the definitions of previously-defined features are cited from an unpublished paper by Swain, Graves, and Morse, with authors' permission. Some previously-defined prominent feature definitions have been revised to reflect their presence in the new genre (novice scientific writing). Others have been identified and defined for the first time.

Two previously-identified prominent features (aside to the reader and hyperbole) were not noted in the current dataset. They are included in the scale, because they technically could occur in scientific writing.

**Positive Linguistic Prominent Features**

*Elaborated details*—use of vivid, appropriate, or striking details; goes beyond a listing of details. Example: "For one, eye contact—a common behavior in normative social interaction—was less likely to happen while interacting with the animal."

*Sensory language*—language addressing the six senses, including direct quotations. Example: "When the participants steadily raised their affected arms, the feedback volume increased. When the participants moved their affected arms faster, feedback volume was produced at the higher rate."

*Metaphors*—all types of metaphoric language (metaphor, simile, etc.); especially noted is the use of common words used in metaphoric ways. Example: "Bilingual children absorb their surroundings like sponges (…)."

*Alliteration*—effective repetition of sound in successive words. Example: "Attention Deficit Hyperactivity Disorder (ADHD) is a chronic condition characterized by distractibility, hyperactivity, and impulsive behavior (Boot, Nevicka, & Baas, 2017)."

*Vivid nouns/verbs*—uncommon diction, very appropriate and descriptive. Example: "The results demonstrate the added value of the dynamic tests in forecasting reading development and predicting responsiveness to reading interventions."

*Hyperbole*—exaggeration. (Not noted in the present writing sample.)

*Striking words*—striking word usage, including appropriate or surprising nouns, verbs, adjectives, adverbs, etc. Example: "The most common characteristics of Apert syndrome appear as a triad: underdevelopment of the midface, craniosynostosis, and symmetrical syndactyly (or fusion) of the digits of the hands and feet."

*Cumulative Sentence*—a sentence with a base clause and one or more free modifiers. Example: "Data analysis did not support the authors' hypothesis in that those who played the active video games did expend more energy, but did not increase food intake, suggesting that active gameplay may be a better option for adolescents when considering energy balance."

*Verb cluster*— type of free modifier (-ing or -ed participle). Example: "These researchers continued their study, the KiVa program, beyond a questionnaire by training and implementing teacher-let interventions for nine months, focusing on bystander intervention (Juvonen & Schacter)."

*Noun Cluster*—type of free modifier; a noun, possibly with attachments. Example: "The result can be an unsuspecting adoptee who feels he/she is missing part of his/her genetic identity, and longs to have hold of this information."

*Absolute*—type of free modifier; an independent noun with its own verb and deleted auxiliary verb. Example: "Ironically, hunger is present within the United States of America, a country known as a 'superpower.'"

*Adverbial lead*—beginning the sentence with adverbial (word, phrase, or clause). Example: "Much like in the previous study, Santen, Sproat, and Hill (2013) tested for echolalia rates (repetitive speech) among children with ASD, communication disorders, and normal children."

*Balance/parallelism*—all types of parallel construction. Example: "Gender-neutral parenting is a type of parenting style that treats boys and girls equally; for example, boys can play with monster trucks, and so can girls. Boys can play with Barbie dolls just as girls can."

*Effective repetition*—repeating the same word, or a form of it, effectively (also includes repetition of phrases, or construction). Example: "Children in stables homes are much more resilient, and resilient children have an ability to form secure attachment (…)."

*Subordinate sequence*—an organizational pattern that follows a "detail of a detail of a detail" order. Example: "(…) In the beginning, the deaf children felt they did not fit in with their hearing peers and did little to interact with them. Not until the hearing students learned to sign and began communicating with their deaf classmates did the deaf students began to feel comfortable in the classroom, and their self-esteem began to rise (Kreimeyer, et al., 2000).

*Coordinate sequence*—an organizational pattern that contains clauses connected by one of the following: for; and; nor; but; or; yet; so. Example: "Problems with social interaction and peer acceptance could lead to depression and other mental health issues throughout adolescence for AS individuals (Elst, et al., 2013)."

*Sentence variety*—effective use of a variety of sentence forms and lengths. Example: "To confirm the interest in filling an unknown family history, research suggests that genetic testing in a direct-to-consumer format has increased in popularity among adoptees. Their motivation? Gathering any bit of knowledge that was left unknown during the adoption process (Baptista et al., 2016)."

*Transitions*—the presence of key function words or phrases to enhance organization. Example: "In comparison to Swanson et al. (2014), Johnson and Lieberman (2007)…"; "In contrast, …"

*Coherence/cohesion*—obvious presence of cohesive devices throughout the writing to create cohesion or coherence. In scientific writing, may take a form of well-structured headings, among others.

*Voice*—the presence of an original, personal or authentic conception of the subject. Example: "There have been many accommodations made for deaf students so that they can attend general schools with their hearing peers. Items such as hearing aids, cochlear implants, as well as a special glove that converts signs into language. But are these technologies really accommodating the deaf? Or are they accommodating the hearing?"

*Narrative Storytelling*—including event sequences and anecdotes to develop ideas. Example: "Imagine a mother at home alone bonding with her newborn. She receives an unexpected call, and is bluntly informed that her child has sickle cell. Sickle cell is as an incurable, inherited, recessive gene, blood disorder in which an individual has irregular shaped oxygen deprived blood cells, periodic pain crisis, severe anemia,

increased possibility of getting infections, having a stroke, and acute chest syndrome (Nemours Foundation, 2015).

Fast forward three years. The mother is excited about her child's first day of school. Out of nowhere the child spikes a fever, which is always to be treated as an emergency (Nemours Foundation, 2015). The ER doctor diagnoses the child with walking pneumonia, gives prescriptions, and sends the child home. By the next day, the child's eyes have turned yellow and he is screaming in pain. They return to the local ER, from which he is sent by ambulance to the state's children's hospital."

*Aside to reader*—direct communication with the audience. (Not present in current writing sample)

*Diction*—effective choice of words and phrases. Example: "Mrug et al. (2014) interviewed the girls and their parents to reveal characteristics of delinquency (…)."

*Effective organization*—clear pattern of organization of writing throughout the paper.

**Negative Linguistic Prominent Features**

*Usage problems*—occurrences of nonstandard, social, regional, or ethnic dialect features. Example: "Based off of the data observed, the working memory has less necessity for meaning and processing as the long-term memory."

*Weak structural core*—sentences that are "derailed" with misplaced awkward elements; also includes sentence fragments. Example: "Juries should understand the fallibility in witness accounts, while being made aware of the influences of outside factors like emotions and human error can contribute to the case."

*Garble*—unintelligible sentence. Example: "There was implication in this experiment such as a student being in school that could show no differences in reading notes or making a mental picture of notes while studying."

*Weak organization*—obvious lack of organization throughout the paper.

*Redundancy*—repeating the same idea or concept over and over, sometimes described as "verbiage" or "mindless filler." Example: "Researchers observed 263 students by observing participants for 15 minutes."

*List technique*—list of ideas related to a topic but not to each other. In present writing sample, this most commonly manifested as a list of study descriptions with no transitions Example: "Smith (2016) investigated (…). Jones (2014) looked at (…). Onslow (1994) examined (…)."

*Faulty punctuation*—a persistent pattern of any/all varieties of punctuation errors.

*Faulty spelling*—a persistent pattern of faulty spelling.

*Shifting point of view*—abrupt changes in the writer's point of view or subject. Example: "For some mother creating this bond may be harder or it feels like it does not come as natural as others, and this is where the postpartum depression may come into play."

*Underdeveloped*—presence of ideas that are introduced, but not fully explained or connected with others. Commonly occurred in discussion sections. Example: "Several studies indicate that creativity is higher in individuals who have been diagnosed with ADHD, while other studies have not found any statistically significant correlation between the two. The conclusions drawn from each study differ depending on the test used to measure creativity, and the varying degree of symptoms associated with an

individual's diagnosis. To more effectively consolidate the evidence supporting greater creativity in individuals diagnosed with ADHD, further research must be conducted on the adverse effects that ADHD medication has on the creative capabilities of the individuals."

*Misplaced modifier*—a phrase or clause occurring in a sentence in such place, that it appears to refer to a word different than intended. Example: "These four therapy sessions were conducted to focus on how couples should behave for the betterment of their baby and their development."

**Positive Scientific Prominent Features**

*Design rigor*—Explicit mention of methods to reduce limitations of the study, including controlling for practice effect, instructor effect, or randomized group assignment, among others. Example: "The researcher randomized the presentation of stimuli among participants to control for the sequence effect."

*Analysis rigor*—identification of a common methodology flaw in a reviewed study (school-based suicide prevention programs were reported as successful, yet study authors did not report pre- and post-intervention suicide rates). Example: "Although two of the studies that researched a specific school program clearly improved attitudes and knowledge, they failed to demonstrate a link between the prevention program and lower suicide rates."

*Well-blended sources*—effectively using multiple scientific sources to support a point or a series of points. Example: "While Durand et al. (2016) believes the onset of puberty is inherited, Davis et al. (2015) and Karaolis-Danckert et al. (2015) believe early life exposures can induce a child's pubertal timing."

*Exhibit of DV task*—including an example of the item used during an empirical study, for example, an image of mazes participants were expected to complete while listening to music with various beat frequencies.

**Negative Scientific Prominent Features**

*Design flaw*—description of a study implies scientifically-flawed study premise or procedure sequence. Example: a student writer investigated recall accuracy differences between words and numbers by testing each condition on a different group of participants. To investigate whether words or numbers are easier to recall, both conditions should be tested on all participants.

*Lack of examples*—study procedure lacks examples to be fully understood. Example: "Hamidah, et al. (2015) carried out a study in which they took 64 individuals in college and put them in exercise programs. The researchers then measured the anxiety responses in the subjects." Neither the exercise programs nor the anxiety measures are specified in the description.

*Procedural omissions/ambiguities*—description of the study does not sufficiently explain the process of gathering data. Example: "Solis et al. (2016) investigated different designs to help with reading comprehension. Anaphoric cueing focuses on using context clues and question development to further explore the reading material. Results showed that by doing these specific treatments randomly a few times a week for two weeks, the ABA therapy was significantly different than not using therapy."

*Required scientific elements missing*—writing sample does not contain an element/elements required for a given genre, for example a hypothesis, a standard heading, or a paper section (i.e., paper title, or discussion).

*Data/analysis misinterpretation*—incorrect conclusions drawn from presented results of a statistical analysis. Example: "The one-way ANOVA yielded a significant effect between the color of the words, and the recall $F(10) = 1.38$, $p = 0.24$."

*Undefined terms/abbreviations*—use of scientific terminology or abbreviations without including their definitions. Example: "As long as a series of letters can form a word that is phonologically congruent, it can be stored, manipulated, thus recalled, through working memory."

*Attribution errors*—lacking, mistaken, or incomplete attributions of ideas of others, including missing in-text references, incorrect citation format, among others.

*Incorrect placement of scientific information*—misplacement of genre-specific elements. Example: including elements of discussion in results section.

*References errors*—format or content mistakes in listing of the works cited in the paper (in references section).

*Extrapolating beyond data/Faulty logic*—statements that, based on cited research, cannot be confirmed true, or that are logically unsound. Example: "Everyone can benefit from counseling."

*Misuse of terms*—incorrect use of scientific terminology. Example: using the word "experiment" interchangeably with "study;" stating that results of behavioral science studies on people "prove" the hypothesis to be true. Example: "However, it is proven that an animal can positively impact mood, stress and anxiety."

*Analysis error*—use of an incorrect statistical test to answer the research question. Example: In a study investigating between-group differences, a paired samples *t*-test was used to conduct the statistical analysis.

*Inappropriate Personification*—the writer, through sentence construction, implies that inanimate entities (typically, research) have human-like attributes. "present study seeks to determine;" "images have a better recall than memories do;" "[the] study looked at."

*Excessive use of passive voice/construction*—typically present when describing research by others, or method/results of one's own study. "Seven different PTSD symptoms were studied… The PTSD checklist was used… In addition, a study was conducted that focused on PTSD and the effects on female military personnel… Data were collected… It can be concluded that PTSD is a significant mental health disorder… It can be understood that both men and women suffer from the disorder."

*Hypothesis/thesis incongruent with reviewed literature*—Selection of reviewed literature does not reflect the topic of the paper, or a hypothesis/thesis of the paper is not supported by the reviewed literature. Example: a paper titled "Methods of treatment for children for dyslexia" contains a seemingly-random summary of a study comparing stress levels in parents of children with dyslexia with stress levels of parents whose children do not have dyslexia.

*Statistical reporting error*—an incorrect or incomplete information cited based on statistical output. Example: "It took significantly longer for participants to complete the number search than the word search, $t(22) = -7.891$, $p = 0000000741$."

# APPENDIX F

# CORRELATIONS AMONG IDENTIFIED PROMINENT FEATURES

Table F1

*Correlations Among Identified Prominent Features*

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Elaborated details | 1 | | | | | | | |
| 2 | Sensory language | .22** | 1 | | | | | | |
| 3 | Metaphors | .08 | .13 | 1 | | | | | |
| 4 | Alliteration | .14* | .20** | .22** | 1 | | | | |
| 5 | Vivid verbs and nouns | **.38**** | **.30**** | .20** | .24** | 1 | | | |
| 6 | Striking words | .06 | .12* | .19** | .14 | .26** | 1 | | |
| 7 | Diction | .15* | .05 | -.01 | -.06 | **.31**** | .19** | 1 | |
| 8 | Cumulative sentence | .08 | .21** | .11* | .06 | .23** | .05 | .27** | 1 |
| 9 | Verb clusters | -.01 | .22** | .09 | .06 | .18* | .11 | .21** | **.62**** |
| 10 | Noun clusters | .11 | .26** | .13 | -.02 | .23** | -.05 | .19** | .28** |
| 11 | Absolutes | .00 | .11 | .13 | -.02 | .17* | .06 | .12 | .21** |
| 12 | Adverbial leads | .28** | .19** | .23** | .08 | **.33**** | .26** | **.36**** | .27** |
| 13 | Balance/parallelism | .28** | .08 | .11 | .12 | **.36**** | .18** | .26** | .19** |
| 14 | Effective repetition | .12 | .07 | -.04 | -.03 | .23** | .02 | .26** | .23** |
| 15 | Sentence variety | **.38**** | .17* | .14* | .06 | **.34**** | .26** | **.30**** | .27** |
| 16 | Effective organization | **.50**** | .03 | .12 | .17* | **.35**** | .15* | **.34**** | .13 |
| 17 | Subordinate sequence | .03 | -.04 | -.03 | -.02 | .02 | .05 | .16* | .12 |
| 18 | Coordinate sequence | .01 | -.04 | -.04 | -.03 | -.05 | .14* | .14* | -.09 |
| 19 | Transitions | .29** | .14* | .05 | .11* | **.36**** | .26** | **.32**** | .14* |
| 20 | Coherence /cohesion | **.34**** | .09 | .12 | .17* | **.31**** | .28** | **.43**** | .16* |
| 21 | Voice | **.34**** | .09 | .18** | .13 | **.40**** | .24** | **.45**** | .11 |
| 22 | Narrative storytelling | .11 | .26** | .13 | .20** | .23** | .17* | .05 | .21** |
| 23 | Design rigor | .02 | -.02 | -.02 | -.01 | .07 | .18* | .21** | .09 |
| 24 | Analysis rigor | .17* | .22** | .05 | .10 | .25** | .06 | .07 | .14* |
| 25 | Exhibit of DV task | .08 | -.01 | -.01 | -.01 | .14* | -.02 | .15* | .16* |
| 26 | Well-blended sources | .30** | .17* | .20** | .04 | **.48**** | .22** | .20** | .24** |
| 27 | Usage problems | .18** | .17* | .02 | .19** | .12 | .05 | .06 | .01 |
| 28 | Misplaced modifier | .04 | .05 | .05 | .03 | .04 | 0.07 | .03 | -.04 |
| 29 | Weak structural core | .15* | .10 | .01 | .13 | .22** | .16* | .22** | -.03 |
| 30 | Garbles | .21** | .01 | .00 | .05 | .09 | .11 | .17* | .04 |
| 31 | Weak organization | .28** | .09 | .09 | .06 | .16* | .08 | .17* | .08 |
| 32 | Redundancy | .12 | .09 | -.07 | .06 | .08 | .12 | .09 | .00 |
| 33 | List technique | .25** | .10 | .09 | .07 | .22** | .09 | .20** | .08 |
| 34 | Faulty punctuation | 0.00 | .02 | -.12 | .00 | -.02 | .10 | .09 | -.12 |
| 35 | Faulty spelling | .15* | -.03 | -.05 | .06 | -.04 | .09 | .16* | .04 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| 36 | Shifting point of view | .05 | -.05 | .05 | .03 | .05 | .00 | .08 | .02 |
| 37 | Underdeveloped | **.31**$^{**}$ | .15$^{*}$ | .08 | -.03 | .21$^{**}$ | .17$^{*}$ | .28$^{**}$ | .09 |
| 38 | Design flaw | -.09 | .03 | -.18$^{**}$ | .02 | .07 | -.11 | .07 | .06 |
| 39 | Thesis incongruent w/ lit. | .05 | .03 | .02 | .02 | -.01 | -.11 | -.11 | .06 |
| 40 | Statistical reporting error | .00 | .05 | -.08 | .03 | .13 | -.02 | -.04 | .05 |
| 41 | Anal. /stats misinterpret. | .02 | .06 | -.04 | -.10 | .05 | -.05 | -.01 | .05 |
| 42 | Wrong analysis | -.02 | .02 | .02 | .01 | .05 | .03 | .05 | .04 |
| 43 | Extrapolating beyond data | .14$^{*}$ | -.01 | -.03 | .07 | .01 | .11 | .00 | -.10 |
| 44 | Procedural omissions | .20$^{**}$ | .04 | .01 | -.03 | .05 | .00 | .01 | -.11 |
| 45 | Req. sci. elements missing | .07 | .06 | -.02 | .03 | .15$^{*}$ | -.04 | .13 | .13 |
| 46 | Lack of examples | .14$^{*}$ | .15$^{*}$ | -.04 | .10 | .16$^{*}$ | .08 | .12 | -.01 |
| 47 | Attribution errors | .07 | .02 | .04 | .03 | .02 | -.05 | .12 | .01 |
| 48 | References errors | -.05 | .01 | .15$^{*}$ | .11 | .00 | .03 | .06 | .02 |
| 49 | Excessive passive voice | .06 | .07 | .00 | .08 | .00 | -.03 | .01 | -.06 |
| 50 | Undefined terms | .03 | .02 | -.01 | -.05 | -.06 | .13 | .08 | -.14 |
| 51 | Wrong plcmt of sci. info. | .11 | .03 | .03 | .02 | .03 | .05 | .02 | .08 |
| 52 | Inappropriate personific. | .04 | .08 | -.05 | -.04 | .02 | -.09 | -.09 | -.16$^{*}$ |
| 53 | Misuse of terms/jargon | .16$^{*}$ | .03 | .01 | .09 | .05 | .04 | .06 | .04 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| 9 | Verb clusters | 1 | | | | | | | |
| 10 | Noun clusters | .22** | 1 | | | | | | |
| 11 | Absolutes | .29** | .11 | 1 | | | | | |
| 12 | Adverbial leads | .20** | .19** | .14 | 1 | | | | |
| 13 | Balance/parallelism | .14* | .02 | .02 | .26** | 1 | | | |
| 14 | Effective repetition | .18** | .07 | -.05 | .11 | **.35**** | 1 | | |
| 15 | Sentence variety | .21** | .11 | .17* | **.39**** | **.36**** | .25** | 1 | |
| 16 | Effective organization | .05 | .09 | .09 | .24** | **.36**** | .21** | **.43**** | 1 |
| 17 | Subordinate sequence | .26** | .10 | -.04 | .06 | .07 | -.05 | .08 | .06 |
| 18 | Coordinate sequence | .24** | -.04 | -.04 | .03 | -.06 | -.05 | .06 | -.01 |
| 19 | Transitions | .12 | -.03 | .03 | **.34**** | .29** | .07 | **.34**** | **.38**** |
| 20 | Coherence /cohesion | .20** | .09 | .09 | .24** | **.35**** | .26** | **.45**** | **.51**** |
| 21 | Voice | .05 | .15* | .09 | **.37**** | **.34**** | .26** | **.34**** | **.44**** |
| 22 | Narrative storytelling | .22** | .11 | .11 | .14 | .15* | .07 | .17* | .09 |
| 23 | Design rigor | -.04 | -.02 | -.02 | .03 | .06 | .19** | .12 | .14 |
| 24 | Analysis rigor | .19** | .03 | .03 | .11 | .06 | - .01 | .10 | .05 |
| 25 | Exhibit of DV task | -.03 | -.01 | -.01 | -.05 | .13 | .28** | .08 | .10 |
| 26 | Well-blended sources | .15* | .17* | .17* | .24** | .26** | .18** | **.30**** | .29** |
| 27 | Usage problems | -.07 | .00 | .00 | -.04 | .07 | -.02 | .17* | .25** |
| 28 | Misplaced modifier | -.15* | .05 | -.16* | .04 | -.03 | -.10 | .03 | .07 |
| 29 | Weak structural core | -.02 | -.01 | .04 | .15* | .08 | .02 | .26** | .23** |
| 30 | Garbles | .08 | -.06 | .08 | .17* | .15* | .11 | .26** | .20** |
| 31 | Weak organization | .08 | .03 | .03 | .17* | .19** | .12 | .20** | **.33**** |
| 32 | Redundancy | -.01 | -.05 | -.05 | .04 | .08 | .06 | .09 | .10 |
| 33 | List technique | .10 | .10 | .10 | .24** | .16* | .13 | .29** | .25** |
| 34 | Faulty punctuation | -.04 | .02 | .07 | -.02 | .02 | .04 | .11 | .12 |
| 35 | Faulty spelling | .03 | -.09 | .04 | .09 | .08 | .08 | .17* | .17* |
| 36 | Shifting point of view | -.04 | .05 | -.05 | -.02 | .07 | .07 | .12 | .01 |
| 37 | Underdeveloped | .05 | .10 | -.11 | .18** | .25** | .15* | **.30**** | **.36**** |
| 38 | Design flaw | .06 | .03 | .03 | .03 | .08 | .03 | -.02 | -.05 |
| 39 | Thesis incongruent w/ lit. | -.04 | .03 | .03 | -.11 | .08 | .03 | .05 | .03 |
| 40 | Statistical reporting error | .11 | .05 | .05 | .06 | -.01 | -.03 | .04 | -.04 |
| 41 | Anal./stats misinterpret. | .05 | -.03 | -.03 | .04 | -.01 | .01 | .05 | -.07 |
| 42 | Wrong analysis | .04 | .02 | .02 | -.03 | .05 | -.19** | .08 | -.03 |
| 43 | Extrapolating beyond data | -.20** | -.07 | .05 | .00 | -.03 | .01 | .04 | .09 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| 44 | Procedural omissions | -.18** | -.01 | -.12 | .02 | .08 | -.02 | .08 | .08 |
| 45 | Req. sci. elements missing | .28** | .01 | .12 | .13 | .16* | .14* | .12 | .13 |
| 46 | Lack of examples | .01 | .04 | .04 | .15* | .18** | .07 | .21** | .18** |
| 47 | Attribution errors | -.02 | -.04 | -.04 | .05 | .07 | .01 | .01 | .12 |
| 48 | References errors | -.03 | .01 | .01 | -.01 | .06 | .09 | .00 | .06 |
| 49 | Excessive passive voice | -.10 | .07 | -.15* | .06 | .05 | -.05 | .03 | .03 |
| 50 | Undefined terms | -.23** | -.03 | -.03 | -.03 | .04 | .03 | .06 | .07 |
| 51 | Wrong plcmt. of sci. info. | .01 | .03 | .03 | .09 | .04 | .05 | .10 | .08 |
| 52 | Inappropriate personific. | -.15* | -.02 | -.02 | -.03 | -.11 | -.08 | .01 | -.03 |
| 53 | Misuse of terms/jargon | -.03 | -.02 | -.08 | .05 | .06 | .06 | .19** | .08 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
|---|---|---|---|---|---|---|---|---|---|
| 17 | Subordinate sequence | 1 | | | | | | | |
| 18 | Coordinate sequence | **.33**$^{**}$ | 1 | | | | | | |
| 19 | Transitions | .16$^{*}$ | .18$^{**}$ | 1 | | | | | |
| 20 | Coherence /cohesion | .23$^{**}$ | .30$^{**}$ | **.44**$^{**}$ | 1 | | | | |
| 21 | Voice | .07 | -.06 | **.35**$^{**}$ | **.50**$^{**}$ | 1 | | | |
| 22 | Narrative storytelling | .10 | -.04 | .08 | .09 | .15$^{*}$ | 1 | | |
| 23 | Design rigor | -.02 | -.02 | .13 | .03 | -.05 | -.02 | 1 | |
| 24 | Analysis rigor | -.06 | .10 | .04 | .09 | .02 | .22$^{**}$ | -.03 | 1 |
| 25 | Exhibit of DV task | -.01 | -.01 | .09 | .10 | -.04 | -.01 | **.70**$^{**}$ | -.02 |
| 26 | Well-blended sources | -.04 | -.11 | .25$^{**}$ | .25$^{**}$ | **.32**$^{**}$ | .17$^{*}$ | -.05 | .26$^{**}$ |
| 27 | Usage problems | .04 | .02 | .18$^{**}$ | .17$^{*}$ | .13 | .12 | .05 | .09 |
| 28 | Misplaced modifier | .05 | .06 | .12 | .03 | .10 | .05 | .03 | .02 |
| 29 | Weak structural core | .12 | .14$^{*}$ | .28$^{**}$ | .22$^{**}$ | .18$^{**}$ | .10 | .11 | .08 |
| 30 | Garbles | .02 | .09 | .30$^{**}$ | .28$^{**}$ | .14$^{*}$ | .01 | .04 | .09 |
| 31 | Weak organization | .10 | .05 | .15$^{*}$ | .20$^{**}$ | .21$^{**}$ | .09 | .05 | .03 |
| 32 | Redundancy | .09 | .10 | .14$^{*}$ | .20$^{**}$ | .11 | .02 | .05 | -.03 |
| 33 | List technique | .05 | .00 | .09 | .26$^{**}$ | .27$^{**}$ | .10 | .05 | .17$^{*}$ |
| 34 | Faulty punctuation | .09 | .10 | .12 | .17$^{*}$ | .09 | -.04 | .07 | -.06 |
| 35 | Faulty spelling | .05 | .11 | .11 | .21$^{**}$ | .10 | -.03 | .05 | .01 |
| 36 | Shifting point of view | .06 | .06 | .09 | .08 | .11 | .05 | .03 | -.04 |
| 37 | Underdeveloped | .12 | .05 | .17$^{*}$ | .26$^{**}$ | .24$^{**}$ | .05 | .11 | .05 |
| 38 | Design flaw | .03 | .03 | .03 | -.12 | -.01 | .03 | .01 | .04 |
| 39 | Thesis incongruent w/ lit. | .03 | .03 | -.04 | .03 | .08 | .03 | **-.34**$^{**}$ | -.08 |
| 40 | Statistical reporting error | -.06 | .05 | -.03 | -.04 | -.11 | .05 | .02 | .08 |
| 41 | Anal./stats misinterpret. | .07 | .07 | .01 | -.04 | -.09 | .06 | .03 | .05 |
| 42 | Wrong analysis | .02 | .02 | .07 | .07 | .05 | .02 | .01 | -.14$^{*}$ |
| 43 | Extrapolating beyond data | -.05 | .07 | .15$^{*}$ | .06 | .04 | -.07 | .06 | -.04 |
| 44 | Procedural omissions | -.03 | -.10 | -.01 | -.03 | .00 | .09 | .11 | .11 |
| 45 | Req. sci. elements missing | .08 | .15$^{*}$ | .14$^{*}$ | .22$^{**}$ | .08 | .17$^{*}$ | .09 | .12 |
| 46 | Lack of examples | .11 | .17$^{*}$ | .24$^{**}$ | .21$^{**}$ | .10 | .04 | .08 | .04 |
| 47 | Attribution errors | .07 | .05 | .24$^{**}$ | .11 | .05 | -.04 | .06 | .02 |
| 48 | References errors | .00 | .04 | .04 | .05 | .04 | .07 | .05 | .08 |
| 49 | Excessive passive voice | .03 | .10 | .03 | .08 | -.01 | -.04 | -.04 | .04 |
| 50 | Undefined terms | .04 | .06 | .06 | .07 | .07 | -.03 | .09 | -.20$^{**}$ |
| 51 | Wrong plcmt. of sci. info. | -.10 | .04 | -.03 | .02 | .04 | .03 | .02 | -.13 |
| 52 | Inappropriate personific. | .00 | .03 | -.04 | -.09 | .03 | .03 | -.10 | .02 |
| 53 | Misuse of terms/jargon | .05 | .02 | .09 | .13 | .05 | .09 | -.03 | -.06 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
|---|---|---|---|---|---|---|---|---|---|
| 25 | Exhibit of DV task | 1 | | | | | | | |
| 26 | Well-blended sources | -.03 | 1 | | | | | | |
| 27 | Usage problems | -.04 | .16* | 1 | | | | | |
| 28 | Misplaced modifier | .02 | .04 | .17* | 1 | | | | |
| 29 | Weak structural core | .08 | .16* | **.37**** | .17* | 1 | | | |
| 30 | Garbles | .03 | .09 | .13 | .09 | .27** | 1 | | |
| 31 | Weak organization | .03 | .16* | .00 | -.08 | .04 | -.08 | 1 | |
| 32 | Redundancy | .03 | -.02 | .21** | .13 | .15* | .00 | .02 | 1 |
| 33 | List technique | .04 | .13 | -.01 | -.05 | .01 | -.01 | .19** | .05 |
| 34 | Faulty punctuation | .05 | .02 | .26** | .02 | .27** | .09 | .04 | .00 |
| 35 | Faulty spelling | .04 | -.02 | .29** | .04 | .23** | **.34**** | .06 | .02 |
| 36 | Shifting point of view | .02 | .05 | .10 | .00 | .11 | .03 | .14* | .11 |
| 37 | Underdeveloped | .08 | .08 | .05 | .05 | .05 | .10 | .23** | -.01 |
| 38 | Design flaw | .01 | .07 | .01 | -.04 | .06 | -.06 | -.07 | -.07 |
| 39 | Thesis incongruent w/ lit. | .01 | -.02 | .01 | -.04 | -.15* | -.06 | .11 | .03 |
| 40 | Statistical reporting error | .02 | .12 | .11 | -.07 | .10 | .01 | -.07 | .15* |
| 41 | Anal/stats misinterpret. | .02 | .04 | .06 | -.09 | .14 | -.05 | -.04 | .19** |
| 42 | Wrong analysis | .01 | .05 | -.05 | -.03 | .09 | -.04 | -.05 | -.05 |
| 43 | Extrapolating beyond data | .04 | .08 | .04 | .06 | .18* | .16* | .09 | .18* |
| 44 | Procedural omissions | .08 | .13 | .11 | .13 | .16* | .03 | .00 | .00 |
| 45 | Req. sci. elements missing | .06 | .10 | -.13 | -.13 | .12 | .11 | .13 | .03 |
| 46 | Lack of examples | .05 | .03 | -.04 | .06 | .15* | .09 | .06 | .10 |
| 47 | Attribution errors | .13 | -.03 | .12 | .10 | .04 | .06 | .10 | .11 |
| 48 | References errors | .11 | .01 | .13 | .15* | .04 | .01 | .14* | .04 |
| 49 | Excessive passive voice | .05 | -.13 | .10 | .06 | .08 | .01 | .02 | .16* |
| 50 | Undefined terms | .07 | .03 | .17* | .20** | .11 | .03 | .01 | .08 |
| 51 | Wrong plcmt of sci. info. | .01 | .03 | .12 | .06 | .12 | .06 | .04 | -.02 |
| 52 | Inappropriate person. | -.07 | -.01 | .16* | .07 | .07 | -.05 | -.09 | .13 |
| 53 | Misuse of terms/jargon | .05 | .09 | .16* | .15* | .21** | .13 | -.06 | .21** |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|
| 33 | List technique | 1 | | | | | | | |
| 34 | Faulty punctuation | .09 | 1 | | | | | | |
| 35 | Faulty spelling | -.01 | **.34**$^{**}$ | 1 | | | | | |
| 36 | Shifting point of view | -.02 | .04 | .03 | 1 | | | | |
| 37 | Underdeveloped | .12 | .13 | .15$^{*}$ | .10 | 1 | | | |
| 38 | Design flaw | -.08 | -.10 | .01 | -.04 | -.16$^{*}$ | 1 | | |
| 39 | Thesis incongruent w/ lit. | .01 | .05 | .01 | -.04 | .06 | -.02 | 1 | |
| 40 | Statistical reporting error | -.08 | -.08 | .02 | -.07 | -.19$^{**}$ | .27$^{**}$ | -.03 | 1 |
| 41 | Anal./stats misinterpret. | -.06 | -.02 | .02 | -.03 | -.13 | .07 | -.05 | .27$^{**}$ |
| 42 | Wrong analysis | -.05 | -.07 | -.05 | -.03 | -.01 | -.01 | -.01 | -.02 |
| 43 | Extrapolating beyond data | -.09 | .07 | .18$^{**}$ | .13 | -.01 | .00 | .00 | -.01 |
| 44 | Procedural omissions | -.05 | .11 | .12 | -.08 | .18$^{*}$ | -.01 | -.08 | -.23$^{**}$ |
| 45 | Req. sci. elements missing | .16$^{*}$ | .03 | .03 | .05 | .02 | -.06 | -.13 | .07 |
| 46 | Lack of examples | .23$^{**}$ | .06 | -.04 | .05 | .17$^{*}$ | -.04 | -.04 | -.07 |
| 47 | Attribution errors | .06 | .10 | .16$^{*}$ | .15$^{*}$ | .14$^{*}$ | -.09 | .00 | -.15$^{*}$ |
| 48 | References errors | .07 | .18$^{**}$ | -.01 | .11 | .03 | .00 | .00 | -.13 |
| 49 | Excessive passive voice | .06 | .06 | .00 | .08 | -.05 | .05 | .05 | .05 |
| 50 | Undefined terms | -.01 | .23$^{**}$ | .18$^{**}$ | .10 | .11 | .01 | .08 | -.24$^{**}$ |
| 51 | Wrong placement of sci. info. | .03 | -.02 | .16$^{*}$ | .05 | .06 | -.03 | .17$^{*}$ | **.30**$^{**}$ |
| 52 | Inappropriate personification | -.09 | .11 | -.01 | -.02 | -.13 | .07 | .14$^{*}$ | .04 |
| 53 | Misuse of terms/jargon | -.01 | .07 | .11 | .05 | -.10 | .04 | -.04 | .06 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 |
|---|---|---|---|---|---|---|---|---|---|
| 41 | Anal. /stats misinterpret. | 1 | | | | | | | |
| 42 | Wrong analysis | .13 | 1 | | | | | | |
| 43 | Extrapolating beyond data | -.01 | -.06 | 1 | | | | | |
| 44 | Procedural omissions | -.03 | -.01 | .10 | 1 | | | | |
| 45 | Req. sci. elements missing | .03 | .01 | -.15* | -.17* | 1 | | | |
| 46 | Lack of examples | -.02 | -.08 | .06 | .08 | .10 | 1 | | |
| 47 | Attribution errors | -.01 | -.06 | .15* | .10 | .03 | .16* | 1 | |
| 48 | References errors | -.21** | -.05 | .00 | .13 | -.06 | .03 | .25** | 1 |
| 49 | Excessive passive voice | .12 | .04 | .16* | .05 | -.10 | .02 | .22** | .12 |
| 50 | Undefined terms | .01 | .00 | .12 | .16* | -.16* | .05 | .13 | .07 |
| 51 | Wrong placement of sci. info. | -.06 | -.02 | .07 | -.04 | -.01 | -.15* | -.15* | -.07 |
| 52 | Inappropriate personification | -.10 | .00 | .17* | .09 | -.11 | -.03 | .02 | .12 |
| 53 | Misuse of terms/jargon | .15* | .03 | .26** | .18** | -.06 | -.06 | .03 | .08 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.

Table F1 (Continued)

*Correlations Among Identified Prominent Features*

| | | 49 | 50 | 51 | 52 | 53 |
|---|---|---|---|---|---|---|
| 49 | Excessive passive voice | 1 | | | | |
| 50 | Undefined terms | .05 | 1 | | | |
| 51 | Wrong plcmt. of sci. info. | .04 | .03 | 1 | | |
| 52 | Inappropriate personific. | .13 | .13 | -.08 | 1 | |
| 53 | Misuse of terms/jargon | .29** | **.31**** | -.03 | .12 | 1 |

*Notes:* Bolded numbers indicate significant correlations of moderate and strong magnitude ($r \geq .3$). * signifies $p$ of .05; ** signifies $p$ of .01 or less. Values for negative features represent successful avoidance, not presence.