Research Article

A Framework for Analyzing Features of Writing Curriculum in Studies of Student Writing Achievement

Kyle Oddis, *Northeastern University*

Jill Burstein, *Duolingo, Inc*[1]

Daniel F. McCaffrey, *Educational Testing Service*

Steven L. Holtzman, *Educational Testing Service*

Structured Abstract

- **Background:** Researchers interested in quantitative measures of student "success" in writing cannot control completely for contextual factors which are local and site-based (i.e., in context of a specific instructor's writing classroom at a specific institution). (In)ability to control for curriculum in studies of student writing achievement complicates interpretation of features measured in student writing. This article demonstrates how identifying and analyzing features of writing curriculum can provide dimensions of local context not captured in analysis of student-generated texts alone. Using a dataset of 48 curricular texts collected from 21 instructors teaching in five disciplines across six four-year public universities in the United States, this article: 1) presents a set of curriculum scoring rubrics developed through qualitative analysis, 2) describes a protocol for training raters to use the rubrics to score curricular texts to achieve rater agreement and generate quantitative data, and 3) explores how this framework

[1] Jill Burstein completed her work on this paper while employed at ETS.

might be amended to more deeply consider feature relationships between curriculum and student writing in studies of student writing achievement.

- **Literature Review:** The literature review provides an overview of existing studies that our research expands upon; grounds rubric development in genre theory, threshold concepts in writing studies, and design thinking; and explores how conducting curriculum analysis in tandem with feature analysis of student writing can benefit writing analytics research programs.
- **Research Questions:**
    - RQ1: What identifiable features of writing curriculum might affect how students approach situated writing tasks?
    - RQ2: How can we categorize features of writing curriculum to help us better understand its role in student writing achievement within and across disciplines?
    - RQ3: Can we produce a quantitative measure of curricular features that can be used in conjunction with natural language processing (NLP) data gathered on features of student writing?
    - RQ4: How can a set of usable, theoretically grounded rubrics offer insight into what research teams interested in studies of writing achievement might consider going forward?
- **Methodology:** The first phase of this study involved qualitative analysis of curricular texts as a guide for creating scoring rubrics. The scoring task in the second phase of the study consisted of three components:
    1. **Development of scoring rubrics.** Rubric development was based on observations from two exploratory rounds of qualitative coding of texts in our dataset. Rubrics addressed five features of writing curriculum: *accessibility*, *applicability*, *actionability*, *situational clarity*, and *overall quality*.
    2. **Annotation protocol training.** Three research assistants with experience annotating linguistic features in texts served as raters and were trained to annotate and score the curricular texts according to the rubrics.
    3. **Application of scoring rubrics.** Each trained rater scored the curricular texts in batches, and the study lead (first author) served as an "expert rater" who also scored the texts so final rater agreements (quadratic weighted kappa) could be calculated.
- **Results:** Qualitative analysis revealed that features of writing curriculum varied widely across learning sites, attesting to a lack of standardization or consistency

of writing curriculum at and across institutions. Quantitative results speak to challenges in producing "usable" statistical data with a limited dataset.

- **Discussion:** Our study illustrates the challenges of applying rubrics to curricular datasets which offer only a partial picture of the realities of teaching and learning writing in multiple disciplines at various institutions. The potential to observe relationships between features of curricular texts and features measured in student writing requires collecting more robust datasets that include assignment grading rubrics, assignment sheets/instructions, and syllabi across disciplines in local contexts where writing happens. Future studies would need to include a sufficiently large number of courses where faculty provide a complete set of relevant curricular materials to allow for course-level analysis.
- **Conclusions:** This study's design is promising for application to larger datasets which may be drawn from single and multi-institutional contexts. Our limited dataset offers inconclusive results for demonstrating relationships between student writing features and features of associated writing curriculum (e.g., student writing motivation and *applicability* of curriculum). However, insights from this process suggest that in order to understand student writing achievement more comprehensively, we must develop more diversified data collection and analysis practices. This would afford deeper insight into the complexities of teaching and learning writing, specifically in terms of how students orient themselves to writing tasks delivered in curriculum. Future approaches to similar kinds of research can offer more insight into how curriculum affects student writing achievement and broader outcomes (e.g., college GPA).

Keywords: student writing achievement, writing curriculum, writing assessment, writing program administration, writing assignments, scoring rubrics, task design, writing analytics

1.0 Background

The relationship between how students write and how curriculum tells them to write in postsecondary contexts (i.e., two-year and four-year colleges and universities) is understudied in the United States. Writing analytics and writing/composition scholars have the methods, tools, and training needed to gain important insights into the relationship between curriculum, pedagogy, assessment, and student writing achievement (Burstein et al., 2017; Collier et al., 2012; Duffy, 2019; Duffy & Agnew, 2020; Elliot, 2016; Gallagher, 2010, 2014; Geller et al., 2016; Gere, 2019; Grouling, 2018; Haswell, 2001; Jo, 2010; Li & Lindsey, 2015; Porter et al., 2000; Ross & LeGrand, 2017; Rourke & Zhou, 2019; Sharer et al., 2016; Shermis & Burstein, 2003; Tannenbaum & Katz, 2021). However, this complex undertaking requires research

approaches that are often limited by availability of curricular and administrative data (Condon et al., 2016; Oddis et al., 2020; Rose & Weiser, 2002).

Since writing curriculum is meant to facilitate student learning of writing genres—particularly those genres students will need for writing in professional and civic life—inattention to the role of curriculum in student writing achievement raises concerns over writing skills transfer to educational and professional contexts (Ringler et al., 2018). It also raises concerns over whether what we claim to value in writing instruction is observable in the texts we (re)produce and (re)use as part of delivered curriculum.[2] When student writing samples are used as the primary sources of data for measuring program efficacy against metrics like retention or student achievement (often through course grades or GPA), we miss out on opportunities to support broader claims that students and teachers are "co-creators" of knowledge. We also miss out on opportunities to adequately perceive benefits and challenges of program reform (Anderson, 2010).

This article describes an exploratory attempt at adding dimension to studies of student writing achievement by attending more closely to the relationship between teaching and learning as it is represented in delivered curriculum. This study reconsiders how curricular texts can provide useful data and explores how that data might be analyzed through qualitative and quantitative approaches. Our study also illustrates what happens when data collection practices are insufficient to yield "usable" statistical results. In this way, our study highlights many of the challenges in conducting "mixed method" research on writing curriculum, especially considering "shifting disciplinary approaches to the study of writing" (Poe, 2019).

Using a dataset of 48 curricular texts (course syllabi, writing assignment sheets, and writing assignment rubrics) collected from 21 instructors teaching in five disciplines across six U.S. postsecondary institutions, our study offers a framework for analyzing features of writing curriculum as a basis for future work. The curricular data for our study were obtained as part of a larger study entitled "Exploring Writing Achievement and Its Role in Success at 4-Year Postsecondary Institutions," funded by the U.S. Department of Education, Institute of Education Sciences (IES).[3] Faculty who agreed to participate in the larger study were invited to supply their course syllabi and descriptions of all writing assignments in their courses.

Four assumptions underly the design of our study:

1. "Successful" completion of situated writing tasks is affected by *how* students have been told to approach those tasks (i.e., how tasks are framed and communicated in curriculum).

[2] Refers to what is "delivered" by an instructor to students which describes institutional, program, or course requirements; texts produced and distributed to students in some tangible material form with which learners interact.
[3] Award Number R305A160115. Opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily reflect the views of the IES. Information on the larger study can be obtained here: https://ies.ed.gov/funding/grantsearch/details.asp?ID=1807

2. Curriculum offers insights into instruction and application of course and program goals and outcomes.
3. Curriculum provides a dimension of local context not captured by student-produced texts or writing samples alone.
4. Curriculum can provide data points which can aid program assessment and faculty development for the benefit of students writing within and across disciplines.

"Curriculum" as defined in this study consists of texts which are composed, curated, recycled,[4] and disseminated across institutional writing sites that express their creators' values regarding what a student is meant to learn about writing and why they are meant to learn it. "Values" can be defined a number of ways, but in context of this study, values are understood as not simply the criteria to which we respond—as administrators, instructors, students, and researchers—through rhetorical performance; values are also about how "ethical, pedagogical, and political commitments" (Broad, 2003) materialize in the texts we create and (re)use in college writing contexts. Aull (2020) writes that "it is not only in theory but also in the cumulative effect of patterned discourse that school assignments shape writing and thinking" (p. 5). The findings from our study support this understanding. Aull (2020) writes that "school genres" which are "sanctioned by institutions and used repeatedly in student learning" are "important examples of institutionalized truth and power," (p. 2) and are therefore instrumental in understanding student writing achievement. It is difficult to imagine that curriculum—how it is designed, articulated, and delivered—does *not* have an impact on NLP-based assessments of student writing since student texts are produced often in direct result of curricular requirements. Indeed, "school discourse maintains institutions and helps shape how we understand those institutions and associated communicative actions" (Aull, 2020, p. 3). We cannot depend on "unconscious assimilation as the primary way students must make the connection between assignment expectations and the written choices they make" (Aull, 2020, p. 4).

Syllabi, assignment sheets, and grading rubrics are what we would describe as "core" curricular texts in that they are essential to student learning in writing in the disciplines (WID) and writing across the curriculum (WAC). They also represent "particular disciplinary developments we can reasonably challenge today" (Aull, 2020, p. 4). These are texts which express the values of their creators to varying degrees. In this study, we classify the curricular texts as illustrated in Table 1.

[4] See "A Text-Analytic Method for Identifying Text Recycling in STEM Research Reports" (Anson et al., 2019) and https://textrecycling.org for information on how researchers reuse materials from their own work. We can also think of some curricular texts as "recycled" in this way.

Table 1

Classification and Characterization of "Core" Curricular Texts

Course syllabi	Assignment sheets	Assignment rubrics
• Outline course goals relative to program goals and outcomes in a student's program of study • Suggest application of knowledge (transfer) to other courses or to students' professional and civic lives • Offer insights into how students are assessed (graded) on writing tasks (e.g., relative weight of writing assignments in calculating overall course grades) • Offer insights into an instructor's disciplinary understanding of writing's role in student learning • May be used to determine course credit equivalencies between institutions	• Provide additional context for writing task completion • May narrativize assignment goals relative to course/program goals and outcomes • Suggest application of knowledge for transfer between assignments (e.g., scaffolding) • Offer insights into an instructor's approach to writing processes or highlight an instructor's pedagogical or disciplinary values (e.g., peer review, formative vs. evaluative feedback) through emphasis of certain practices (e.g., deadlines, submission format, citation, use of analog or digital tools)	• Provide a point-value breakdown of elements for which students are scored with greater specificity or nuance • Reveal relative importance (weight) of certain writing features in grading (i.e., what is weighted more in calculating an overall assignment grade via distribution of points) • Can offer more detail on assignment scoring when assignment sheets provide context or theoretical justification for writing tasks but no point breakdown to demonstrate the importance of certain elements that "should be" present in final written products

These core curricular texts are important to an institution's functional ecosystem and offer fundamental insight into course design and instructor practices. Inattention to analysis of curriculum becomes consequential for researchers seeking to understand student achievement in postsecondary writing contexts, particularly when values are difficult to identify in curriculum with enough consistency to make valid claims about what actually "matters" in writing instruction. Difficulties in determining what instructors value and how those values are communicated to students make it all the more difficult to determine whether students have "achieved" the goals and outcomes assigned to them. In other words, how students receive, interpret, and attend to the values (both disciplinary and personal) communicated in curriculum can impact their ability to complete writing tasks "successfully."

Moreover, inconsistencies in curriculum design and language can result in student learning inequities—i.e., students may not be on equal footing to complete an assignment absent certain curricular features—which can compromise successful completion of coursework writing tasks. For instance, assignment instruction sheets are often used to frame the rhetorical situation of tasks (what students are being asked to write, for whom students are being asked to write, and why students are being asked to write). Determining whether a rhetorical situation is new for students presents a major challenge for instructors composing curriculum in any discipline where writing happens. For students, lack of genre awareness while attempting to complete situated writing tasks is a result of a number of complex life factors such as previous schooling (McComiskey, 2012), cultural and socioeconomic differences (Andersson, 2018), neurological differences (Ortiz, 2020), and linguistic diversity (Heugh, 2014; Rose & Weiser, 2018). While researchers interested in quantitative measures of student success and mixed method approaches to writing studies research cannot control completely for situational and local contextual factors which are site-based (i.e., in context of a specific instructor's writing classroom at a specific postsecondary institution), our study suggests that an inability to control for curriculum in studies of student writing achievement adds complexity to interpretation of features captured by NLP analyses of student writing.

Further justification for our study's attention to curricular features is based in awareness of the educational realities of today's students who are "highly mobile" (Gürüz, 2011; Shkoler et al., 2020; Stubblefield, 2016), meaning that they are physically and virtually moving between places, jobs, and social classes (Shkoler et al., 2020). "Mobility" is the new order of the day as students participate in a global network of learning augmented by pandemic conditions. Movement between institutional contexts and course delivery methods (Zhang-Wu, 2020) also necessitates greater attention to the relationship between writing curriculum and student writing achievement. Looking at features of student writing and features of writing curriculum becomes especially helpful in understanding and characterizing "equivalency" of coursework, whether in context of college-to-college equivalency or dual-enrollment equivalency in high schools that provide college credit to students taking certain courses. This is true in the United States, where institutional outcomes are often shaped by the *type* of institution at which students are learning.

Two-year and four-year public and private colleges and universities cater to diverse student populations that tend to vary by regional location, and face an array of constant funding challenges that are influenced by both state and federal politics. Add this to the challenge of determining equivalency when evaluating international curriculum for awarding transfer credits either to or from U.S. institutional contexts, and the picture of factors that influence student "success" in writing becomes quickly and overwhelmingly complex. Not all postsecondary institutions have articulation agreements (Bers, 2013; Payne et al., 2021; Worsham et al., 2021), which are essentially "curriculum roadmaps" (Payne et al., 2021). Writing program directors and administrators are often tasked with deciding if cross-context courses bear credit equivalency, and as we will see from the results of our study, that is a difficult task—not least of all because, even when courses present the same *content*, expression of that content in its presentation and design can render educational experiences nonequivalent for diverse writers.

What should administrators look for when determining credit equivalencies for the "highly mobile" students who are writing today? What should we actually value in determining credit equivalencies for courses where writing happens, and how/can we agree on what *to* value? We believe that we can identify curricular features that can point us in a clearer direction for determining equivalency and equity of writing coursework without placing additional limitations on instructor freedom to design writing tasks that meet course and program goals.

2.0 Literature Review

The literature review provides a foundation for considering how writing analytics researchers working with student writing sample data might benefit from conducting in-tandem analyses of associated curriculum. The first of these benefits comes from considering the potential for studies of curriculum to aid writing program administrators (WPAs) in determining transfer credit equivalences when addressing student mobility in a digital-hybrid world. The second benefit has to do with addressing transparency in how reliability is achieved between raters and how it precedes validity (i.e., when we attempt to calculate "reliable" statistical data but cannot readily see how difficult it is to achieve). The third benefit relates to applying curriculum scoring rubrics to generate insights about application of disciplinary values as students attempt to achieve outcomes associated with these values in their writing. When considered within the taxonomy of writing analytics research, this study has promise for research programs across all four areas: educational measurements, massive data analysis, digital learning ecologies, and ethical philosophy (Lang et al., 2019).

2.1 Influential Studies of Writing and Assignments

> Since Frederiksen et al.'s (1957) "In-basket test," some writing studies scholars—and now writing analytics scholars—have taken up the call for a clear need for instruments which will measure such complex skills as the ability to organize discrete pieces of information, to discover the problems implicit in a situation, to

> anticipate events which may arise because of such problems, and to arrive at decisions based on a large number of considerations. (p. 1)

Aull (2020) does this by showing how "empirical methods from linguistics which scale" can "make visible the discourse expectations in academic writing instruction" (Marcellino, 2020). Aull (2020) uses both distant and close reading to help characterize what makes "good" academic writing for students through qualities of "civility," "cohesion," and "compression." By mapping certain features of student writing achievement onto discourse conventions (drawing upon genre theory and Swales' articulation of "discourse communities" [5]), Aull examines discourse at lower- and upper-level points in students' argumentative writing development. By focusing on sentence-level features for analysis, Aull draws attention to how measurement of these features can contribute to an instructor's ability to design effective assignments and rethink/challenge disciplinary values and practices that do not serve student writers. Aull supports implications for assignment design in this comprehensive study of argumentative writing as a genre students practice throughout their educational careers by offering a typology of writing assignments that instructors might use to help explain and clarify a genre's conventions to students. The usefulness of this typology is supported by Aull's argument, and while the study does not apply analytic methods to the curriculum itself, the relationships between presence of conventions, discourse expectations, and the role of curriculum in instruction are logically inferred. This study lays groundwork for future work that might be done to understand these relationships more deeply.

Melzer's (2014) foundational curriculum study inspired our approach in its pursuit of deeper understanding of writing assignments as a necessary component of addressing student writing achievement. Melzer's work responds to Anson's (1993) call for more large-scale research on WAC and traces the impact of the WAC movement through analyzing rhetorical situations (with particular attention to "purpose") of writing assignments using "quantitative distributions of purpose and audience" and a "textual analysis of representative assignments and related materials available on class websites, such as grading rubrics and course outcomes" (2014, p. 20). Melzer's results suggest substantial "differences between the distribution of purposes and audiences for WAC courses" (pp. 72-74) relative to other courses in his dataset of 2,101 writing assignments. Melzer's study engages in an important discussion of the relationship between assignments and instructor pedagogy and makes a case for WAC courses as transformative initiatives in higher education.

Isaacs (2018) explores "evidence of recent changes to the higher education landscape," noting that we are "more public with our practices and processes" now with the rising interest in higher education in "accountability" alongside "concerns over debt and graduation rates" which parallel a rise of increased "oversight and requirements" (p. 9). Isaacs' study is shaped, therefore, by the "quality and quantity of data provided in reliable public sources" and decidedly "does not

[5] Swales articulates the concept of a discourse community in *Genre Analysis* (1990) and later revisits it (2017).

benefit from the kind of insider and deep knowledge that an interview or review of a range of syllabi would present" (2018, p. 9). However, Isaacs' coding of institutional documents highlights the repetition of phrases like "purpose," "audience," "mechanics," "rhetorical awareness," and others that contribute to "readers' sense of the construct of writing each institution is presenting" (2018, p. 9). Covering issues of writing studies' practices as a field, presence of institutional support, and the evolution of first-year composition (FYC), Isaacs' comprehensive overview and application of close reading, text mining, descriptive reporting, and content analysis methods offers a take on "mixed method" analysis of curriculum which provides a useful model for our study. Isaacs' method is based on a two-rater system to provide "greater reliability" for "reporting that relied on interpretation" while adding that a single rater was "sufficient" for determining aspects like institutional requirements and terms like "Standard English" (pp. 177-178).

Section 4.2.2 describes the process of achieving rater agreement for our study's curriculum scoring tasks and expands on the work of the studies discussed above by approaching analysis of curriculum through calculating quadratic weighted kappas (QWK) rather than relying on the analysis of curricular texts by single or dual raters. This was done in an attempt to provide expanded considerations of reliability to our insights. This additional step in our research design offers a path for research teams working across institutional and organizational contexts—something that serves writing analytics researchers particularly well in that the field is developing with a vested interest in multidisciplinary collaborations (Moxley et al., 2017). To our knowledge, no study to date has attempted QWK in studies of writing curriculum.

2.2 Concepts for Curriculum Scoring Rubrics

The QWK metrics presented in the results section were calculated based on scores generated through application of rubrics designed to score the curricular texts in our dataset. The three overarching concepts reviewed in this section informed development of our scoring rubrics: 1) genre theory, 2) threshold concepts in writing studies, and 3) design thinking. Rationales for applying each concept to rubric development are provided below.

2.2.1 Genre Theory

A key component of evaluating curricular texts involves evaluating how an assignment helps a student understand a rhetorical situation,[6] particularly if a rhetorical situation is new or unfamiliar. The presence or absence of a clear rhetorical situation in curriculum can help or hinder a student's ability to write in certain—especially unfamiliar—genres. It is beneficial to code curricular texts through the lens of genre theory because genre identification can be troublesome for student writers that do not possess a firm understanding of what it means to write in different rhetorical situations, especially if the audience or exigence is not clearly

[6] Defined as the context of a rhetorical act, made up (at a minimum) of a rhetor (a speaker or writer), an issue (or exigence/need), a medium (such as a speech or a written text), and an audience.

defined for novice writers by experienced instructors. Determining whether a rhetorical situation is new for students presents a major challenge for instructors composing writing curriculum in any discipline.

In curricular texts like assignment instruction sheets that *mention* audience but do not clearly *define* it, for example, a student's working assumption might be that the primary audience for a writing task is the instructor, even if they are writing in genres with additional (usually hypothetical) external audiences. This can become confusing for students if the audience is not clearly defined *and* if students do not come into a writing classroom with prior genre knowledge (Devitt, 2004). This is where genre theory offers a key component to interpretation of curriculum: Genre theory is used to inform our understanding of how audience is defined in curriculum (if it is defined at all), and the presence or absence of a clear rhetorical situation in curriculum can help or hinder a student's ability to write in certain genres, which might ultimately influence the presence of certain student writing features that researchers are interested in measuring. While Devitt (2004) was the primary inspiration for coding for genre in the curricular dataset of this study, plenty of literature in the field emphasizes the importance of understanding genre in writing tasks (e.g., Bawarshi, 2000, 2003; Duff, 2002; Sullivan & McConnell, 2018). It makes sense, then, to consider genre as a critical component for evaluating writing curriculum in terms of how well a curricular text presents the requirements of a given genre and its rhetorical situation.

2.2.2 Threshold Concepts in Writing Studies

Faculty tend to structure their ideas of learning and teaching around threshold concepts from their fields. Adler-Kassner and Wardle (2016) define threshold concepts for the field of rhetoric/composition or writing studies as ideas that writing teachers have about writing, framed as the ways in which writing teachers understand the practices and processes of writing that have the potential to change a learner's stance. Threshold concepts help learners see things (e.g., the world, themselves, their abilities) differently. The difference between understanding/applying a threshold concept and learning/memorizing a technical term, for instance, is that learning a term does not necessarily change a person's point of view—it just provides new language to talk about that point of view. Learning a threshold concept, on the other hand, and seeing it in action, can fundamentally change how a student understands a concept. This can affect the way a student approaches a rhetorical situation when writing about any given topic. (Adler-Kassner & Wardle, 2016)

In writing studies, threshold concepts help us better understand how instructors understand and value writing in all its complexity. These general concepts (divided in Adler-Kassner and Wardle's collection into five larger categories and 36 sub-concepts) describe most of what scholars in the field of rhetoric and composition currently believe is important in teaching writing. The dataset of curricular texts from this study—which included writing assignments

from several disciplines—show that writing happens in courses across postsecondary curricula, and these threshold concepts are expressed to varying degrees in each curricular instance.

Attending to these concepts in rubric development and application allows us to see how instructors might choose to frame writing tasks in curriculum based on what they generally believe about writing as a process, practice, or set of practices within and beyond their disciplines relative to how experts in writing/composition studies understand writing. The choices students make in composing often reflect what their instructors have articulated as valuable in their framing of tasks. Therefore, threshold concepts are theoretically and rhetorically useful for informing any assessment of writing curriculum since expression of these concepts might influence presence of student writing features and might also provide insight as to what a writing program or writing instructor values in evaluating student "success."

2.2.3 Design Thinking

Universal Design for Learning (UDL) suggests that there is more contributing to curriculum quality than theoretical uptakes of genre or threshold concepts alone. Because UDL is about how concepts are presented and how this presentation on a design level can impact a student's ability to complete a writing task (Meyer et al., 2014), coding for it is important in terms of identifying where curriculum attends to accessibility and usability. To design an accessible curricular text, an instructor must consider basic formatting elements of a document; for example, the use of the Americans with Disabilities Act (ADA) standard-compliant fonts, inclusion of headings that enable screen reader compatibility, and use of color schemes visible for colorblind students (Burke et al., 2016; Null, 2013). Instructors may choose to rely on color, for example, to differentiate important instructions in curriculum, which can be unintentionally hindering to a student's ability to identify and act upon this information. But accessibility does not stop with basic formatting elements. Critical design approaches (Gonsher, 2016; Purdy, 2014; Wiggins & McTighe, 2005) highlight the importance of framing design thinking as more than just "a useful myth" (Sheffield, 2018). Inclusive education cannot stop at adoption of basic UDL principles; critical design lenses augment our understanding of accessibility by considering more deeply how UDL might (dis)connect with disability studies (Baglieri, 2020; Osorio, 2020).

Accessibility of texts is also connected to usability for students across the spectrum of difference (Tomlinson & Newman, 2017). "Usability" refers to learnability; in other words, how straightforward the instructions are such that it is easy for students to figure out what they need to do with information to which they have access. Usability of curriculum matters because it speaks directly to a student's ability to learn the content presented and act (write) accordingly in response. Usability may be enhanced by additions to curricular texts like outlines, bulleted lists, checklists, and examples—these are some of the ways instructors can proactively consider neurodiversity and disability in curriculum design and development.

At many institutions, professional development that covers these aspects of curriculum design is underattended to, and proactive planning for accessible and usable course texts has

been found to be inconsistent at best (Fisher & Wright, 2010; Scott et al., 2017). As observed during this study, some syllabus templates do not reflect even basic UDL thinking. Often, accommodations at the institutional and instructional level are reactive instead of proactive (Yergeau), and many instructors do not know about or do not have access to on-campus resources that might aid them in creating more accessible and usable curricular materials. Moreover, it can be difficult for instructors to know what students need explained and in how much detail, and it is rarely easy for instructors to identify neurodiversity in students unless the students are willing to divulge this information. Some students may not even know that they are neurodivergent absent a proper and accurate diagnosis, so proactive design considerations are crucial in nurturing student success. Students are also not mandated to disclose diagnoses (Tomlinson & Newman, 2017), and, if curriculum is designed in a true spirit of equity, they should not have to (Brewer et al., 2014; Yergeau, n.d.). This is why it is even more important to proactively consider curricular features in terms of both accessibility and usability when interpreting the role of curriculum in measuring student writing achievement. How accessible and usable a curricular text is for any student directly impacts that student's ability to execute the writing task in front of them.

2.3 Expanding Insights from NLP Technologies

The concepts and ideas discussed in the previous sections offer grounds for more deeply considering the role of curriculum in influencing what we see in feature measurements of student writing samples generated by NLP. It is difficult to understand *why* we are seeing feature presence in some writing samples more than in others if we do not take a more comprehensive view of all the factors that influence feature presence—especially when curriculum so directly affects students' understandings of what they are meant to do and how they should go about doing it. Ling et al.'s (2021) study of motivation in student writing came from NLP data on writing features present in the writing samples collected from the larger IES study that initiated this article's project.

Ling et al.'s (2021) study demonstrates the depth of insight that can be gained from understanding student writing motivation through NLP technologies like automated writing evaluation (AWE), but also highlights the limitations of this approach when contextual factors which are site-based are not made evident. While "understanding connections between motivation and linguistic features of writing has the potential to improve instruction" and provides some "directions for pedagogical interventions" (Ling et al., 2021, p. 2), the role of curriculum and potential curricular interventions could theoretically influence the presence of certain student writing features that contribute to motivation. In their literature review, Ling et al. (2021) cite research on writing motivation in terms of predictors like "self-efficacy" relative to achievement goal theory which includes "three contrasting goal orientations toward learning: mastery, performance-approach, and performance-avoidance" (p. 2). They note that "people tend to engage in activities that make them feel competent" (Ling et al., 2021, p. 2), so a natural

question that emerges is: How might curriculum contribute to or impede students' feelings of competency when approaching a situated writing task?

We hoped the approach we took to analyzing curriculum in our study would help us understand why NLP analysis of student writing in the larger IES study was not always substantive enough to make all the claims we wanted to make despite its provision of insights on features contributing to writing motivation as offered in Ling et al. (2021). We hoped that analyzing the assignments associated with the student writing samples would allow us to revisit the results of NLP feature analysis and understand in more depth why the study team saw the kinds of results it observed in student writing samples.

3.0 Research Questions

Our motivation for this study was to find out if there are observable relationships between features of writing curriculum and features measured in student writing produced as uptakes[7] of specific curriculum. Operating with an understanding that curriculum provides a dimension of local context not captured by student writing alone, we addressed the following questions:

- RQ1: What identifiable features of writing curriculum might affect how students approach situated writing tasks?
- RQ2: How can we categorize features of writing curriculum to help us better understand its role in student writing achievement within and across disciplines?
- RQ3: Can we produce a quantitative measure of curricular features that can be used in conjunction with natural language processing (NLP) data gathered on features of student writing?

As the study progressed, we realized that we did not have adequate curricular data to answer all of these questions completely or to map our observations onto data obtained as part of the larger IES writing achievement study. Therefore, the research questions we initially posed for the study produced results that raised a new question that describes what this study actually produced:

- RQ4: How can a set of usable, theoretically grounded rubrics offer insight into what research teams interested in studies of writing achievement might consider going forward?

In what follows, we present our methodology as a framework for taking up and expanding this work in future writing analytics research projects.

4.0 Research Methodology

[7] "Uptake" refers to the concept as articulated by Anne Freadman in "Anyone for Tennis?" (1987) and "Uptake" (2002), which describes the relational complexities of genres within activity systems, drawing on Carolyn Miller's (1984) idea of genre as social action.

As with most exploratory studies, the impetus for ours began with attempts to figure out what could be gleaned from the curricular data we had without a clear picture of what it might reveal. Seeking answers to our research questions began with two rounds of qualitative coding[8] of the collected curricular texts submitted by instructors participating in the IES study. The coding process was grounded in the concepts reviewed in section 2.2, and these concepts influenced the design of the scoring rubrics presented in section 4.2—as did questions from institutional site partners who offered feedback on findings from the rounds of coding over the course of the study's development.

4.1 Qualitative Coding

Fifty initial codes were identified in six groups based on the concepts reviewed in section 2.2. Each curricular text in the set of 48 texts was coded using HyperRESEARCH software.[9] This first large, broad set of 50 codes was used to establish a baseline understanding of common features and elements in the curricular texts in our dataset. In this first round of coding, the following set of exploratory questions served as a guide to develop the initial "codebook" which allowed for observation of general trends and traits across curricular texts:

1. What are students being asked to *do* in this assignment?
2. How are students being asked to do it?
3. Why are students being asked to write *this way*?
4. What is the specific writing task or set of writing tasks outlined in the curriculum? How can we identify and categorize these tasks?
5. Who is the audience for the assignment? Is the audience clearly defined?
6. What is/are the genre(s) students are writing in? How is the writing process articulated for certain genres?
7. How is writing *itself* discussed? (Is it discussed as a process/in a series of steps?)
8. How might the addition of sources, style guides, and formatting requirements clarify or obfuscate what students are being asked to do?
9. Do certain features of curriculum make the task easier or more difficult to understand? How do design/formal elements contribute to this?
10. What characterizes a "good" or effective assignment text? Can effective assignment texts be characterized by greater frequencies of certain text features?
11. Are some text features more prominent in assignments from certain genres?
12. What might frequencies of certain features in assignment texts suggest about delivered writing curriculum across the college writing landscape?
13. How might the presence of certain assignment text features affect student learning?

[8] See Saldaña (2015) for multiple approaches to coding qualitative data.
[9] See HyperRESEARCH: http://www.researchware.com/products/hyperresearch.html

The six initial categories based on these guiding questions consisted of codes that identified the following curricular features:

1. Assessment (scoring; e.g., extra credit, penalties)
2. Concepts (context; e.g., shared knowledge or required text, theory, and scholarly references to source materials)
3. Design (format; e.g., font, headings, bold, italics, visual elements)
4. Function (logistics; e.g., how to submit, due dates, required style guides)
5. Institutional requirements (goals; e.g., course or program objectives)
6. Writerly accomplishments (tasks; e.g., audience awareness, grammar/mechanics, source use and citation practices)

In the first coding round, what stood out across all assignment text types were "design" elements (388 total coded instances out of 1,007)—specifically, formatting and the visual organization of information. It became apparent during coding that most assignment texts were not proactively employing UDL principles; the use of ADA-compliant fonts was consistent, but there was no evidence that this was an active consideration on the part of instructors given that word processing programs like Microsoft Word default to the use of fonts like Arial and Times New Roman, which are ADA-compliant fonts compatible with screen readers. Whenever instructors did use features like headings and subheadings, they were used to create stylistic features rather than functional features that would be amenable for students using screen readers to navigate a document.

The other design element that stood out immediately was the overuse of bolded text. Instructors consistently relied on formatting elements like bold and italics to highlight important information; however, this did little more than draw the eye to too many places, especially when an instructor made a choice to, for example, bold approximately one-third of the text in question. In the first round of coding, "bold" was by far the most frequently occurring code, coded 161 times for the set of 48 texts. There is no clear evidence to suggest that bolding has any positive effect on a student's ability to understand a writing assignment, and in fact, it may distract learners from readily understanding what it is they need to do.

Other design elements coded in the first round included:

- LegibleFont (48)
- LongBlockQuote (5)BulletN
- umberedList (24)
- ColorFont (7)
- ItalicFont (45)
- UnderlineFont (63)
- Outline (4)
- VisualTitle (23)
- HeadingFeature (3)

- VisualElements (5)
- CaptionsAltText (0)

What these counts suggest is that instructors employed design elements mostly to draw attention to information they deemed important, but they did not consider accessibility elements like use of headings for screen readers (HeadingFeature) or captions/alt text for any images they included (CaptionsAltText).

There was also noticeable emphasis in assignment instructions on functional tasks, such as how to submit assignments (e.g., where to upload, in what document format, when it was due, and what style guide to use). Instructors also emphasized how to incorporate sources—specific sources they wanted students to include and how they wanted students to cite those sources (e.g., APA or MLA style). Thesis statements were also observed as an important feature; however (and surprisingly), few curricular texts included examples of what a "strong thesis statement" should look like. When examples were used in some assignment sheets (which was quite rare), they were offered in list form and often offset by ambiguous language like "may" (e.g., "you may do this, or you may do that").

Finally, there was evidence of genre complexity: A noticeable portion of curricular texts worked to outline or explain how to write in certain genres while also listing instructor expectations. While many instructors mentioned "genre" and "audience" as terms in their texts, there was often little elaboration on what a genre meant (i.e., what the rhetorical situation required). Instead, instructors usually opted to challenge their student writers to make assumptions based on prior knowledge. As previously mentioned, prior knowledge is not a given; instructors cannot assume students have prior knowledge of how to enact certain (especially hybrid) genres, even if students had to "demonstrate comprehension" in some other way before taking a course (e.g., through a placement or AP exam, which are unique writing genres of their own). In some cases, instructors would identify multiple genres in the instructions for a single assignment, asking students to, for example, write a "letter" for the task, then refer to the same task as writing an "essay" that requires a "thesis statement."

After the first round, codes were modified from the initial 50 codes to a new set of 15 codes and categorized into three groups instead of six groups. The new codes and groups (in italics) consisted of the following:

1. *Genre* (8 codes): AudienceMention; AudienceStatement; Complexity; Example; Features; SituationDefined; SourceUse; Thesis
2. *Task Completion* (5 codes): Feedback; Grading; Parameter; Resource; SourceText
3. *Transfer*[10] (2 codes): Goal (vertical transfer); Scaffold (horizontal transfer)

[10] Vertical and horizontal transfer definitions are adapted from the International Bureau of Education's definition of "Transfer of learning": http://www.ibe.unesco.org/en/glossary-curriculum-terminology/t/transfer-learning

The groups and individual codes used in the second round of coding were determined based on general patterns observed in the first round. The results from coding round two are displayed in Table 2. See Appendix A for a list of code definitions and examples.

Table 2

*Code Frequency Counts from the Second Round of Qualitative Coding of Texts (*N*=48)*

Genre group (249)	TaskCompletion group (202)	Transfer group (22)
• AudienceMention (13) • AudienceStatement (5) • Complexity (27) • Example (31) • Features (77) • SituationDefined (64) • SourceUse (19) • Thesis (13)	• Feedback (10) • Grading (40) • Parameter (108) • Resource (19) • SourceText (25)	• Goal (16) • Scaffold (6)

The findings from both rounds of coding demonstrate a general lack of consistency and standardization across all assignment text types, emphasized in the distribution of codes presented in Table 3, which shows substantial variability in the presence of curricular features at and across all six institutional sites.

Table 3

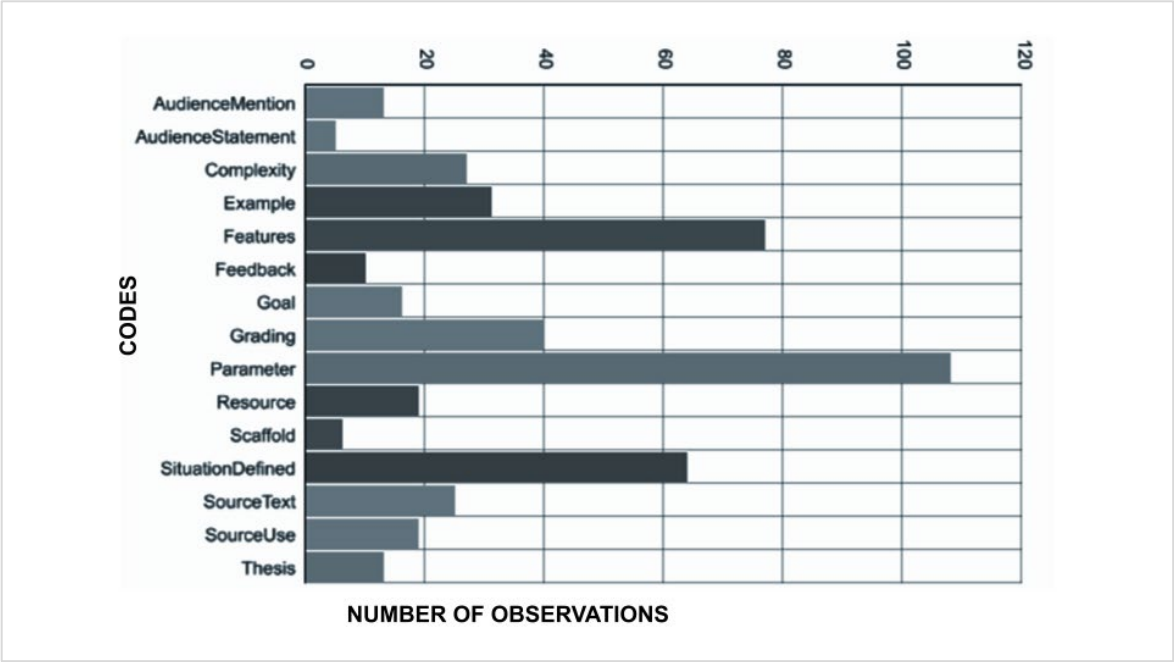
Codes and Results of Code Frequency Counts Across All Core Curricular Texts and Code Frequency by Side (Indicates the Number of Texts in Which the Code Occurred at Each Site)

Code name	Site A	Site B	Site C	Site D	Site E	Site F	Code subtotal
AudienceMention	7	0	5	0	1	0	**13**
AudienceStatement	4	0	0	0	0	1	**5**
Complexity	11	1	8	2	0	5	**27**
Example	3	0	9	1	18	0	**31**
Features	25	2	16	1	31	2	**77**
Feedback	4	1	2	1	0	2	**10**
Goal	1	3	5	2	1	4	**16**
Grading	17	4	10	4	2	3	**40**
Parameter	32	3	37	6	21	9	**108**
Resource	3	1	6	1	4	4	**19**
Scaffold	1	1	2	0	0	2	**6**
SituationDefined	28	3	22	2	7	2	**64**
SourceText	11	2	8	0	0	4	**25**
SourceUse	7	0	1	0	10	1	**19**
Thesis	8	0	4	0	0	1	**13**
Site subtotal	**162**	**21**	**135**	**20**	**95**	**40**	

Assignment sheets generally began with an attempt to define the rhetorical situation of the assignment ("SituationDefined"). Most followed with some outline of "Features," but fewer than anticipated included examples of what presence of these features might look like in a student's final written product. The most frequent use of the "Example" code came from assignment sheets in biology. "Complexity" and "AudienceStatement" codes appeared less frequently than anticipated. Assignment sheets made mention of "an audience" (ambiguous) over twice as frequently as audience was actually explicitly named or defined. Various combinations of other *Genre* and *TaskCompletion* group codes were used to try to explain or clarify writing task objectives. Assignment grading rubrics focused almost exclusively on "Grading," "Parameters," and "Features," as did syllabi that discussed assignments. Some syllabi mentioned "Feedback," but most only offered "Resources" specific to writing. The bulk of the coded "Goal" instances came from syllabi. Figure 1 illustrates that "Parameter" was the most frequently observed code across all six sites, followed by "Features" and "SituationDefined."

Figure 1

Number of Observations for Each Code Across All Six Sites



Attention to the disparities represented in Figure 1 can help programs and administrators create templates for assignments and offer additional opportunities for professional and faculty development even without application of scoring rubrics like those offered in Section 4.2. This information is also useful in helping researchers develop tools that might assist students in determining whether their writing is accomplishing the goals and completing the tasks set for them by their programs and instructors. This also provides incentive to conduct similar work in the future with larger datasets, even if that work is limited to qualitative analysis.

Information from theory-based qualitative coding alone can inform how writing programs and instructors construct curriculum and understand the relationship between curriculum, pedagogy, and academic resource/writing centers, which have historically served as "mediators" between assignment texts, instructors, and student writers attempting to achieve writing goals and complete the tasks assigned to them amidst complex situational factors (Bromley et al., 2013; Denny & Towle, 2017; Newman & Dickinson, 2017). Because the two rounds of coding revealed such a startling lack of consistency in curriculum both at and across writing sites, site partners and advisory board members for the study were keen to understand how such findings might offer opportunities for faculty development and writing program assessment at their own institutions. Site partner questions and concerns were considered in designing the scoring rubrics presented in the next section following a discussion of coding findings at an annual site partner meeting held at Educational Testing Service. Observations of curricular feature presence represented by code frequencies influenced the development of the scoring rubrics, which were

then applied to score each curricular text in our dataset individually. Essentially, we wanted to use insights from the qualitative analysis to create tools (rubrics) that could generate additional quantitative data.

4.2 Curriculum Scoring Task

The curriculum scoring task included rubric development, rater training, and application of rubrics to yield numeric scores for each of the 48 curricular texts, which we hoped to use as quantitative data in conjunction with statistical measures of student writing features classified in the larger IES study.

4.2.1 Development of Scoring Rubrics

Site partner questions about the results of the qualitative analysis described in the previous section guided development of the "Curricular Text Scoring Rubrics" we used to score curriculum for *accessibility*, *applicability*, *actionability*, *situational clarity*, and *overall quality*. The hope was that researchers who were using corpus linguistics and NLP to produce student writing feature metrics might be able to map relationships between curricular features onto the metrics generated from student writing samples composed in response to the associated curricular texts.

It should be noted that the decision to design analytic feature rubrics and one holistic *overall quality* rubric is grounded not in a desire to rigidly standardize curricular texts, but instead to address equity in a demonstrable set of transparent goals for designing writing tasks that instructors, administrators, or researchers can accept or reject. For example, a task that does not attend to UDL is what we would consider a "diminished" writing task in delivered curriculum. Simultaneously, we also understand that there are many ways to meet objectives and design a writing task to meet program standards or course objectives. A decision to include separate rubrics for *accessibility* and *situational clarity*, for instance, speaks to an understanding that 60 years of research in writing studies tells us that to design a writing task in ignorance of genre (*situational clarity*), for example, is potentially harmful for student learning, and so, it is a separate measure from *accessibility* that is still relevant to understanding situated writing tasks as presented in delivered curriculum. The separate analytic measures can be considered in tandem or in isolation, so creating a set of analytic and holistic rubrics accounts for sources of variability in task design. Separating analytic measures allows us to understand some of that task variability and consider it in holistic scoring; moreover, separating these measures allows us to consider why we might give more weight to one measure over another in holistic scoring based on research goals and overall study design.

The final rubrics are presented below. Note that "writing assignment" encompasses all three categories of "core" curricular texts in our dataset as the syllabi included in the dataset all contained descriptions of writing assignments. The "holistic assessment" (*overall quality*) considers all four analytic quality measures: *accessibility* (attending primarily to design

thinking), *applicability* (attending primarily to threshold concepts), *actionability* (attending primarily to design and usability), and *situational clarity* (attending primarily to genre).

Curricular Text Scoring Rubrics

Holistic assessment: *I agree that this writing assignment incorporates: 1) accessibility, 2) applicability, 3) actionability, and 4) situational clarity.*
4 (Strongly agree)
3 (Agree)
2 (Disagree)
1 (Strongly Disagree)

Accessibility: I agree that this writing assignment considers accessibility.

Table 4

Accessibility *Scoring Rubric*

3 (Strongly Agree)	2 (Agree)	1 (Strongly Disagree)
Highly Accessible Designed considering UDL principles; ADA-compliant fonts, headings, and color schemes; judicious use of formatting choices (e.g., bold or italics); any visual or multimodal components included are captioned and hyperlinks are used according to accessibility standards; options for multiple means of expression; clear process for regular feedback; possible use of lists (numbered or bulleted) and readable blocked text	**Moderately Accessible** Designed considering some UDL principles, ADA-compliant fonts, and color schemes; some overuse of bold or colors for formatting and, if document is in MS word, no use of "headings feature"; no use of multimodal components; hyperlinks may or may not be embedded correctly; no options for multiple means of expression or semi-clear process for regular feedback; possible use of bulleted lists, numbers, or outlines	**Not Accessible** No evidence of UDL thinking; commits one or more of these errors: non-ADA-compliant fonts, overuse of bold or difficult-to-read colors that would pose issues for colorblind or visually impaired students or screen readers, no use of headings features on documents if composed in MS Word, links inserted with full URL; no mention of feedback; no use of bulleted lists, numbers, or outlines; assignment prompt may be only a single broad question

Applicability (Disciplinary, Institutional or Professional): I agree that this writing assignment considers applicability.

Table 5

Applicability *Scoring Rubric*

3 (Strongly Agree)	2 (Agree)	1 (Strongly Disagree)
Highly Applicable Assignment is framed in a disciplinary or professional context, is clearly scaffolded with other assignments or readings from the course, and explicitly lists or articulates program and/or course outcomes; clearly connects this assignment to a student's goals in college or in their eventual career, discipline, or industry	**Somewhat Applicable** Assignment is vaguely framed in a disciplinary or professional context, but connections or goals are not specific; there is mention of a required reading to be used in the assignment, but it is unclear how the assignment itself works within the larger ecosystem of a program of study, the institution, or the professional world	**Not Applicable** Assignment seems to exist in isolation from other course/program materials; is not clear how it fits into sequence of course assignments; no mention of course/learning goals or outcomes; no mention of relationship to other works in the field or what students should "gain" from this activity (unclear "why")

Actionability: I agree that this writing assignment considers actionability.

Table 6

Actionability *Scoring Rubric*

3 (Strongly Agree)	2 (Agree)	1 (Strongly Disagree)
Highly Actionable Tasks are clearly outlined with due dates and times—including dates for peer review sessions and/or instructor feedback deadlines; drafting steps are made apparent in an easy-to-follow sequence; submission protocols (e.g., file type, where/how to upload) are easy to find	**Mostly Actionable** Describes tasks in paragraph form but no lists or clear due dates and/or times at the top of the text; lists or emphasizes penalties for late submission and/or due dates but does not discuss dates for peer review or revision activities; does not mention either acceptable file formats or programs students will use to upload (e.g., Blackboard, SafeAssign, Turnitin)	**Not Actionable** Vaguely describes the task, mentions a final due date but does not include a time; clear that students would need to obtain information elsewhere regarding opportunities for peer feedback, resources, or due dates (assignment text suggests or indicates this information is posted elsewhere or was discussed only in class); focuses mostly on penalties; does not mention submission format or how to turn in

Situational Clarity: I agree that this writing assignment considers situational clarity.

Table 7

Situational Clarity *Scoring Rubric*

3 (Strongly Agree)	2 (Agree)	1 (Strongly Disagree)
Very Clear Genre and audience are explicitly articulated/defined in the context of this assignment or other course readings *and are not simply mentioned or named*; multiple or unique genres are explained, and examples are provided; conventions or style guides (APA, MLA, etc.) are mentioned as features of a particular discipline, genre, or discourse community	**Mostly Clear** Genre and audience are mostly articulated/defined in the context of this assignment or other course readings *and are not simply mentioned or named*; multiple or unique genres are explained, and examples are provided; conventions or style guides (APA, MLA, etc.) are mentioned as features of a particular discipline, genre, or discourse community. The instructor uses the word "genre" or "audience" but does not provide clarifying explanation or examples	**Unclear** No mention of audience or attempt to define the rhetorical situation; offers only a single question as a prompt; no examples are provided

4.2.2 Annotation Protocol Training

The holistic assessment was designed by the first author (study lead) in consultation with the second author (P.I.) and a writing studies expert consultant.[11] Three research assistants (H1, H2, and H3) with experience annotating linguistic features in texts served as raters. The research assistants were trained by the study lead to annotate and score the set of curricular texts according to the rubrics. An annotation protocol was developed following rater training and calibration where rater agreement was computed regularly to ensure raters had sufficient agreement.

4.2.2.1 Rater Agreement. In preparation for the training, raters first reviewed training materials and protocol documentation composed by the study lead alongside the final scoring rubrics. Raters were trained over two, non-consecutive seven-hour days during a single week. Training was conducted remotely by the study lead via video conference. Following the initial two days of training, raters scored a practice set of ten writing curriculum texts. The practice texts were pulled from publicly available course materials online and included syllabi, assignment sheets, and rubrics from six institutions that reflected the kinds of curricular texts

[11] We thank Dr. Norbert Elliot for his guidance in helping us to develop and refine these rubrics.

raters would be annotating for the study. In addition, the study lead's assignments from her teaching were used for training. Practice documents were de-identified before being shared with raters via a shared Google Drive folder.

During training, raters reviewed the full set of training materials and then completed rating activities. Raters were asked to take notes throughout the training. These notes were shared with the study lead after each training session for the purpose of logging the training process and updating the protocol.[12] The study lead also took notes to keep track of questions each rater had regarding the scoring procedure and met with the P.I. regularly to discuss progress. All training documentation was stored in a shared Google Drive folder that was accessible to the P.I., the raters, and the study lead.

During Week 1 of training, raters scored the ten practice texts and recorded scores for each practice text along with notes and questions to discuss during a follow-up call with the study lead. The practice texts, rater scores, and notes were discussed during a two-hour video conference that took place two weeks after the initial training sessions. The study lead also scored the practice texts as the Human Expert (HE) rater; these scores were shared with the raters, and any discrepancies between raters were discussed. Raters shared their reasoning for scores given in each rubric category. In Week 2, each rater's notes from this follow-up session were shared with the study lead and compiled with scores into a shared spreadsheet for reference.

Once raters were trained, the study lead sent the first "batch" of six texts from the larger set of 48 to the raters for scoring. These texts were labeled "Batch 1," and the texts therein were labeled texts "A-F." Once agreement was calculated for Batch 1 scores, the study lead observed score discrepancies in some categories between raters. Score discrepancies were defined as two points difference in scoring (e.g., H1 scored a 1 and H3 scored a 3). To address these discrepancies, the study lead and the P.I. decided to include Batch 1 as an additional training set in order to recalibrate raters, which meant that scores for Batch 1 would be revealed to raters and discussed amongst them. Texts A-F represented authentic documents from study sites, so we determined that their differences from the initial training documents might have led to discrepancies in scoring. Upon consideration as of writing this article, we feel that this might also speak to the general observations of widespread inconsistencies in curricular texts, which make it difficult to calibrate raters and produce usable statistical data (we discuss this further in section 6.0).

In a recalibration video conference, the study lead and the raters discussed scoring discrepancies, and raters shared their scores and notes for texts A-F. Raters were instructed to use this set of texts as examples to discuss as proxies for challenging texts in subsequent batches. Batch 1 scores and notes were made available in a spreadsheet in the shared Google Drive folder. During recalibration, the study lead observed that *situational clarity* was the most difficult category to score for raters. She then held one-on-one discussions with each rater via video

[12] This was an "anticipatory" step, i.e., anticipating if there were discrepancies so the study lead could go back and review rater notes to talk through issues after the first 10 practice documents were scored.

conferences to better understand rater difficulties that were both procedural and conceptual. These meetings provided insights into whether discrepancies were due to misunderstandings or other factors related to conceptual perspective. Rater notes were then reviewed by the study lead and addressed with each rater in terms of what might have influenced discrepant scores. Notably, these discussions revealed that some raters were more inclined to provide lower scores based on additional conceptual factors, illustrated in the following quotes from raters documented in notes:

- "I've had some guilt about how few assignments are shifting over to agree or strongly agree, [and I'm] not sure how to curb the 'this is the best one I've seen in a while' bias."
- "I need to remind myself that the rubrics were made for all assignments and not just for this sample; I need to score them individually and not relatively."
- "I have additional training in accessibility, so I think I am inclined to score these relatively low for accessibility."
- "As a parent, when I look at my son's assignment, if some minor details aren't clear, I can't figure it out, but if the situational clarity isn't there . . . maybe content becomes more important in my scoring."

It was observed that these kinds of biases could create a "perfect storm" for disagreement and that raters were scoring relative to distribution of perceived quality rather than the set standard. Galdas (2017) discusses how recognizing research bias is crucial for determining the utility of study results but questions the issue of bias in qualitative research given that it is a paradigm drawn from quantitative research. In the course of conducting a mixed qualitative/quantitative study like ours, "managing bias" does not always directly translate between qualitative and quantitative methods. Desire to manage bias, as Galdas (2017) points out, is a challenge for qualitative research because of disciplinary pressures to demonstrate research outputs that lead to quantifiable impact. With this in mind, we felt it was important to have discussions with raters that might help them be more aware of biases in their scoring and to use these examples as a means of understanding how biases influenced their scores in order to attend to both the qualitative and quantitative aims of our study.

It was determined, following additional discussion and training, that raters were sufficiently trained based on agreements calculated in Table 8 and could proceed with scoring the remaining batches.

Table 8

Agreement for Added Practice Set of 6 Texts (Batch 1, Texts "A-F")

	Accessibility		***Applicability***		***Actionability***		***Situational Clarity***		***Holistic***	
Agreement	**exact (ex)**	**exact + adjacent (adj)**	**ex**	**ex+adj**	**ex**	**ex+adj**	**ex**	**ex+adj**	**ex**	**ex+adj**
H1/H2	50%	100%	70%	100%	90%	100%	70%	100%	90%	100%
H2/H3	80%	100%	70%	100%	100%	100%	70%	100%	90%	100%
H3/H1	70%	100%	80%	100%	90%	100%	90%	90%	100%	100%
H1/HE	70%	100%	80%	100%	80%	100%	80%	100%	90%	100%
H2/HE	80%	100%	70%	100%	70%	100%	90%	100%	100%	100%
H3/HE	60%	100%	80%	100%	70%	100%	80%	100%	90%	100%

4.2.3 Application of Scoring Rubrics

During scoring of the curricular text dataset[13] in subsequent weeks following this initial two-week training period, batches were small (approximately six to ten texts per batch) and were separated into four batches total. Each text included in a batch corresponded to a single document type (a syllabus that contained assignment descriptions, an assignment grading rubric, or an assignment instruction sheet) that was included in the instructor's complete "text package" provided to researchers. A "text package" could have contained one or more documents (e.g., a syllabus, rubric, or assignment sheet) associated with a given writing task. While the majority of assignments corresponded to a single text type, there was not always a one-to-one match between writing tasks (e.g., essays students were expected to write for a class) and curricular text types (i.e., the documents created by instructors to convey expectations for tasks to students). In some cases, one assignment might have been discussed across multiple text types (e.g., both within the context of a syllabus and in a separate assignment sheet), which together would constitute a text package. Conversely, an instructor may have referred to multiple writing tasks within a single curricular text, as was the case for syllabi. In all cases, raters assigned a score to each unique curricular text document rather than scoring at the text package level. All curricular texts were de-identified so that no information was available about the site or instructor.

Batches were delivered to raters over two months in a labeled .zip folder with a specified completion deadline. Based on their other work priorities, raters scored the curricular texts in each batch over one to two weeks per batch. Each rater was provided with a blank Excel spreadsheet with separate pages labeled with each individual rating category which corresponded to the scoring rubrics (i.e., *accessibility*, *applicability*, *actionability*, *situational clarity*) and the

[13] The scoring dataset was now *N*=42, reduced from *N*=48 due to Texts A-F becoming additional training documents used during rater recalibration.

holistic (*overall quality*) score. The spreadsheet contained three columns for each category: "TextID," "Rating," and "Comments." Raters were instructed to record scores for each text in each batch and return their spreadsheets to the study lead by certain deadlines. Raters could not see each other's scores. Following lessons learned from Batch 1, instructions were given to raters to write comments when they felt unsure about the score they were giving or to offer justification for a particular score.

For each curricular text in the set of 42, the analytic rubrics required raters to provide a Likert scale score for each trait on a scale of 1 to 3 and a holistic score (*overall quality*) on a scale of 1 to 4. A text could receive a higher or lower overall holistic score than the scores given for each individual trait scored in the analytic rubrics. Raters were instructed to give more weight in the holistic score to *actionability* and *situational clarity* in determining *overall quality*. These two traits were identified by the study lead as more influential for the holistic assessment based on training discussions.

After each batch was scored, the first author calculated exact and exact plus agreement in Excel between each pair of raters (H1 and H2, H1 and H3, H2 and H3) and her ratings as HE. The study lead created a "Master Notes and Updates to Protocol" document to address questions that were not clearly outlined in initial training protocol documentation, which could be referenced as raters worked to complete scoring. Raters had continued access to this document via the shared Google Drive folder. The folder also contained the practice texts, final scoring rubrics, training documentation, and the spreadsheet of practice text scores and notes from each rater. Raters were expected to consult this documentation during scoring if they needed reminders about scoring decisions or if they needed to recalibrate.

5.0 Results

Once the full dataset was scored, we computed rater agreement (see Tables 9-13) and quadratic weighted kappa (QWK) between all rater pairs and all rating categories (see Table 14). Scores from the study lead were used as the "gold standard" ratings (Human Expert, or HE). The HE updated her original scores using the final version of the rubrics to ensure that her scores included the updated protocol and rubric criteria. HE-o indicates original ratings from the study lead, and HE-f indicates the final ratings of the study lead after recalibrating.

Table 9

Agreement for Accessibility *in All 5 Rubric Categories Between H1, H2, H3, and HE-f*

Raters	**Agreement excluding training set (Texts A-F)**			
Accessibility				
H1/H2	exact	0.69 (69%)		
	adjacent	0.30 (30%)		
	discrepant	0 (0%)		
	exact+adjacent	0.99 (99%)		
H1/H3	exact	0.83 (83%)		
	adjacent	0.14 (14%)		
	discrepant	0.02 (2%)		
	exact+adjacent	0.97 (97%)		
H2/H3	exact	0.64 (64%)		
	adjacent	0.33 (33%)		
	discrepant	0.02 (2%)		
	exact+adjacent	0.97 (97%)		
	Agreement excluding training set (Texts A-F)		**Agreement for all texts (*N*=48)**	
H1/HE-f	exact	0.73 (73%)	exact	0.68 (68%)
	adjacent	0.26 (26%)	adjacent	0.31 (31%)
	discrepant	0 (0%)	discrepant	0 (0%)
	exact+adjacent	0.99 (99%)	exact+adjacent	0.99 (99%)
H2/HE-f	exact	0.95 (95%)	exact	0.95 (95%)
	adjacent	0.04 (4%)	adjacent	0.04 (4%)
	discrepant	0 (0%)	discrepant	0 (0%)
	exact+adjacent	0.99 (99%)	exact+adjacent	0.99 (99%)
H3/HE-f	exact	0.64 (64%)	exact	0.62 (62%)
	adjacent	0.33 (33%)	adjacent	0.35 (35%)
	discrepant	0.02 (2%)	discrepant	0.02 (2%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.97 (97%)

Table 10

Agreement for Applicability *in All 5 Rubric Categories Between H1, H2, H3, and HE-f*

Raters	**Agreement excluding training set (Texts A-F)**			
Applicability				
H1/H2	exact	0.54 (54%)		
	adjacent	0.38 (38%)		
	discrepant	0.07 (7%)		
	exact+adjacent	0.92 (92%)		
H1/H3	exact	0.73 (73%)		
	adjacent	0.26 (26%)		
	discrepant	0 (0%)		
	exact+adjacent	0.88 (88%)		
H2/H3	exact	0.71 (71%)		
	adjacent	0.21 (21%)		
	discrepant	0.07 (7%)		
	exact+adjacent	0.92 (92%)		
	Agreement excluding training set (Texts A-F)		**Agreement for all texts (*N*=48)**	
H1/HE-f	exact	0.64 (64%)	exact	0.60 (60%)
	adjacent	0.30 (30%)	adjacent	0.33 (33%)
	discrepant	0.04 (4%)	discrepant	0.06 (6%)
	exact+adjacent	0.94 (94%)	exact+adjacent	0.93 (93%)
H2/HE-f	exact	0.79 (79%)	exact	0.80 (80%)
	adjacent	0.18 (18%)	adjacent	0.16 (16%)
	discrepant	0.02 (2%)	discrepant	0.02 (2%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.96 (96%)
H3/HE-f	exact	0.70 (70%)	exact	0.73 (73%)
	adjacent	0.22 (22%)	adjacent	0.19 (19%)
	discrepant	0.06 (6%)	discrepant	0.07 (7%)
	exact+adjacent	0.92 (92%)	exact+adjacent	0.92 (92%)

Table 11

Agreement for Actionability *in All 5 Rubric Categories Between H1, H2, H3, and HE-f*

Raters	**Agreement excluding training set (Texts A-F)**			
Actionability				
H1/H2	exact	0.57 (57%)		
	adjacent	0.38 (38%)		
	discrepant	0.04 (4%)		
	exact+adjacent	0.95 (95%)		
H1/H3	exact	0.54 (54%)		
	adjacent	0.45 (45%)		
	discrepant	0 (0%)		
	exact+adjacent	0.99 (99%)		
H2/H3	exact	0.59 (59%)		
	adjacent	0.40 (40%)		
	discrepant	0 (0%)		
	exact+adjacent	0.99 (99%)		
	Agreement excluding training set (Texts A-F)		**Agreement for all texts (*N*=48)**	
H1/HE-f	exact	0.54 (54%)	exact	0.58 (58%)
	adjacent	0.40 (40%)	adjacent	0.37 (37%)
	discrepant	0.04 (4%)	discrepant	0.04 (4%)
	exact+adjacent	0.94 (94%)	exact+adjacent	0.95 (95%)
H2/HE-f	exact	0.71 (71%)	exact	0.72 (72%)
	adjacent	0.26 (26%)	adjacent	0.25 (25%)
	discrepant	0.02 (2%)	discrepant	0.02 (2%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.99 (99%)
H3/HE-f	exact	0.71 (71%)	exact	0.72 (72%)
	adjacent	0.28 (28%)	adjacent	0.27 (27%)
	discrepant	0 (0%)	discrepant	0 (0%)
	exact+adjacent	0.99 (99%)	exact+adjacent	0.99 (99%)

Table 12

Agreement for Situational Clarity *in All 5 Rubric Categories Between H1, H2, H3, and HE-f*

Raters	**Agreement excluding training set (Texts A-F)**			
Situational Clarity				
H1/H2	exact	0.54 (54%)		
	adjacent	0.45 (45%)		
	discrepant	0 (0%)		
	exact+adjacent	0.99 (99%)		
H1/H3	exact	0.64 (64%)		
	adjacent	0.30 (30%)		
	discrepant	0.04 (4%)		
	exact+adjacent	0.94 (94%)		
H2/H3	exact	0.61 (61%)		
	adjacent	0.35 (35%)		
	discrepant	0.02 (2%)		
	exact+adjacent	0.96 (96%)		
	Agreement excluding training set (Texts A-F)		**Agreement for all texts (*N*=48)**	
H1/HE-f	exact	0.64 (64%)	exact	0.66 (66%)
	adjacent	0.33 (33%)	adjacent	0.31 (31%)
	discrepant	0.02 (2%)	discrepant	0.02 (2%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.97 (97%)
H2/HE-f	exact	0.78 (78%)	exact	0.79 (79%)
	adjacent	0.21 (21%)	adjacent	0.20 (20%)
	discrepant	0 (0%)	discrepant	0 (0%)
	exact+adjacent	0.99 (99%)	exact+adjacent	0.99 (99%)
H3/HE-f	exact	0.69 (69%)	exact	0.66 (66%)
	adjacent	0.28 (28%)	adjacent	0.29 (29%)
	discrepant	0.02 (2%)	discrepant	0.04 (4%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.95 (95%)

Table 13

Agreement for Overall Quality *Between H1, H2, H3, and HE-f*

Raters	**Agreement excluding training set (Texts A-F)**			
Overall Quality				
H1/H2	exact	0.21 (21%)		
	adjacent	0.52 (52%)		
	discrepant	0.26 (26%)		
	exact+adjacent	0.73 (73%)		
H1/H3	exact	0.47 (47%)		
	adjacent	0.42 (42%)		
	discrepant	0.10 (10%)		
	exact+adjacent	0.89 (89%)		
H2/H3	exact	0.47 (47%)		
	adjacent	0.42 (42%)		
	discrepant	0.10 (10%)		
	exact+adjacent	0.89 (89%)		
	Agreement excluding training set (Texts A-F)		**Agreement for all texts (*N*=48)**	
H1/HE-f	exact	0.50 (50%)	exact	0.54 (54%)
	adjacent	0.38 (38%)	adjacent	0.35 (35%)
	discrepant	0.11 (11%)	discrepant	0.10 (10%)
	exact+adjacent	0.88 (88%)	exact+adjacent	0.89 (89%)
H2/HE-f	exact	0.47 (47%)	exact	0.50 (50%)
	adjacent	0.50 (50%)	adjacent	0.47 (47%)
	discrepant	0.02 (2%)	discrepant	0.02 (2%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.99 (99%)
H3/HE-f	exact	0.62 (62%)	exact	0.60 (60%)
	adjacent	0.35 (35%)	adjacent	0.35 (35%)
	discrepant	0.02 (2%)	discrepant	0.04 (4%)
	exact+adjacent	0.97 (97%)	exact+adjacent	0.95 (95%)

Table 14

*QWK (Strength of Agreement) Excluding Training Set (*N*=42)*

Rubric	H1/H2	H1/H3	H1/HE-o	H1/HE-f	H2/H3	H2/HE-o	H2/HE-f	H3/HE-o	H3/HE-f	HE-o/HE-f
Accessibility	0.150	0.389	0.115	0.115	0.166	0.659	0.659	0.062	0.062	1.000
Applicability	0.406	0.662	0.454	0.533	0.610	0.798	0.790	0.561	0.541	0.868
Situational Clarity	0.431	0.339	0.523	0.520	0.459	0.780	0.743	0.592	0.596	0.977
Actionability	0.434	0.480	0.377	0.419	0.494	0.588	0.616	0.611	0.648	0.894
Overall Quality	0.010	0.013	0.325	0.194	0.430	0.586	0.647	0.620	0.554	0.763

The strongest agreement overall was between H2 and HE-f, where all agreement was in the moderate range, 0.60-0.79. All other interrater agreement was either weak (0.40-0.59), minimal (0.21-0.39), or random (0.0-0.2). Across all rater pairs (excluding HE-o and HE-f), *accessibility* and *overall quality* scores had the lowest interrater agreement. The *applicability* category seemed to have the highest agreement between weak to moderate across the board, followed by *situational clarity*. There was no agreement in the strong range (0.80-1.0), with the exception of HE-o and HE-f.

As a final step, the study lead calculated mean scores for all curricular texts in the five rubric categories. Across the board, mean scores were quite low across all six institutional sites. The mean scores for all (*N*=48) assignments are as follows:

- *Accessibility:* 1.2 out of a possible 3
- *Applicability:* 1.6 out of a possible 3
- *Actionability:* 1.6 out of a possible 3
- *Situational Clarity:* 1.5 out of a possible 3
- *Overall Quality:* 1.9 out of a possible 4

The third and fourth authors (data analysts who worked on student writing feature measurements of the larger IES study) determined that QWK calculations were not usable for application to the collected writing sample feature data due to the small sample sizes for correlations once curricular texts were aggregated to the course level and separated into assignment types. In other words, we did not have a large enough curricular dataset to produce reliable results even though our attempt implies results might be usable if this method is applied to a larger dataset with a clearer path to aggregate texts to the course level.

These results attest to the importance of considering what insights can be gained from *not* being able to make certain statistical claims. The attempt to render usable statistical data from our dataset provides an opportunity to consider why things did not work out as we had hoped. Both the rating process and its results demonstrate how difficult it is to reach agreement on what

makes a writing assignment "good" and can aid in conversations about what we actually value when we talk about effective curriculum that contributes to student success. What we learned in our attempts to generate usable results is that in order to understand the impact of curriculum on student writing achievement, we need more data, and we need to consider more proactively in the initial phases of exploratory study design how different types of data might speak to each other.

6.0 Discussion

Creating equitable learning conditions for students approaching situated writing tasks requires careful consideration of design, formatting, and language choice in writing curriculum. While much work has already been done to identify, classify, and measure features of student writing (as referenced in sections 1.0 and 2.0), less has been done to consider how presence of these student writing features might correspond to how the writing tasks themselves are framed to students in delivered curriculum and therefore how this framing might affect student writing achievement. This is likely because doing so quickly gets complex—as our study demonstrates—and "few researchers have access" to the kinds of corpora they would need to make reliable claims about the complex "relationship between discourse and task design during students' transition into college-level writing" (Aull, 2017, p. 11). This suggests that many students are left struggling to understand what is being asked of them when they approach college writing tasks.

We observed that, overall, many curricular texts were not as clear or as comprehensive as they could have been. There was not a lot of apparent proactive application of even basic UDL principles, which puts neurodivergent and disabled student writers at a material disadvantage. Essentially, our exploration of local site-specific materials suggests there is much room for improvement in curricular text design and much to consider when applying these observations to faculty development as it pertains to composing and developing curriculum. While classroom instruction is cited as something that fills in gaps in learning (Camburn & Han, 2011; Lerner, 2019), we could not see evidence of that from the curricular texts in this study. Arguably, delivered curriculum *should* demonstrate expectations clearly, because not every student attends all class sessions, and it is difficult to know (without being physically or digitally present during a class session) how effectively an instructor communicates information and whether that communication works for students of all abilities.

This is especially relevant now during the COVID-19 pandemic, where new instructional modalities rely heavily on digital platforms and necessitate careful, proactive decision-making about curriculum design and delivery. Moreover, not every institution requires instructors to submit assignment sheets/instructions or grading rubrics alongside syllabi, so it is unclear whether this kind of data exists at most institutions.

While there are efforts currently underway to construct accessible and digital public archives of curricular data (Oddis et al., 2020; the FYC Archive[14]; the Dartmouth '66 Seminar Exhibit[15]; the Open Syllabus Project[16]), these data have not yet been applied to research projects that seek to better understand relationships between curricular features like those we identified and specific features measured in studies of student writing achievement. We were also limited in what we could do with our data because we did not have access to instructor grades or scores for given assignments.

There are some obvious limitations in this study that should be discussed to help others who may want to pursue this type of research in the future. First, some instructors did not opt to share their curricular texts. Further, instructors provided different sets of texts that articulated only some parts of writing tasks, so we did not have consistent data from each instructor. One instructor may have provided a grading rubric, syllabus, and assignment sheet, while another may have provided only the assignment sheet. In addition, we had no information about whether instructors also posted examples of "successful" writing task completion on Blackboard/Canvas/another learning management system (LMS) or on other course sites (e.g., Digication or Google Classroom). Therefore, only the texts provided by instructors could be coded and scored using the rubrics we developed.

We also attribute our results (and lack thereof) to how we classified texts by grouping text types under a large umbrella of "curriculum" conceived broadly. Our decision to include three types of curricular texts—syllabi, assignment sheets, and grading rubrics—in a larger dataset was made because of the already limited availability of texts to score. We believe that our framework might render more usable results when curricular texts are treated separately rather than as part of a larger set of generalized "curriculum." We wonder to what extent we would see certain scores as higher or lower overall in different feature categories for different types of curricular texts, and so, we suggest that researchers think proactively about collecting curricular materials in the earliest stages of study design—enough to render usable data in each *separate* category of curriculum (i.e., a set of syllabi, a set of assignment sheets/instructions, and a set of assignment grading rubrics). We also see potential for applying rubrics to text packages, which is another potential avenue for writing analytics researchers applying a "mix" of qualitative and quantitative analysis methods. Finally, individual instructors might use these rubrics as a means of critically analyzing and improving their own instructional design in ways that could be mapped onto larger institutional outcomes and goals.

In the final phase of this study, following annotation training and rater scoring of curricular texts, our team attempted to answer the question of how much the presence or absence of detected curricular features affects student writing achievement. Ideally, to investigate this, correlations between curricular features (as identified through qualitative analysis) and student

[14] See the First-Year Composition Archive: https://fyca.colostate.edu/
[15] See the Dartmouth '66 Seminar Exhibit: https://wac.colostate.edu/resources/research/dartmouth/
[16] See the Open Syllabus Project: https://opensyllabus.org/

writing quality[17] (as identified by NLP) would be explored. However, our small dataset meant that statistical analyses could not produce usable results for many reasons—the first of which can be attributed to oversights in initial data collection and communicated requirements for site partner participation in the study. In the initial stages of the larger exploratory IES study, the original research team was not specific about materials to provide that described the assignments used to generate student writing samples. Many faculty did not provide materials related to writing assignments because they were not required to for participation, and these materials may have included assignment instructions, assignment scoring rubrics, and syllabi that contained additional assignment descriptions that would have contributed to a larger dataset.

The limited number of curricular documents collected and differing curricular text types submitted for different courses meant that our team did not have the necessary data to do what we ultimately hoped to do. To generate more usable results for our desired outcomes, curricular text types would have to be considered separately for statistical analysis. In our study, this requirement severely limited the number of courses that could be included in statistical calculations. Additionally, the ideal analyses of this data would be on the assignment level, demonstrating the association of curricular features of a particular assignment's instructions to student writing quality on the corresponding completed assignment. However, our team was also not provided with consistent labeling for each assignment it received from students, so data could only be matched on the course level. This further limited our sample as well as introduced additional sources of error, as a student's average writing scores across different assignments could vary by assignment. For this reason, to further investigate relationships between curricular features and student writing features, a single type of curricular document for each assignment as well as the resulting assignment submissions should be carefully collected and matched early on. Then, under this research design, correlations between curricular document measures and measures of student writing quality for these assignments could be explored.

The inability to have access to this level of data is a serious potential limitation to the productive work that writing analytics researchers might otherwise be able to do and requires deeper consideration of how and why we protect and provide certain forms of data. We identify a need to critically discuss what data we make available for research and how scholars might collaborate with research teams that have access to greater resources and larger and more comprehensive datasets. This limitation also provides a rationale for encouraging more instructors to share their curricular materials—not to leave them open to admonishment, but instead to more deeply consider how we can improve writing curriculum and help each other negotiate the challenges of teaching writing as our institutional and global realities evolve. We see a great need to consider more deeply how the data instructors and programs create in the form of curricular texts might better serve researchers in understanding how and why students write in response to situated tasks.

[17] See Burdick et al. (2013) for a discussion of validity of computer-analytic developmental scales that measure constructs of "writing ability" and "quality."

A natural question going forward is: How much does the presence or absence of identifiable curricular features affect student writing achievement? The small number of assignments and the lack of links between assignments and individual student writing samples did not support investigating this question fully with this study's data. Such analyses are important for future research in writing analytics and writing studies more broadly. Studies seeking to address this question could evaluate assignments and syllabi using the rubrics developed here and compare the scores of students' written responses to their corresponding curriculum. Such studies would need to include students in a sufficient number of courses to allow course-level analysis of the impact of syllabi or other curricular texts. They would also need the ability to link students' written work to specific assignments. A challenge to such studies is the writing evaluation conducted by instructors. Course grades would be a potential source of evaluation, but differences in grading policies could add error to any analysis that uses grades. Automated writing evaluation (AWE), in which computer algorithms are used to evaluate various features of writing samples, would provide a means of standardized evaluation of responses.

It is worth considering how a framework similar to what we have outlined in this article might be applied not only to curricular texts but to pedagogical texts as well. We can define "pedagogical texts" as artifacts which capture what *actually happens* at local writing sites beyond what is expressed in delivered curriculum. Examples of pedagogical texts which might be explored using a similar method could include recorded/transcribed lectures, program meeting notes and minutes, or online chat transcripts from writing center sessions (e.g., Lerner & Oddis, forthcoming). Applying a mixed method approach to analyzing pedagogical texts might also help illuminate aspects of the relationship between intended, hidden, and extra-curriculum (Gere, 1994; Lerner, 2019), which would be highly beneficial to the fields of writing studies and writing analytics. Our current pandemic moment is undoubtedly generating all kinds of pedagogical texts created through use of various digital tools and platforms. Digital pedagogical texts can offer unprecedented insight into how curriculum and pedagogy intersect with student writing features—especially because digital tools *enable* certain pedagogies, all of which have curricular content (i.e., knowledge that is generated and applied through diverse uptakes and experiences).

There are many ways in which curriculum and pedagogy intertwine within and across institutional writing sites; however, it is crucial that researchers going forward also understand that one is not a stand-in for the other. Lerner (2019) notes that "literature intended to represent the collected knowledge of the field" gives "short shrift" to "writing curriculum in comparison to writing pedagogy" (p. 7). Lerner's observation of the field's "reluctance" to address curriculum is also potentially consequential for researchers given that mistaking pedagogical uptakes of curriculum *as* curriculum can reinforce harmful assumptions about teaching and learning writing across disciplines. For instance, inattention in research to curricular features reinforces the idea that any adept pupil will simply be able to interpret and enact curriculum in response to a "question of the will" (Salvatori, 1996, p. 35). When students are left trying to "will" themselves into understanding what is expected of them by various stakeholders simultaneously, affected

also are the instructors and administrators tasked with helping these students improve or meet standards and broader outcomes. For example, as Aull (2017) concludes in a corpus analysis of writing tasks, "A-graded students are responding, consciously or not, to the discursive demands of different tasks" which are "genre-specific," and these "discourse choices help contribute to the students' writing success" (p. 33). This all adds up to a sobering realization: We often create conditions in institutions that do not create the consistency that many students need to succeed.

The degree to which any research team addresses the potential consequences of this lack of access to both curricular and pedagogical data might explain some of the highly variable rates of student "success" in college writing that continue to mystify many (often public) voices in higher education. Researchers have opportunities to consider in future study designs how to collect more robust data which better considers and reflects the realities of teaching and learning writing at the college level and accounts more comprehensively for the relationships between discourse and task design to observable features in student writing. Abbott and McKinney (2012) emphasize that "good" research means careful and accessible approaches to integrating research design and statistics, and our study—despite its limitations—is one which we believe moves toward realizing this goal.

Elliot et al. (2013) demonstrate how, in considering the evolution of writing assessment, it is not enough to outright reject the reality of automated writing evaluation and its use and impact on higher education. The researchers who build AWE tools often have more resources to conduct the high-impact research that many WPAs simply do not have the time, funding, or desire to do (Strickland, 2011). Tool developers working to build platforms like ETS's Writing Mentor[18] can also benefit by designing a methodology in the early stages of their research that more dynamically considers the relationship between student writing, curriculum, and pedagogical uptakes of curriculum in the development of tools \ that assist students in completing their situated writing tasks. Finally, there is potential to more closely examine whether writing tasks as articulated in assignment sheets, for instance, show more potential for transfer within and across certain disciplines and across institutions.

7.0 Conclusions

The findings, limitations, and challenges of this study reveal that while we still have much to learn about the relationship between features measured in student writing and features of curriculum, there are many promising opportunities for writing analytics researchers to develop and apply new methods which will help us better grasp these relationships. Researchers interested in supporting student achievement in writing should not only collect data from analysis of student writing; they should make curricular texts (syllabi, assignment sheets, and grading rubrics) *and* pedagogical texts (e.g., recorded/transcribed lectures, presentations, lesson plans, classroom activities, observations, and interviews) part of evidence-centered techniques

[18] See Writing Mentor: https://mentormywriting.org/

during the earliest stages of study design, which could help control for local site-oriented relevance. In order for writing analytics, writing/composition studies, and education researchers to work together more effectively, we need to learn from what curriculum has to teach, and we need to know what happens alongside curriculum on a pedagogical level that accounts for what is often hidden on the local stage from a more global view.

One portion of the larger IES study in which this study is situated revealed that in response to a student survey question "How much did the writing assignment help you to explore or to better understand the topic that you wrote about?" 43.6 percent of students responded "Not at All" and 37.2 percent responded "Somewhat." Fewer than 15 percent of students said the assignment actually helped them "a lot," and our study sought answers to why that might be. Further work needs to be done to determine if there is a relationship between genres of assignments and helpfulness of delivered curriculum for diverse learners, but for now, this initial exploration demonstrates a sobering reality: There is little consistency in how instructors frame writing tasks.

We suggest that a "good" assignment is not just an assignment that demonstrates awareness of threshold concepts in any discipline. In fact, while nearly all assignment texts examined in this study appeared to value a writing studies threshold concept in some way, concepts did not have a lot of bearing on factors like usability or accessibility, which often have a lot more to do with a student's ability to complete a task based on elements such as formatting and design of the curricular document itself. While nearly all of the 48 assignment texts in our study demonstrated some evidence of threshold concepts, few showed much evidence of UDL thinking or indicated explicit evidence of transferability—at least, not in a way that was measurable.

When student writing achievement is studied based on feature presence in student writing samples *alone*, researchers miss an important opportunity to validate the power of the relationship between teaching and learning through quantitative means. There are varying degrees of resistance to applying quantitative methods to help make these claims. For example, Goldstone and Underwood (2014) discuss the "antagonism toward counting" that has historically defined literary studies. Many writing programs are housed under the banner of English departments, so it can be challenging from different disciplinary perspectives to assert the potential of attempting to "quantify" teaching/learning relationships regarding what content is taught about writing. The fact remains, however, that higher educational institutions distribute resources based on what *is* presented in a quantifiable way, particularly with regard to metrics like retention (Dougherty & Reddy, 2013; Yi, 2019).

For those unwilling to acquiesce to this, there are other reasons to attend to the usefulness of analyzing curriculum that align with many of the disciplinary values often articulated by writing/composition scholars and instructors. Heard (2014) asserts that greater attention to curriculum design can foster "desirable habits of composition" which "encourage inventiveness (innovation) and creativity" (p. 315) not just for students, but for instructors, and especially for junior scholars and faculty. In Heard's view, to foster design as an essential part of instructors' professional development is an "act of invention—an act that prolongs our engaged inquiry into

the values, habits, and assumptions we practice as students and teachers" (2014, p. 316). This is undoubtedly consequential for how students subsequently learn to practice, inquire, and invent in their writing beyond classroom walls (or screens). Heard makes a strong case for the importance of greater attention to design of curricular materials along these lines, and many others cited in this article detail what design thinking can offer writing studies as a field. Based on what our study team uncovered over the duration of our research, we agree with an insight offered by Lerner (2019): "inattention to curriculum ultimately hampers our effort to enact meaningful reform and to have an impact on larger conversations about education and writing."

Writing analytics offers a potential home for this kind of work in its commitment to multidisciplinarity, which extends to partnerships between writing studies scholars and educational research organizations that put our methods in conversation, as the authors of this article did. While in many ways, this study did not provide the kinds of results we were hoping for in terms of reliable statistical data, our study strongly points toward the need for more of this type of work. We invite researchers to use or augment our rubrics or develop their own based on shared sets of values. We feel researchers working together in writing analytics are particularly equipped and positioned to do this type of work, especially on collaborative research teams whose approaches are importantly and necessarily informed by diverse methodologies and complementary values which center on a desire to help students achieve "success" in writing—whether that means simply meeting the requirements of the writing tasks designed for them or discovering that they are capable writers whose contributions to knowledge-making are essential for the advancement of teaching and learning.

Acknowledgments

We would like to thank the advisory board and site partners for the IES study through which the data for our research was obtained. We would like to specifically thank Dr. Norbert Elliot for guidance in helping to refine and develop the curriculum scoring rubrics used in this study and Dr. Mya Poe for connecting the study lead to researchers at Educational Testing Service. We recognize the importance of mentorship and facilitating connections between multi-institutional and organizational researchers as essential for encouraging new generations of scholars to conduct collaborative research. Finally, we thank participants of the 2021 Writing Analytics Virtual Symposium who attended the presentation of this study's findings from locations around the world and who offered insight and feedback on the direction of this article.

Author Biographies

Kyle Oddis is a Ph.D. candidate in writing studies and the project manager for the Northeastern University Writing Program Digital Public Archive. Her research explores evolving roles of writing assessment, curriculum, and pedagogy within frameworks of institutional and organizational values. Focusing also on neurodiversity and design, Kyle's work foregrounds

proactive and innovative approaches to the teaching and learning of writing and encourages cross-institutional/organizational collaboration in multidisciplinary teams.

Jill Burstein, Ph.D. is a Principal Assessment Scientist at Duolingo. Jill conducts assessment innovation research for The Duolingo English Test in the Assessment Research area, and co-leads a validity research team. Jill's research career has been motivated by a strong interest to make a social impact through equitable and widely accessible education technology. Her interests lie at the intersection of artificial intelligence (AI) and natural language processing (NLP), educational measurement, equity in education, language assessment, and linguistics. As a leader in AI in education, Jill has led NLP teams that have invented automated writing evaluation systems used in large-scale, high-stakes assessment and digital writing support applications. She has led federally funded writing analytics research exploring relationships between writing achievement and postsecondary academic success outcomes. Jill holds a B.A. in linguistics and Spanish from New York University and M.A. and Ph.D. degrees in linguistics from the Graduate Center, City University of New York.

Dan McCaffrey is Associate Vice President for Psychometric Analysis and Research at Educational Testing Service (ETS). He conducts research on automated writing evaluation and automated scoring of constructed responses, causal modeling, and measuring student growth in addition to overseeing the psychometric and data analysis for ETS's testing programs and contracts.

Steven Holtzman serves as a Principal Research Data Analyst at Educational Testing Service. He has an M.A. in statistics and a B.A in statistics and economics from Boston University. His work at ETS concentrates on using study design, data collection, data management and data analysis methods to help promote research in education. He has also co-authored numerous publications and presented at many conferences. His recent projects have examined noncognitive assessments, writing skills, teacher evaluation, and workforce selection assessments.

References

Abbot, M. L., & McKinney, J. (2012). *Understanding and applying research design.* Wiley & Sons.

Adler-Kassner, L., & Wardle, E. (Eds.). (2016). *Naming what we know, classroom edition: Threshold concepts of writing studies*. Utah State University Press.

Anderson, P. V. (2010). The benefit and challenges of adopting a new standpoint while assessing technical communication programs: A response to Jo Allen. In M. N. Hundleby & J. Allen (Eds.), *Assessment in technical and professional communication* (pp. 57-64). Routledge.

Andersson, M. A. (2018). Higher education, bigger networks? Differences by family socioeconomic background and network measures. *Socius.* https://doi.org/10.1177/2378023118797217

Anson, C. (1993). *Writing across the curriculum: An annotated bibliography.* Greenwood.

Anson, I. G., Moskovitz, C., & Anson, C. M. (2019). A text-analytic method for identifying text recycling in STEM research reports. *The Journal of Writing Analytics*, *3*, 1-47.

Aull, L. L. (2017). Corpus analysis of argumentative versus explanatory discourse in writing task genres. *The Journal of Writing Analytics*, *1*, 125-150.

Aull, L. L. (2020). *How students write: A linguistic analysis*. Modern Language Association.

Baglieri, S. (2020). Toward inclusive education? Focusing a critical lens on universal design for learning. *Canadian Journal of Disability Studies*, *9*(5), 42-74.

Bawarshi, A. (2000). The genre function. *College English, 62*(3), 335–-360.

Bawarshi, A. (2003). *Genre and the invention of the writer: Reconsidering the place of invention in composition.* University Press of Colorado; Utah State University Press.

Bers, T. H. (2013). Deciphering articulation and state/system policies and agreements. *New Directions for Higher Education*, *162*, 17-26.

Brewer, E., Selfe, C. L., & Yergeau, M. (2014). Creating a culture of access in composition studies. *Composition Studies*, *42*(2), 151-154.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.

Bromley, P., Northway, K., & Schonberg, E. (2013). how important is the local, really? A cross-institutional quantitative assessment of frequently asked questions in writing center exit surveys. *The Writing Center Journal*, *33*(1), 13-37.

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (2013). Measuring students' writing ability on a computer-analytic developmental scale: An exploratory validity study. *Literacy Research and Instruction*, *52*(4), 255-280.

Burke, D. D., Clapper, D., & McRae, D. (2016). Accessible online instruction for students with disabilities: Federal imperatives and the challenge of compliance. *Journal of Law & Education*, *45*(2), 135-180.

Burstein, J., McCaffrey, D., Beigman Klebanov, B., & Ling, G. (2017). Exploring relationships between writing & broader outcomes with automated writing evaluation. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 101-108). Association for Computational Linguistics.

Camburn, E. M., & Han, S. W. (2011). Two decades of generalizable evidence on U.S. instruction from national surveys. *Teachers College Record*, *113*(3), 561-610.

Collier, D., LaPorte, J., & Seawright, J. (2012). Putting typologies to work: Concept formation, measurement, and analytic rigor. *Political Research Quarterly*, *65*(1), 217-232.

Condon, W., Iverson, E. R., Manduca, C., Rutz, C., & Willett, G. (2016). *Faculty development and student learning: Assessing the connections*. Indiana University Press.

Denny, H., & Towle, B. (2017). Braving the waters of class: Performance, intersectionality, and the policing of working class identity in everyday writing centers. *The Peer Review*, *1*(2). https://thepeerreview-iwca.org/issues/braver-spaces/braving-the-waters-of-class-performance-intersectionality-and-the-policing-of-working-class-identity-in-everyday-writing-centers/

Devitt, A. J. (2004). *Writing genres*. Southern Illinois University Press.

Dougherty, K. J., & Reddy, V. (2013). *Performance funding for higher education: What are the mechanisms? What are the impacts? ASHE Higher Education Report*. Wiley & Sons.

Duff, D. (2002). Intertextuality versus genre theory: Bakhtin, Kristeva and the question of genre. *Paragraph, 25*(1), 54–73.

Duffy, J. (2019). *Provocations of virtue: Rhetoric, ethics, and the teaching of writing*. Utah State University Press.

Duffy, J., & Agnew, L. (Eds.). (2020). *After Plato: Rhetoric, ethics, and the teaching of writing*. Utah State University Press.

Elliot, N. (2016). A theory of ethics for writing assessment. *The Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/36t565mm

Elliot, N., Gere, A. R., Gibson, G., Toth, C., Whithaus, C., & Presswood, A. (2013). Uses and limitations of automated writing evaluation software. *WPA-CompPile Research Bibliographies*, *23*. https://comppile.org/wpa/bibliographies/Bib23/AutoWritingEvaluation.pdf

Fisher, E. A., & Wright, V. H. (2010). Improving online course design through usability testing. *MERLOT Journal of Online Learning and Teaching*, *6*(1), 228-245.

Freadman, A. (1987). Anyone for Tennis? In I. Reid (Ed.), *The place of genre in learning: current debates* (pp. 91–124). Deakin University (Australia): Centre for Studies in Literary Education.

Freadman, A. (2002). Uptake. In R. M. Coe, L. Lingard & T. Teslenko (Eds.), *The rhetoric and ideology of genre: strategies for stability and change* (pp. 39-53). Hampton Press.Frederiksen, N., Saunders, D. R., & Wand, B. (1957). The in-basket test. *Psychological Monographs: General and Applied*, *71*(9), 1-28.

Galdas, P. (2017). Revisiting bias in qualitative research: Reflections on its relationship with funding and impact. *International Journal of Qualitative Methods.* SAGE.

Gallagher, C. W. (2010). Assess locally, validate globally: Heuristics for validating local writing assessments. *WPA: Writing Program Administration*, *34*(1), 10-32.

Gallagher, C. W. (2014). All writing assessment is local. *College Composition and Communication*, *65*(3), 486-505.

Geller, A. E., Eodice, M., & Lerner, N. (2016). *The meaningful writing project: Learning, teaching, and writing in higher education*. Utah State University Press.

Gere, A. R. (1994). Kitchen tables and rented rooms: The extracurriculum of composition. *College Composition and Communication*, *45*(1), 75-92.

Gere, A. R. (2019). Knowledge making and writing analytics: MLA special session. *The Journal of Writing Analytics*, *3*, 312-316.

Goldstone, A., & Underwood, T. (2014). The quiet transformations of literary studies: What thirteen thousand scholars could tell us. *New Literary History*, *45*(3), 359-384.

Gonsher, I. (2016) Beyond design thinking: An incomplete design taxonomy. *Critical Design Critical Futures*. http://www.cd-cf.org/articles/beyond-design-thinking/

Grouling, J. (2018). Training writing teachers: An assignment in mapping writing program values. *Prompt: A Journal of Academic Writing Assignments*, *2*(1). https://doi.org/10.31719/pjaw.v2i1.16

Gürüz, K., & Zimpher, N. L. (2011). *Higher education and international student mobility in the global knowledge economy.* State University of New York Press.

Haswell, R. H. (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Greenwood Publishing Group.

Heard, M. (2014). Repositioning curriculum design: Broadening the who and how of curricular invention. *College English*, *76*(4), 315-336.

Heugh, K. (2014). Epistemologies in multilingual education: Translanguaging and genre—companions in conversation with policy and practice. *Language and Education*, *29*(3), 280-285.

Isaacs, E. (2018). *Writing at the State U: Instruction and administration at 106 comprehensive universities*. Utah State University Press.

Jo, A. (2010). Mapping institutional values and the technical communication curriculum: A strategy for grounding assessment. In M. N. Hundleby & J. Allen (Eds.), *Assessment in technical and professional communication* (pp. 39-56). Routledge.

Lang, S., Aull, L., & Marcellino, W. (2019). A taxonomy for writing analytics. *The Journal of Writing Analytics, 3,* 13–37.

Lerner, N. (2019). *Reformers, teachers, writers: Curricular and pedagogical inquiries*. University Press of Colorado.

Lerner, N., & Oddis, K. (forthcoming). "Hi, no worries at all!": Rhetorical listening as expression of emotional knowledge in online synchronous writing conferences. In J. Morris & K. Concannon (Eds.), *Affect and emotion in the writing center*. Parlor Press.

Li, J., & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing Writing*, 26, 67-79.

Ling, G., Elliot, N., Burstein, J. C., McCaffrey, D. F., MacArthur, C. A., & Holtzman, S. (2021). Writing motivation: A validation study of self-judgment and performance. *Assessing Writing*, *48,* 1-15.

Marcellino, W. (2020). Book review of Aull, L. (2020). *How students write: A linguistic analysis*. The Modern Language Association of America. *The Journal of Writing Analytics*, *4*, 243-245.

McComiskey, B. (2012). Bridging the divide: The (puzzling) *Framework* and the transition from K–12 to college writing instruction. *College English*, *74*(6), 537-540.

Melzer, D. (2014). *Assignments across the curriculum: A national study of college writing*. Utah State University Press.

Meyer, A., Rose, D. H., & Gordon, D. (2014). *Universal design for learning: Theory and practice*. Cast Incorporated.

Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech, 70,* 151-167.

Moxley, J., Elliot, N., Eubanks, D., Vezzu, M., Elliot, S., & Allen, W. (2017). Writing analytics: Conceptualization of a multidisciplinary field*. The Journal of Writing Analytics*, *1*, v–xvii.

Newman, D., & Dickinson, M. (2017). Preparing students for success in hybrid learning environments with academic resource centers. *New Directions for Teaching and Learning*, *149*, 79-88.

Null, R. (2013). *Universal design.* CRC Press.

Oddis, K., Blankenship, A., Lanham, B., & Lerner, N. (2020). Possibilities of a public-facing digital archive in the age of analytics. *The Journal of Writing Analytics*, *4*, 159-192.

Ortiz, L. A. (2020). Reframing neurodiversity as competitive advantage: Opportunities, challenges, and resources for business and professional communication educators. *Business and Professional Communication Quarterly*, *83*(3), 261-284.

Osorio, R. (2020). A disability-as-insight approach to multimodal assessment, 26-36. In D. Kelly-Riley & N. Elliot (Eds.), *Improving outcomes: Disciplinary writing, local assessment, and the aim of fairness* (pp. 26-36). Modern Language Association.

Payne, B., Cigularova, D., Costanzo, J., Das, S., Mann, M., & Perez, K. (2021). Impact of articulation agreements on student transfer between higher education institutions: A case study of a cybersecurity program. *Community College Journal of Research and Practice*. https://doi.org/10.1080/10668926.2021.1887007

Poe, M. (2019). Research in the teaching of English: From alchemy and science to methodological plurality. *The Journal of Writing Analytics*, *3,* 317-333.

Porter J., Sullivan, P., Blythe, S., Grabill J., & Miles, L. (2000). Institutional critique: A rhetorical methodology for change. *College Composition and Communication*, *51*(4), 610-642.

Purdy, J. P. (2014). What can design thinking offer writing studies? *College Composition and Communication*, *65*(4), 612-641.

Ringler, H., Klebanov, B. B., & Kaufer, D. (2018). Placing writing tasks in local and global contexts: The case of argumentative writing. *The Journal of Writing Analytics*, *2*, 34-77.

Rose, S. K., & Weiser, I. (2002). The WPA as researcher and archivist. In L. L. Gaillet, H. D. Eidson, & D. Gammill (Eds.), *Landmark essays on archival research* (pp. 145-155). Routledge.

Rose, S. K., & Weiser, I. (2018). *The internationalization of U.S. writing program*s. Utah State University Press.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *The Journal of Writing Analytics*, *1*, 227-257.

Rourke, L., & Zhou, X. (2019). When scores do not increase: Notes on quantitative approaches to writing assessment. *The Journal of Writing Analytics*, *3*, 264-285.

Saldaña, J. (2015). *The coding manual for qualitative researchers*. Arizona State University Press.

Salvatori, M. R. (1996). *Pedagogy: Disturbing history, 1819-1929*. University of Pittsburgh Press.

Scott, L. A., Thoma, C. A., Puglia, L., Temple, P., & D'Aguilar, A. (2017). Implementing a UDL framework: A study of current personnel preparation practices. *Intellectual and Developmental Disabilities, 55*(1), 25–36.

Sharer, W. B., Morse, T. A., Eble, M. F., & Banks, W. (Eds.). (2016). *Reclaiming accountability: Improving writing programs through accreditation and large-scale assessments*. Utah State University Press.

Sheffield, J. P. (2018). More than a useful myth: A case study of design thinking for writing across the curriculum program innovation. *The WAC Journal*, *28*, 168-188.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Erlbaum.

Shkoler, O., Rabenu, E., Hackett, P. M. W, & Capobianco, P. M. (2020). *International student mobility and access to higher education.* Palgrave Macmillan.

Strickland, D. (2011). *The managerial unconscious in the history of composition studies*. Southern Illinois University Press.

Stubblefield, M. L. (2016). *International student mobility in higher education: Examining decision making factors for international students choosing to study in the Regional University System of Oklahoma* [Doctoral dissertation, University of Oklahoma]. ProQuest Dissertations Publishing.

Sullivan, D. F., & McConnell, K. D. (2018). It's the assignments—A ubiquitous and inexpensive strategy to significantly improve higher-order learning. *Change: The Magazine of Higher Learning, 50*(5), 16-23.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Swales, J. M. (2017). The concept of discourse community: Some recent personal history. *Composition Forum, 27.* https://compositionforum.com/issue/37/swales-retrospective.php

Tannenbaum, R. J., & Katz, I. R. (2021). Validity considerations in complex task design. *The Journal of Writing Analytics*, *5*, 196-214.

Tomlinson, E., & Newman, S. (2017). Valuing writers from a neurodiversity perspective: Integrating new research on autism spectrum disorder into composition pedagogy. *Composition Studies*, *45*(2), 91-273.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design.* Association for Supervision and Curriculum Development.

Worsham, R., DeSantis, A. L., Whatley, M., Johnson, K. R., & Jaeger, A. J. (2021). Early effects of North Carolina's Comprehensive Articulation Agreement on credit accumulation among community college transfer students. *Research in Higher Education, 62*(7), 942-975.

Yergeau, M. R. (n.d.). Reason. *Kairos*, *18*(1). https://kairos.technorhetoric.net/18.1/coverweb/yergeau-et-al/pages/reason/index.html

Yi, A. Y. (2019). The weight of the metric: Performance funding and the retention of historically underserved students. *The Journal of Higher Education*, *90*(6), 965-991.

Zhang-Wu, Q. (2020). Teaching hybrid online college composition classes to international students during COVID-19: Equity, diversity, inclusiveness, and community building. *English Leadership Quarterly*, *43*(2), 9-13.

Appendix A: Qualitative Analysis Codebook

Definitions for each individual code are provided below within the context of their overarching categories and their relationship to the three theoretical frameworks.

Genre Group Codes

AudienceMention refers to an instance in which the instructor uses the word *audience* directly but does not define who that audience is (for example, "show that you understand your audience"); who the audience is supposed to be is not immediately clear.

AudienceStatement, on the other hand, refers to an instance in which the instructor explicitly lists the audience (or multiple audiences) and offers some kind of definition or indication of who the student should have in mind while writing (for example, "you are writing a letter *to me* . . . " or "your audience is yourself and your peers").

Complexity is used to code instances where the instructor introduces information that might confuse a student's existing understanding of the rhetorical situation (for example, the instructor asks the student to write a letter, and then calls it an essay and requires a thesis statement; letters don't typically require thesis statements). This code is also used in instances where a professor articulates a new genre that is a hybrid of multiple genres, or changes the rhetorical situation to include ambiguity (for example, "you may do this, but it is not required").

The **Example** code indicates an instance in which an instructor provides a direct example of a successful or unsuccessful attempt at some aspect of the assignment; e.g., they include an example of a poorly constructed thesis statement or offer a sample outline of how the student's text should be composed ("this is what a clear thesis statement looks like").

The **Features** code is used in instances where an instructor enumerates or lists conventions of a genre or articulates elements that typically constitute a given rhetorical situation (for example, "Traditionally when people write memoirs, they relate a chronological series of important events . . . ").

The **SituationDefined** code, on the other hand, indicates when the instructor defines a situation by outlining its unique contours in context of a particular course; the defined situation *does not always* map onto what has been articulated as features of a given genre. To use the previous example, the Features code would be applied to a sentence in an assignment text that describes what is *typical* of a memoir; the SituationDefined code, on the other hand, would be applied in an instance where the instructor followed that up with something like, "The memoirs we will be writing are going to be a little different. Instead of speaking just about life events, you will also have to x, y, z . . . " Defining a rhetorical situation for a genre helps students understand how the writing assignment they are completing maps onto or differs from existing genres, which are defined by certain features.

The **SourceUse** code refers to the source the instructor is using in the assignment text to situate the assignment or help articulate the genre in which the student is being asked to work

(for example, "I'm going to explain what I mean using some quotes from this person . . . "). In other words, they are using an external source to help explain what they are asking students to do.

Finally, the **Thesis** code is used simply to indicate whether or not the instructor invokes a "thesis statement" as a component of the assignment. Not all genres require thesis statements. By naming a thesis statement as an important element of writing, the instructor is setting a standard that the student is being asked to meet, whether or not that alters a student's understanding of a genre or rhetorical situation. This code may or may not be present where Genre Complexity is present.

A Note on Genre Complexity

Among the most important codes identified within the Genre group is Complexity. Complexity is generally understood by scholars in the field as a *given* when discussing genre because genres don't exist in static forms or ways (Devitt, 2004); in other words, there is no such thing as a genre without complexity. That said, examples of assignments that ask students to write a letter *and also an essay at the same time* are potentially confusing for students who *don't* have the same scholarly level of understanding of genre discussed earlier (this is a reflection of instructors holding tacit knowledge). This can be hindering to writing task completion, particularly in the absence of an example that allows students to understand how to respond to the given/articulated rhetorical situation. If a rhetorical situation is unfamiliar, students may not meet the *undemonstrated* standards and may not understand how to correct their mistakes because they do not have the vocabulary of knowledge to know what they are doing "wrong." This is complicated even more when instructors add the requirement to integrate sources when source use is also usually contingent upon understanding rhetorical situation *within a disciplinary context*.

It is not possible to eliminate Complexity, and that is not the recommendation being made here; instead, what the ubiquity of Complexity suggests is that there is a lot more going on when students are asked to write an "essay" than what might be clear on the surface of an assignment. How students *interpret* what they're being asked to do could hold a lot of weight in their ultimate ability to successfully execute a writing task. One instructor's definition of a memoir, for example, could be entirely different from a student's prior knowledge understanding of the memoir genre; if the instructor isn't articulating the particulars of *their* rhetorical situation—if they are not providing clear examples or outlining a clear process through which students can demonstrate comprehension—students that *already* have a more nuanced understanding of genre can pull farther ahead. Depending upon a student's socioeconomic status or educational history, that prior knowledge can vary *widely*—that's why it's helpful to look at assignment texts across multiple institutions as we have done here.

Task Completion Group Codes

The **Feedback** code is used to indicate instances in which the instructor discusses the modes of feedback the student will receive; for example, if there are peer review sessions, draft stages with formative instructor comments (not letter grades), or a writing center appointment requirement. The Feedback code indicates if there is a clear *process* through which students must move while completing a given task.

The **Grading** code indicates where instructors have provided actual numerical or lettered scoring criteria, such as "a strong thesis statement is worth one point." The Grading code also indicates where instructors have specified penalties for lateness or bonus points for extra credit, as these elements all directly contribute to a final grade received on any given assignment and help students understand what the instructor considers to be most important in assessing a writing task.

The **Parameter** code is used for instances where an instructor has provided some kind of word count or formatting requirement that is based on either personal preference or a disciplinary standard (for example, "700-1,000 words in length" or "Times New Roman font").

The **Resource** code is applied when the instructor provides a link to book a writing center or academic resource center appointment, or when the instructor references guides that may be located on Blackboard or a course site that exist to help a student get the task done (for example, "See *this thing* posted on Blackboard to help you with *this*.").

The last code in this category is **SourceText.** SourceText, which is different from SourceUse in the Genre group, refers to the source texts the *student* needs to use to complete the task (as opposed to a source the instructor is using to assist them in articulating a genre). A SourceText instance would be coded when an instructor says something like, "you will need to use quotes from Maya Angelou's essay."

Transfer Group Codes

This group has only two codes. The first is **Outcomes**, which is used in instances where the instructor has connected an assignment with a goal or an outcome that is listed as central to success in a student's program of study, as part of the writing program's goals, or framed as some larger societal goal (as in, success beyond the classroom) which is generally connected to some larger institutional outcome that will have been previously defined. This could be mapped onto a concept of *vertical transfer*, for example, which requires transferring a lower-level skill to a higher level of cognition.

Finally, the **Scaffold** code is used to indicate where an instructor has explicitly stated how this assignment fits into the architecture of their course or alongside other assignments in the local classroom context (for example, "you will draw on the skills you developed in the first assignment to complete this next assignment"). This could be mapped onto a concept of *horizontal transfer*, which is more sequential and happens in the same context (in this case, the classroom).