A ZOOMABLE ASSESSMENT: NAVIGATING THE ECOLOGIES OF WRITING PROGRAM ASSESSMENT

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Matthew Warner

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
English

May 2022

Fargo, North Dakota

North Dakota State University
Graduate School

Title

A Zoomable Assessment: Navigating the Ecologies of Writing Program Assessment

By

Matthew Warner

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Bruce Maylath
Chair

Sean Burt

Enrico Sassi

Megan Orr

Approved:

July, 11 2022
Date

Sean Burt
Department Chair

ABSTRACT

This dissertation project explores the potential for using an inferential statistics test (t-tests) within an existing writing program assessment design. The purpose of using inferential statistics is to provide several perspectives on a data set collected using the existing assessment design thereby improving what a writing program administrator can learn about the program. To demonstrate the use of statistical tests, I selected as a variable of interest participation in an international collaboration, the Trans-Atlantic and Pacific Project (TAPP). Based on this variable, I asked, can inferential statistics identify whether participation in TAPP created a difference in student portfolio scores for a program outcome? To perform the t-test, I calculated the mean portfolio scores for TAPP and for Non-TAPP groups. Then, after sorting the program data by course, two courses, a writing in the health professions and a writing in the technical professions, had enough sections participate in TAPP to conduct two more tests, one for each course. The tests posed the same question, whether participation in TAPP had a difference in portfolio scores for a program outcome, but had zoomed from the program level into the course level.

The tests indicated that at the highest level (the program) participation in TAPP did not have a statistically significant difference on portfolio scores. The tests at the other level (the course) indicated that participation in TAPP did not have a statistically significant difference on writing in the health professions but did have a statistically significant difference on writing in the technical professions. Possible explanations for these results are examined in relation to existing writing studies literature.

The approach of examining several levels is dubbed a zoomable assessment because the statistical tests allow for more nuanced examinations, that is, the tests zoom into the data set.

Based on the findings, I propose further uses and possible limits of uses of inferential statistics as a complement to existing assessment designs. As part of the proposal, I advocate for assessment design, such as zoomable assessment, that is accessible, meaning the design does not require special software or extensive knowledge of advanced statistical analysis methods.

ACKNOWLEDGMENTS

I would like to thank Bruce Maylath and Enrico Sassi for their support, guidance, and patience over the many, many years that they have supported me with this project; Sean Burt for taking a plunge to learn about writing program assessment; Megan Orr for being a wonderful statistics teacher; MK Laughlin for years of friendship; and a small group of academic Twitter followers who offered insights and tips about scholarship and projects related to writing programs and writing program assessment. I would like to thank NDSU for facilitating my work, especial the English Department and a variety of committees, and the NDSU Center for Writers for offering a Disquisition Boot Camp that motivated me to complete large portions of the writing during the summer.

DEDICATION

Dedicated to my family, Richard, Mary, Kevin, and Kate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IKI	Integrating knowledge and ideas (the programmatic outcome).
TAPP	Trans-Atlantic and Pacific Project.
UDW	Upper division writing program.
WPA	Writing program administrator.

1. INTRODUCTION

For writing program administrators (WPA), the assessment of a writing program is often fraught with uncertainties and complications. The underlying purposes of an assessment might range from appeasing administrators or accreditors to helping ensure that courses and pedagogy are beneficial for students. The broad range of purposes has proliferated the development of assessment designs and theories. Each design and theory purports to provide insight into a program and satisfy an underlying purpose. While the proliferation of designs and theories has addressed some needs of WPAs, the abundance has also exacerbated some problems for WPAs. The abundance can produce decision paralysis about how to design an assessment and what can be part of the design.

An emerging component of writing program assessment is to include statistical analysis as part of or as the entire assessment design. Reasons to include statistical analysis vary, from external entities seeking specific kinds of evidence to a WPA's personal interest to try a new approach in order to learn more about the program. Complications that WPAs can experience when incorporating statistical analysis include worries about if they need to revisit math courses or learn computer programming, or even what to know about statistics to include the analysis in the assessment design. Statistical approaches to writing program assessment exist, but the approaches require knowledge of sophisticated statistical tests, such as principal component analysis and multiple regression analysis (White et al. 2015). In the context of working as a WPA, the more sophisticated tools could dissuade WPAs from incorporating statistical analysis. I also believe embedding advanced statistical tests into the assessment design might create unsustainable assessments because a program might change WPAs and the new WPA might not know how to perform the statistical analysis or have the resources to learn how to perform the analysis. In this dissertation, I demonstrate that inferential statistical analysis can be beneficial for WPAs to target specific aspects of a writing program when conducting program assessment. Furthermore, I argue that inferential statistical analysis can be accessible, requiring minimal knowledge of statistics and statistical software, thereby avoiding some of the problems associated with more complex statistical analyses.

For the dissertation, I created a writing program assessment method, dubbed a zoomable assessment, that incorporates an inferential statistical analysis. The analysis targeted one instructional project, participation in the Trans-Atlantic and Pacific Project (TAPP). TAPP is incorporated into some upper division writing courses of the writing program. Consequently, two groups are available to analyze whether there is a difference in portfolio scores between the two, the TAPP-participating courses and the Non-TAPP participating courses. To illustrate how to apply a zoomable assessment, I collected student portfolio data in order to assess participation in the TAPP. The illustrated application shows that a WPA can modify an existing assessment method in one simple way (tracking two groups for an inferential statistical test) and gain insight into a pedagogical activity and, by extension, the writing program. However, before delving into the project, I want to provide a condensed overview of the origins of writing courses in US universities, writing assessment, and writing programs to set the backdrop for how the designs, theories, and purposes of writing assessment have developed and informed the dissertation project. While my dissertation focuses on determining whether there was a difference in portfolio scores based on participation in the TAPP, an accompanying purpose is to limit how much change to the assessment design is necessary in order to minimize increased workload for a WPA and allow for versatility in applying statistics to examine other facets of a writing program. In that sense, the dissertation is a project of maintenance addressing a gap in methodology within writing program assessment scholarship, which is part of an underlying argument that writing program assessment does not require extensive innovations in order to provide insights or to examine program performance, only creativity when applying the design.

1.1. Zoomed Out: Brief Origin Story of Writing Assessment

Before writing programs existed as a concept and before composition existed as a discipline, student writing was a source of disputes about learning and literacy. In *On a Scale* Norbert Eliot (2008) identifies a start to the disputes about student writing to be in 1894 when Barrett Wendell wrote a piece for *The Dial*, a popular literary criticism publication of the period, about university education in America. Wendell surveyed faculty from "venerable Eastern [United States] institutions" (p. 16) in order to understand what instruction was happening at these institutions. The survey offered only "bland analysis" (Eliot, 2008, p.21). Though the analysis was devoid of importance, Wendell still concluded that there was a need to investigate "'the evils of bad training' [in writing instruction]" (Eliot qtd. Wendell, p. 21). Thus, Wendell wrote an early example of a now familiar social commentary about education. The commentary follows a particular pattern that Robert Connors (1997) characterizes as a struggle between periods when reform of writing courses was prominent and periods when abolition of writing courses was prominent. According to Connors, the mood about writing in university settings vacillated between Reform and Abolition. Connors notes each period has distinguishing features, therefore not all reform periods are identical, and similarly, not all abolition periods are identical. However, common tropes occur during each type of period. Though Wendell only identifies a problem, because he concludes a need for change in writing courses rather than the removal of them, he would be a reformist.

Reform periods tended to follow instances of distress about the literacy (writing skills) of students. Though writing courses are the “thin red line protecting the very life of literacy” (Connors, 1997, p. 47), the courses require improvement in order to enable each student to be the best writer possible. A frequent solution to the perceived distress was to become more rigorous. Rigorous was a poorly defined term but served the purpose of assuring individuals that a solution was available. Rigor, therefore, tended to mean that writing instruction required adhering to highly prescriptive ideas of writing conventions. An early example of a rigorous solution to a perceived literacy problem occurred in 1885 at Harvard. Several admissions classes had what some administrators deemed to be low entry exam scores. The low scores became the source for a crisis. To address the lack of rigor, Harvard instituted a required course, English A, for all first-year students. The first reform period had started.

Harvard taught English A for five years before another group of scholars questioned the value of the requirement, not the course necessarily. English A was the sole required course for all first-year students, which contributed to the debate about who should teach it and what should be taught. It marked the first abolitionist period, when groups of people “declare the large sums expended on this all-but-ubiquitous course [first-year writing] a gross waste” (Connors, p. 47). The abolitionist group tends to be individuals within the university, often from units that do not teach writing courses but also from the instructors of writing courses. The group of abolitionists declaring English A to be a “gross waste” cited two concerns about a required writing course: “first, the required freshman course was never meant as a permanent English offering but was instead a temporary stopgap until the secondary schools could improve; and second, the teaching of required composition was tiresome, labor-intensive, and a bad use of trained literary scholars”

(Connors, 1997, p.48-9). [1] It was the 19th century instance of "blame the high schools," and an early acknowledgement that writing instruction is difficult – *labor-intensive* – and probably required skills not possessed by a literary scholar, trained in and more interested to teach criticism or hermeneutics. Furthermore, one group of writing-course abolitionists argued that writing instruction ought to be within the purview of specific disciplines because the scholars in a discipline would be best situated to teach about the writing of that discipline. This group would eventually be formalized into writing-in-the-discipline advocates (Bordelon, Wright, & Halloran, 2012). The reform-abolition cycle had completed a full iteration. An issue of literacy was identified, or arguably, manufactured (Elliot, 2008). A solution was proposed. The solution was questioned and an alternative suggested. The cycle reflects an interest in what constitutes writing and writing pedagogy. Is a course dedicated to writing necessary? Who should teach it? What should be taught? What constituted student success? Though the phrase did not appear amid the discussion, the cycle reflects the types of questions to ask when assessing a writing program.

And the cycle would iterate and continues to loop even to the present. For example, a contemporaneous iteration of the cycle has unfolded in *The Chronicle of Higher Education*. Joseph Teller (2016) wrote "Are We Teaching Composition All Wrong?", a version of the abolish argument; and Doug Hesse (2017) wrote a response, "We Know What Works in Teaching Composition," acknowledging writing courses have issues but nothing that reform

[1] I chuckle because the "lower scores" were not significantly lower than previous years, which some of the early abolition advocates maybe intuitively realized. The intuition that individuals who wanted a required course, the reformists, fabricated the crisis by encouraging a perception of low scores might have motivated the abolition advocates to push against English A. To my awareness, they did not include it as a reason, even after five years of English A being taught. There also is no indication whether English A improved the writing of Harvard students. A lack of inferential statistics has a long history.

cannot address. Teller (2016) argues writing instruction has erroneously followed three assumptions: writing involves a process, topics for writing ought to be complex issues with which students have a connection, and trends of literacy have combined reading and writing into one course when two separate courses should exist. Teller proposes writing pedagogy follow a high frequency model, in which students write a lot of short essays and receive feedback. The essays ought to follow a thesis-driven form of argumentation. The feedback should target specific elements of the thesis-driven argument model. In particular, Teller (2016) wants most reading removed from the writing course because "the more time a course focuses on 'critical reading' and content, the less time it spends on structure, argument, evidence, logical reasoning, and concise, clear prose — the tools a composition class should give undergraduates."

In the response, Hesse (2017) accuses Teller of misrepresenting the scholarship in writing studies and as evidence of the misrepresentation examines how the model Teller proposes ignores findings of writing studies scholarship, then outlines what the scholarship has established as effective practices. Hesse characterizes these practices as carefully sequenced tasks, "coach the process," readings as context and source material, and instruct students in all aspects of writing (genre, invention, grammar, style, logic, accuracy, and fitting a piece to audience). In the end, Hesse and Teller appear less at odds about what constitutes writing, but rather whether or not it is possible to determine if writing programs are successful and defining what constitutes evidence of success. The exchange between Teller and Hesse also illustrates an important reason to develop assessment tools that adapt to local contexts. Writing program administrators have no shortage of commentary regarding their programs. However, this commentary does not offer much insight when considering how or why to perform assessment. Either proposed approach to teaching writing requires an assessment design and informing assessment theory.

To think about reasons to perform assessment, I want to revisit the reported motivation of the survey of US universities by Wendell. In the 19th century, a desire to learn about how education was or was not structured was a response to The Grand Tour. For the Grand Tour, educators and scholars from America toured European universities aiming to learn about various subjects, but "what they uniformly absorbed was the form of the German university" (Eliot, 2008, p.4). The American education tourists perceived an impartiality in the German university because the structure (form) of university included precision through various measurements of learning. Consequently, the tourists returned advocating for precision in the US universities. Furthermore, the measurement held a promise of egalitarianism among students because the measurement would be clearly defined. Measurement as an impartial evaluation would eliminate or, at least, significantly decrease the ostensibly capricious subjectivity of feedback from instructors. A number was tidier than a comment, and the numbers would be calculated the same way for every piece of writing by any student. Thus, "testing – with its origin in the European science of psychological measurement – would become America's unique contribution to education" (Eliot, 2008, p. 4). Through the hope to structure learning around unbiased evaluations, an obsessive pursuit of finely tuned measurement seized the US education system – or, corrupted it (Watters, 2021).

The promise of measurement (assessment) needs reification, though. Measurement, in any form, requires the creation of rankings, scales, instruments, and all manner of methods and processes to make the abstraction into something concrete. The reification often takes the form of a number, a test score, or essay grade. The number becomes the access point to determine whether or not learning has occurred. The pedagogical practices linked to higher numbers, therefore, become presumed as the better practices. This 19th century idea flows through the majority of writing assessment approaches including the use of writing portfolios, which are the data source for this dissertation. Therefore, the dissertation has inherited a legacy feature from assessment design and theory: the numbers obtained through assessment designs indicate whether learning has happened.

The impulse to quantify learning began with behaviorist theories of psychology, notably Edward Thorndike (Gold, Hobbs, & Berlin, 2012; Elliot, 2008; Lynne, 2004[2]), which advocated that mental capacity was measurable. A major element of writing course abolition efforts that most histories of education testing and assessment discuss only implicitly is the distrust of educators or writing course instructors as a source for creating the rankings, scales, instruments, and so forth. Psychologists were presumed to have a better understanding of the development of the mind (Watters, 2021). Historically, the concern associated with educators creating a measurement process or choosing a metric tended to be a lack of objectivity or insufficient rigor in the method or metric (Hamp-Lyons, 2016; Strickland, 2011; Chenworth, et al., 1999). In part, this concern resulted in the emergence of what Stephen North (1987) dubs The Experimentalists within writing studies scholarship. North points to *Research in Written Composition* by Braddock, Lloyd-Jones, and Schoer in 1963 as the first instance of “the Experimental rubric” (p. 141) explicitly applied as a method of inquiry which purported that “there are no hidden features in the design [of research]; that data collection and analysis are not confused or fudged” (p. 155). In other words, writing research should and could be rigorously conducted, including the assessment of student writing and writing programs. I interpret the emergence of experimental

[2] Lynne (2004): In 1921, Edward L. Thorndike, a colleague of [Milo] Hillegas, demonstrated that the objective “mental” tests from the field of psychology were better *predictors* of college performance than the College Board’s existing essay entrance examinations or the student’s high school record (p. 27, emphasis added).

approaches as a manifestation of a type of writing course reform that, though emerging from within writing studies, distrusts that writing instructors (the Practitioners in North's nomenclature) possess the necessary knowledge and expertise to design and conduct assessments *unless* they apply scientific approaches (experimental). One consequence of this distrust is a dialectic has emerged between what Brian Huot (2002) dubs technological assessments and research-based assessments.

> Technological assessments involve an application of a set of methods developed by others and used across sites and contexts. Research-based assessments, on the other hand, require that the community of teachers, students and administrators come together to articulate a set of research questions about student performance, teaching, curriculum or whatever they are interested in knowing more about (Huot, p. 178)

In a way, the two types of assessment that Huot proposes return to the reform-abolition periods that Connors (1997) had identified. Sometimes the technological assessments have the advantage by purporting to remove biases. A technological assessment purports to deliver an objective and repeatable measurement across time and situational circumstances ("across sites and contexts"). Sometimes the research-based assessments have the advantage. A research-based assessment foregrounds the expertise of writing instructors and WPAs, and therefore positions them as able to ask the appropriate questions and identify the means to address the questions.

The debate about English A at Harvard, the vacillation between reform and abolition, and the friction between technological and research-based assessments, I argue, are culminations of two competing attitudes toward the place of writing in education, and by extension the resulting tools common in the design of writing program assessment: predictive tools (such as placement exams) and descriptive tools (such as program narratives). Ideally, the two types of tools harmonize, and the overall assessment of a program will involve both predictive and descriptive tools. However, I think underlying tensions of impartiality (predictive) and tacit working knowledge (descriptive) have kept the two approaches in intractable conflict. Administrators will distrust descriptive narratives seemingly devoid of concrete evidence. Instructors will distrust a spreadsheet populated with numbers seemingly devoid of meaning unless a person is versed in interpreting numeric data.

As often happens amid false dichotomies, other options exist beyond the two options most frequently invoked. For example, Donna Strickland in *The Managerial Unconscious* (2011) notes that composition studies as a field has struggled with a false dichotomy of narratives: tragic fall or romantic heroism. The tragic fall is a result of narratives “telling of the marginalization of teaching and writing in departments that privilege the interpretation of texts (criticism) over the production of texts (rhetoric) and thus the study of literature over the teaching of writing” (Strickland, 2011, p. 5). The observation transports us to the debate over English A in the 1890s and to the debate in *The Chronicle of Higher Education* between Hesse and Teller. In contrast, the romantic heroism is the narrative of “rescuing composition from its degraded and marginal status by repositioning the composition class as a unique site of democratic politics and pedagogical commitment” (Strickland, 2011, p. 6). And kaleidoscope-like, written composition, composition courses, and writing programs have morphed and continue to shapeshift due to the influence of these various theories and ideas about learning and knowing whether learning has occurred.

At this point, efforts to unpack writing program assessment involve more spectrums than are easily recalled: reform-abolition, criticism-rhetoric, technical-research, Experimentalists-Practitioners and others contained throughout the development and eventual emergence of writing studies and writing programs. It is similar to a nature-nurture debate but applied to writing skills, instruction, and pedagogy. The abolitionist view presumes writing is a natural skill and no writing-dedicated course can improve writing skills. In contrast, the reformists seek to find a configuration of instructional activities to support students to become better writers. The advocates of predictive tools suppose that there are inherent factors that will indicate success or failure as a writer, and these advocates also suppose that they can identify the factors even if they themselves are not writing studies scholars (Experimentalists). Writing instructors and WPAs create descriptive tools based on their experiences and accumulated knowledge to provide input for program formation and writing instruction. Though writing instructors and WPAs might use predictive tools, they lean toward experience as writers (Practitioners). Those individuals who want writing courses grounded in literary criticism presumed that enough reading will yield effective writers. In comparison, rhetoric scholars presumed that enough guided experience creating and delivering texts can improve any students' communication skills. Is it nature – inherent talent of students, factors tipping the scale of potential heavily in favor of some students to *be* effective writers? Is it nurture – refining instructional practices through the study of writing itself and realizing students can improve? At the crux of much of the debate resides a pair of questions: can writing be taught and how might we know if it has been effectively taught?

Of course, amidst the debate about whether predictive or descriptive, or abolishing writing courses or reforming them, alternatives exist that draw upon the various concepts informing writing course and writing program designs and theories. Much as Strickland (2011) sought to propose a third narrative lens, operative managerial reasoning, I hope to demonstrate that a third type of assessment tool will help writing program administrators to better understand

and discuss their programs. The proposed tool is an inferential statistical design, a zoomable assessment.

1.2. Zoomable Assessment

Among the many available assessment tools, inferential tools are a rarely discussed complement to the predictive and descriptive tools. Inferential tools have tremendous potential as an accessible statistical application, thereby helping writing program administrators to assess their programs by expanding what can be assessed and how it can be assessed. Furthermore, inferential tools do not require discarding an existing assessment design or even much modification to the assessment design, aside from categorizing collected data, such as I have done for the TAPP and the Non-TAPP courses. However, the predictive and descriptive tools have functioned as the basis for an abundance of available writing program assessment designs, which raises a question about why to propose another one simply because it uses inferential statistics as its basis. I certainly am not looking to inspire a revolution. I am too pragmatic to be revolutionary. Instead, the main reason is an effort to provide writing program administrators access to an option that is not widely implemented or, possibly, known. Writing program administrators will be familiar (or will become familiar upon entering the role) with using predictive tools to place students within a writing program or anticipating student performance in a writing course. Similarly, they will have familiarity with various descriptive tools to write a report about the writing program, whether it is a narrative about the program or providing descriptive statistics about trends on student portfolio scores.

However, these two types of tools create a gap in the types of questions that a WPA might ask about the program. Perhaps most pressing among the questions is whether or not an instructional approach has a detectable difference on sections of a course. To answer this question, a WPA needs inferential tools. Predictive tools rely upon working from “the past.” That is, based on the performance of students who had similar high school grades, choice of degree major, and household income, how might a student who has similar characteristics perform in a course? Descriptive tools create only an instance, a static impression of the current status of the program. These tools have roles within program assessment. Both of these have important uses. However, following Huot’s claim about research-based assessment, the more questions that WPAs can help people ask about the program, the more utility everyone might find in an assessment.

However, more questions might yield more problems when attempting to implement the methods to obtain answers. Therefore, a new design should not be unduly burdensome to use when attempting to obtain insights about the program. Questions might not have definitive answers, but the answers should not be cryptic requiring specialization in multiple areas of research. Furthermore, an assessment design ideally should not create new financial dependencies for the program, such as needing to purchase new software that requires license fees and introduce new constraints based on what data the software can obtain. Too often, a dependence on software creates what I dub a “data hostage” scenario in which program data is accessible only via the software, asking program administrators to either continue paying for the software or to risk losing collected data or struggle to migrate from one system to another one. The assessment design is confined, a “hostage” to the software. However, I want assessment to shift in perspective with greater ease in order to see bigger perspectives or minutiae of targeted interest. That is, I want it to zoom in focus to deliver what WPAs will want or need depending on their situation. The path is winding given the many factors identified, but the first step for a zoomable assessment is a clear demonstration of its potential, which is the purpose of this study.

1.2.1. A Working Definition of a Zoomable Assessment

A zoomable assessment seeks to address a gap in the methods used to conduct writing program assessment. The gap is that writing program assessment tends toward a narrowed focus on how to analyze collected data, or not analyze the data at all. Consequently, many writing program assessment designs have severe limitations regarding what a WPA can know about a program, such as only descriptive statistics or narratives about the program. The analysis and data types require methodological improvements. A zoomable assessment is a program assessment design that encourages an ensemble of tools and data to support WPAs by providing more and improved insights into a program and actionable points based upon the collected data and subsequent analysis. The ensemble of tools includes existing assessment methods, such as descriptive statistics and program narratives, but expands the group to include inferential tools. In comparison to predictive tools, an inferential tool examines program data for the purposes of identifying patterns in the data that might inform a variety of writing program features, including curricular design (how many courses), course design, professional development activities, and pedagogical activities – such as participation in TAPP.

The purpose for using an ensemble of tools is to provide changes in the resolution when examining the writing program, zooming in to examine courses or sections of a course or zooming out to examine patterns related to pedagogical activity throughout all courses and sections. Each tool enables a WPA to gain perspective, and in collaboration – descriptive tools, inferential tools, and predictive tools – the WPA can navigate from shifting perspective, zooming, from different points of view when assessing the writing program. In this dissertation project, I explore the initial boundaries of what a perspective shift, possible by adding one tool to the ensemble, an inferential statistical test, thereby illustrating how a writing program assessment can be zoomable.

1.3. Overview of Chapters

To make the case for a zoomable assessment, I begin in Chapter 2 by surveying the trends on writing program assessment then describe 'zooming' as an interdisciplinary trend that has a place in writing program assessment. The survey of writing program assessment categorizes the scholarship into three clusters: theories, designs, and trajectories. A reason to group writing program assessment scholarship into these three areas is that an assessment theory does not always yield the same type of design, and the trajectories do not always involve current theories or designs. The effort is to be as comprehensive as possible in the survey while steering toward the development of a zoomable assessment as a method. The brief overview of theories, designs, and trajectories examines how a zoomable assessment fits within this constellation of scholarship. Furthermore, I conclude by providing interdisciplinary support for the notion of 'zooming' or 'scale scholarship' within the humanities, for example "The Trans-Scalar Challenge of Ecology" by Zach Horton (2018) and the zoomable reading proposed by Ryan Cordell (2013).[3]

In Chapter 3, I transition the focus to the current study by describing the method used to collect the samples and to conduct an inferential statistical test based on the collected samples. For three academic years, I obtained portfolio scores from end of semester assessments. The

[3] Cordell, R. (2013). "Taken Possession of": The Reprinting and Reauthorship of Hawthorne's 'Celestial Railroad' in Antebellum Religious Press". *DHQ* 7(1). However, Dr. Cordell was also a visiting summer scholar at North Dakota State University, and he discussed the method of 'zoomable reading' of literary texts as part of a graduate seminar course. Aside from Cordell (2013), the phrase 'zoomable reading' is not widely applied, instead the common nomenclature is 'scalar reading' (Horton, 2018).

portfolio scoring follows a process that combines outcomes derived using dynamic criteria mapping (Broad, 2005) and scoring following conventions for multi-trait version of holistic scoring (Elbow, 1996; Elbow & Yancey, 1994). For my purposes, I tracked which portfolios were from sections that participated in a TAPP project. The selected statistical test is a *t-test* for differences in mean portfolio scores. Chapter 4 organizes the results of the test. The main patterns indicate trivial or no difference in most instances. However, the significant results indicate a non-trivial difference in portfolio scores for some courses, and by extension the role that a zoomable assessment serves within an existing assessment design. Finally, Chapter 5 establishes conclusions from the results and notes limitations as well as possible further research.

2. LITERATURE REVIEW: WRITING PROGRAM ASSESSMENT, SCALAR METHODS, AND INTERNATIONAL COLLABORATIONS

2.1. Introduction

In this chapter, I discuss previous research related to writing program assessment, the concept of scalar reading within the humanities which informed the decision to design a "zooming" assessment, and the Trans-Atlantic and Pacific Project (TAPP). The discussion of writing program assessment addresses what constitutes a writing program, how assessments have developed, and what trends are emerging to inform further developments in the design of writing program assessments. The discussion of scalar reading situates my concept of a zoomable assessment among the applications of and calls to apply scalar reading in a variety of contexts. Finally, the discussion of the TAPP describes typical activities during a collaboration and the findings of research of previous TAPP collaborations. The challenge of the literature review is to address the intersections of several scholarly areas (writing studies, writing program design, text interpretation, statistical analysis, international writing partnerships, and linguistics) informing the current dissertation project.

2.2. Writing Programs: Working Definitions, Existing Assessments, and Emerging Trends

No consensus exists for what constitutes a writing program, largely because the development and evolution of writing programs must respond to localized circumstances. However, the features of a writing program are generally considered to be writing courses, policies from the administering group of the program and from the university administration, professional development and personnel management, the students enrolled in the courses including procedures for placing them in accordance with policies, and program review. Writing program assessment factors into the program review, and often the two are conflated for one another. Program review examines the totality of the program, including the policies and procedures and instructor evaluation and professional development. The scope of a program review depends on how the university structures the writing program or how the WPA chooses or is informed to structure program review. Writing program assessment examines one facet of the overall program: course fulfillment of select outcomes, whether university outcomes, general education outcomes, departmental outcomes, or a combination. Other facets of the program through program review can have a part in the assessment (for example graduate instructor preparation), but the generally acknowledged convention is to understand program assessment as dedicated to whether or not what has happened in the courses within the writing program for a semester (or an academic year) has fulfilled the outcomes associated with the writing courses.

Even this understanding remains contentious within the scholarship about writing program administration. For example, John Brereton (1995) argued in *Composition Studies in American Colleges, 1875-1925* that an accurate understanding of student and instructor performance in fulfilling outcomes associated with writing tends to depend entirely upon assessment mechanisms addressing only first-year composition. "First-year composition came to dominate teaching and professional discourse about college writing instruction, but it in no way dominated writing" (Brereton, 1995, p. xvi). Consequently, any understanding of student writing and effectiveness of writing instruction addresses a narrowed view of all possible writing activity. Students often have to cope with competing ideas about writing. Lea and Street (1998) labeled this phenomenon course-switching, after the linguistic concept of code-switching. "In the case of 'course switching' students are having to interpret the writing requirements of different levels of academic activity. Such switching may range from academic disciplines in a traditional sense (such as physics and anthropology) to what we see as 'fields of study', such as modular programmes that incorporate elements of different disciplines and of interdisciplinarity" (Lea & Street, 1998, p. 161). While Lea and Street focused on higher education in the United Kingdom, Haswell (2009) identifies similar patterns within the U.S. higher education systems – emphasizing the plurality of *systems* to denote the variety of institutional types and variety possible within each U.S. state. Even amid this variety, upon reflecting on writing studies scholarship, Haswell (2009) finds patterns that indicate improvement: "The changes often are eccentric, erratic, and marked by periods of quiescence and even backsliding, but students leap ahead when they decide on a major, develop a more realistic sense of authorship and academic voice, and discursively construct a more viable interface between private and public identities" (p. 342-3). The challenge is to transform what is known about writing and writing pedagogy into a program.

In accordance with this variety, Bordelon, Wright, and Halloran (2012) note definitions of writing shifted depending upon institution type. Consequently, the assessment fluctuated based upon institution. Though writing instruction and courses and, therefore, writing programs tend to be associated with English departments, writing programs have existed beyond English departments. In *The Idea of a Writing Laboratory,* Neal Lerner (2009) describes writing programs housed in a department of biology at MIT in the 1990s, the General College Writing Laboratory at the University of Minnesota, and the Dartmouth Writing Clinic (1930s-1959). Pushing the concept of a writing program being beyond ready definitions, Joseph Harris (2006) describes how Duke University structured an "undisciplined" writing program, which appoints writing fellows from any discipline to instruct writing and propose a course design. The presence of writing programs beyond English departments has sufficient establishment within writing program administration scholarship that between the first edition (2013) and the second edition of *A Rhetoric for Writing Program Administrators* (2016), editor Malenczyk included a chapter by Barry Maid "What is an independent writing department/program?" that addresses the challenges and benefits of situating a writing as its own academic department (which the University of Minnesota has done since 2007) or the writing program as an entity separate from academic units, such as student support – an existence known to many writing center administrators. The understanding of all types of writing programs escape ready definitions. Condon and Rutz (2012) conducted a survey of writing across the curriculum programs. Condon and Rutz identified variety in the structure of programs in several dimensions (status within university, course situation within degree curricula, instruction staff, among others) that a tentative taxonomy of WAC programs was possible. However, more recently, in *Sustainable WAC*, Cox, Galin, and Melzer (2018) discovered through surveying that attrition among WAC programs was so significant that a large majority of programs no longer existed or had been dissolved only to be reformed through integrating courses or course activities elsewhere.

Regardless of how a writing program is construed, the assessment of the program has significant implications. "Assessments, like research methods, can produce the larger scene of a writing program with consequences for student placement, course goals, institutional perceptions of writing and the function of writing programs, and the terms of labor of teachers" (Scott & Brannon, 2013, p. 277). While this sweeping concept of Scott and Brannon (2013) seems more like program review, the thrust of the argument is to foreground the importance of empowering WPAs and writing faculty to shape programmatic assessment. "Power manifests in the questions we ask about literacy, the methods we deploy to answer the questions, and the vocabularies and narratives we use to articulate what we find" (Scott & Brannon, 2013, p. 292). The questions asked are derived from the underlying theories about writing (the concept of writing itself, teaching writing, and learning how to write) used to create a lens through which an assessment sees writing.

2.2.1. Theories of Writing Program Assessment

Since the 1940s, the scholarship on writing program assessment tends to cluster into three conceptual frames. Each frame represents a way to understand writing, students, instructors, and pedagogy. The three frames represent the historical developments of writing program assessment, and through conflicts and compromises have informed the theory and design of writing program assessment. Yancey (1999) refers to these conceptual frames as the waves of writing program assessment, noting that the history is a “narrative of incomplete and uncompleted waves” (p. 500). For example, the frames often overlapped, but each one has at some point been a predominate model for how to theorize writing assessment, to create means to perform the assessment, and to interpret the results of assessment.

The first frame focused primarily on the creation of test instruments, such as multiple-choice questions or fill-in-the-blank prompts, primarily serving the function of placing or admitting students into courses. For some writing programs, a second test instrument (often a final paper) or a re-administering at end of the semester of the initial instrument determined whether a student successfully passed a course. This particular frame involved indirect assessment of writing skill, and it tended to avoid influencing writing pedagogy, so was largely tolerated as a necessary means to filter students into and out of courses or to predict student success. The second frame shifted emphasis to direct assessment of writing by administering and scoring writing exams (Huot, 1990b). The focus had shifted toward student learning and effectiveness of instruction by administering the exams at the conclusion of a period of writing instruction, either the end of the course or at specific instances during a course. The third frame focused on expanding the direct assessment of writing through more situated writing other than exam-based written composition, such as incorporating portfolios as the instrument of assessment.

All three frames have influences from the two constructs associated with conducting research, particularly experiment or laboratory research: validity and reliability. The influence has taken the form of rejecting experimentation for allegedly ignoring the available research on how to score writing and the form of admiration for precision attributed to experimental methods as part of the pursuit to find "the best yardstick for literacy" (Lynne, 2009, p. 33). Lynne (2009) describes the pursuit of a "yardstick" for measuring writing as a series of "experiments" regarding what to evaluate (timed essays, research papers, portfolios, reading exams, and on and on), how to evaluate the writing and the writer (multiple-choice, feedback and revision loops, and on and on), and how often to measure (start of semester and end of semester, middle of semester and end of semester, only the end of semester, several times during a semester). The series also reflects the conflict between what scholars and instructors deem a valid assessment and a reliable assessment.

Most writing program assessment has these two constructs, validity and reliability, as points to configure a theory and design of writing program assessment. To my knowledge, only one theory of writing assessment attempts to divorce assessment from these two constructs (Lynne, 2009). The privileging of one construct or the other tends to become the inflection points as writing program assessment has developed.

2.2.2. Validity and Reliability: Changes in Two Ubiquitous Words in Assessment Scholarship

In brief, validity is whether the means of assessment is an appropriate measurement. This working definition of validity is widely accepted. Teasing apart that brevity, I note that *means* and *appropriate* are two sites for much debate. The concern tends to concentrate upon who defines what constitutes being appropriate and who develops the means. Someone who wants to measure correlations might argue that a direct measurement of writing, such as scoring a portfolio, cannot be an appropriate measurement because the scoring process and the scored items (portfolio materials) have too much variability (Hayes & Hatch, 1999), which has produced a sub-field of assessment design dedicated to interrater development (Dixon & Moxley, 2013; Penny & Johnson, 2011; Huot & Schendel, 1999). Someone who wants to measure attributes of writing might argue that the indirect measurement by an exam is not appropriate because it removes important contextual factors from the processes of written composition and therefore misses the mark when trying to determine writing skills (Wardle & Roozen, 2013; Fleckenstein, Spinuzzi, Rickly, & Papper, 2008).

The consequence of these disagreements is what I term the arguments by adjective. That is, the assessment scholarship addressing validity has produced several types of validity. The three prominent types are *construct validity*, *consequential validity*, and *face validity*. Construct validity denotes whether the instrument designed to collect data has adequate definitions and connection between the definitions and measurement of a variable of interest (Creswell, 2014). Consequential validity notes whether the measurement provides an indication of broader context. For example, “what happens to the students who pass – or don’t pass – a test?” (O’Neill, 2015, p. 162). Consequential validity does not stop at the test, which is the limit of construct validity.

Face validity is contentious because, unlike construct and consequential validity, face validity involves situated expertise. For example, the decision of whether to use a single piece of writing, such as a research paper, or a collection of written works in a portfolio requires awareness of what each instrument can measure in terms of construct and of consequence. The contention is that face validity is the most subjective; however, all three types of validity are open to extensive interpretation. Diogenes and Lunsford (2006) argued, as the definition of writing has shifted that the means of measurement tend to lag behind the composing activities happening in writing sections. One consequence of the mismatch is that more adjectives are joining the validity debates in an effort to assure that the written composing activities are addressed within programmatic outcomes. For example, *racial validity* has gained traction as an element of assessment design (Inoue, 2009), which has allowed for the formation of broader theorizing about writing assessment design that factors race and anti-racism into the design (Inoue, 2015). Most recently, Randall et al. (2022) proposed an antiracist validity framework using a heuristic for transforming assessment questions from traditional validity into antiracist validity questions. Similarly, Gere et al. (2021) proposed a framework for writing assessment focused on critical language awareness, striving toward a validity that recognizes linguistic variety as opposed to a constructed, standardized academic language. With the framework proposed by Gere et al. (2021), validity arises from an assessment design that distinguishes language conventions and language appropriateness as well as situates the distinctions between a descriptive grammar and a prescriptive grammar.

Randall et al. (2022) and Gere et al. (2021) propose frameworks, which is helpful because the arguments by adjective do not help to operationalize the concept of validity. Toward an operational definition, Messick (1989) offers the following: validity is “the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (qtd. White et al., 2015, p. 153-4). According to this definition, validity represents how well the measurement takes into consideration the obtained information (empirical evidence) and the types of information sought (theoretical rationales) in order to enable some response (action) and, important for my dissertation, interpretation of information and response (inference). I would re-phrase the definition in this way: a writing assessment is valid when the design permits an informed response based upon the theoretically supported collection of evidence *from sources of considerable variability*, namely the processes of writing, which has yielded theories about written composition about relationships. Based on this re-phrasing, I argue all theories and designs of assessment incorporate to some extent validity, and it aligns with the assertion by Murphy and Yancey (2009) that validity amounts to what an assessment allows instructors to learn from conversations *during* scoring rather than *from* scoring (p. 375). Again, turning to grammatical units, the verbs matter more than the nouns, that is, the scoring matters more than the scores when considering program-level assessment. What determines the extent of incorporation of the resulting discussions is the resources available to act upon what the assessment delivers, which happens through larger program review. However, intentionally or not, validity as a construct always seems to deny or minimize how much variance human activity such as written composition contains, possibly delegating this aspect of assessment to the other ubiquitous construct in assessment scholarship, reliability.

Reliability is whether the means of assessment is consistent. If the assessment has too much volatility in some element of the design, it is not considered reliable. With regard to volatility, the concern tends to be the fairness of the assessment design (Lynne, 2004), which gives way to disagreements about who defines what constitutes an assessment being fair (Inoue, 2015; Huot, 2002; Hayes & Hatch, 1999; White, 1993; Huot, 1990a). For example, with regard to fairness and by extension reliability, in “The Legal and the Local” Poe, Elliot, Cogan & Nurudeen (2014) have presented a thorough examination of previous efforts for a fair (equitable) assessment design then propose applying a contractual legal heuristic *disparate impact analysis.* Disparate impact is “a legal analysis to determine unintentional discrimination” (Poe et al., 2014, p. 590) that could function for composition studies as “a robust entry point into complex economic, social, cognitive, and affective contexts that shape [writing] assessment scenes” (Poe et al., 2014, p. 591). Poe et al. (2014) propose a tool that consists of several steps for evaluating disparate impact in a writing program. An ambitious idea for an interdisciplinary approach to understanding reliability (or fairness), but I am unaware of any writing scholarship that has reported an application of the heuristic. Even so, the concept has traction as apparent in the recent “Disrupting White Supremacy in Assessment” in *Educational Assessment* by Randall, Slomp, Poe, and Oliveri (2022) that proposes a justice-oriented anti-racist validity framework for assessments of all types, not only writing assessment. Conceptually, reliability increasingly means a consistent application must also include how the means of assessment are fair. Inferential statistics has a role to realize whether or not an assessment is fair by equipping WPAs with quantitative evidence to indicate patterns otherwise concealed within a data set comprised of descriptive and predictive statistics.

2.2.3. Portfolio of Student Writing as Assessment Instrument

The use of portfolios of student writing as the instrument to assess a writing program has its origins in the shift from cognitive theories about writing toward social construct theories about writing. In “Taking Stock of Portfolio Assessment Scholarship,” Lam (2017) summarized the shift in a historical review of the use of writing portfolios as a response to shortcomings of written exams or multiple-choice tests. Elbow and Belanoff (1986) are largely seen as the first instance of scholarship devoted to explicitly challenging the use of an exam to assess student writing skills. At State University of New York, Stony Brook, Elbow and Belanoff instituted the use of portfolios as a substitute for a proficiency exam, arguing a collection of writing materials better represented whether a student has developed as a writer. Hamp-Lyons (2001) dubbed this shift the third generation of writing assessment, which is similar to what Yancey (1999) characterized as a third frame – situated-writing that emphasizes the social dimension. Hamp-Lyons (2011) anticipated the fourth generation would involve extensive political aspects to language and its uses (and abuses).

The origins of the portfolio are associated with a social construct turn, but the portfolio itself remains loosely defined. White (2009) likens the change from exams to portfolios as "most evaluation instruments provide a snapshot of student performance, the portfolio can give a motion picture" (p. 163). The conceptual principle of a portfolio is sound; by providing more content, a student can better illustrate their effort to improve as a writer. Obviously, a portfolio ought to collect student writing. However, the parameters for what constitutes a collection remain debated: how much writing is enough, whether the collection should include only end products and not drafts, whether to include a reflective piece such as a letter or preface. These questions have not been settled, and the frequent response is to structure the portfolio based on local circumstances (Scott, 2005; Murphy, 1994). Murphy (1994) supports the use of portfolios but wonders if "establishing a portfolio culture may require teachers to make substantial changes in their instructional practices" (p. 178). The reflexive answer might be, no, the portfolio does not impinge upon instructor autonomy. However, by defining what constitutes a portfolio –

number of genres, types of genres, reflection activities – instructors do encounter mild control over course structure. A portfolio-based course guided by genre theories of composition would not permit an architecture-like portfolio of a project in all phases of production, though it could. By defining a portfolio, the course design and by extension the writing program design encounters constraints. Hamp-Lyons (2002) anticipates these quandaries noting constraints form the very social pact of communication practices, written or otherwise. “The ethical dilemmas and challenges we [writing instructors] face in balancing society’s need for assessments with our determination to do our best for learners are very great. Accepting a shared responsibility for the impact of writing assessment practices will put consideration of our own ethical behavior at the top of our agenda” (Hamp-Lyons, 2002, p. 14). Ethically, the local instructors through diligence in forming and reforming the writing program create fairness (reliable) and appropriate for actions and inferences (valid). The development of a portfolio accomplishes the balancing act between fairness for students (inclusive of a variety of written material) and appropriate actions for instructors (choice of written materials). Dunn, Jr., Luke, & Nassar (2012) consider the questions about how to parameterize portfolios a question of institutional infrastructure. That is, what can the institution support in terms of delivering students quality and instructors support. For Eastern Michigan University, where Dunn Jr., Luke, & Nassar (2012) designed a first-year program, the first question was to select a platform for eportfolio. The selection process revealed too readily the infrastructure issues. “Each instructor had an independent vision of what eportfolios looked like in the context of their course” (Dunn Jr., Luke, & Nassar, 2012, p. 67). Consequently, several rounds of evaluative scoring of systems and designs was necessary. The primary factor became the system that enabled, according to administration, “large-scale collection and sampling of student work for future research and scholarship beyond the more immediate need for program-wide assessment" (Dunn Jr., Luke, & Nassar, 2012, p. 67). In some respects, the portfolio resembles a compromise between institutional interest and evidence-based writing pedagogy scholarship. A written exam allows for expedient collection and comparison of data (test scores), but written exams have minimal support in writing pedagogy scholarship. A portfolio system allows for more robust collection of data but are more time-intensive.

For now, the portfolio is the best of available instruments. However, as an instrument, the portfolio has become the subject of scrutiny. While Lam (2017) supports the use of portfolios, he concludes by noting one persistent problem that previous research has not addressed: "we [writing instructors and scholars] have no idea of how students actually make sense of the learning evidence generated by themselves, peers, teachers and other external source to improve text quality, and how they interpret, internalize and capitalize on this evidence to support growth in continued writing development" (p. 91). White (2009) shares this concern by noting that the holistic scoring approach to portfolio assessment retains techniques developed for scoring exams for writing proficiency. Holistic scoring involves a scorer "to give a 'general impression' score after a quick reading" (White, 2009, p. 166). Scorers compare general impression scores and then resolve any disagreements following one of several approaches. One undesired experience of this scoring approach is scorers often feel as though they are re-grading student material. One response to this criticism has been incorporating reflective elements to the portfolio, such as introductory letters or presenting the collected work. The continued use of holistic scoring is perhaps best explained by Hamp-Lyons (2016):

> The felt need for a single assessment instrument that could show both high reliability and explicit validity, and that could make sense to all stakeholders in an assessment context were very important reasons for the emergence of trait-based approaches to judging

> writing. But another important reason for the move in this direction has been to bring assessments back into the hands of teachers. (p. 2).

The noted criticisms by Lam and by White are indicative of a trend among scholars, in writing studies but other disciplines too, questioning the rigor of research methods. Specifically, within writing studies, Haswell (2005) metaphorically characterized the lack of rigor as a war. The belligerents, according to Haswell, are the National Council of Teachers of English (NTCE) and the Conference on College Composition and Communication (CCCC) against "empirical inquiry, laboratory studies, data gathering, experimental investigation, formal research, hard research, and sometimes just research" (p. 200). For research, Haswell offers the replicable, aggregable, and data-supported (RAD) framework. More recently, Raucci (2021) has issued a call for scholarship in writing studies to undertake a replication agenda. While noting the polysemy of *replicate*, Raucci (2021) argues striving toward replication shifts emphasis from overly praising results from a single study toward "mutually constitutive relation between procedures and results" (p. 449). Given the widespread use of student portfolios and assessment designs for portfolios, it is debatable whether incidental replication has already happened but no one has taken the time to review the scholarship and realize writing studies scholarship has resonance between procedures and results. One possible endeavor to determine if such resonance has happened may reside in scalar approaches to humanities.

2.3. Scalar Reading in the Humanities

The concept of scalar reading gained prominence when Franco Moretti wrote "The Slaughterhouse of Literature" in 2000 for *Modern Language Quarterly*. Moretti proposed the idea of distant reading as opposed to close reading. A close reading would involve explication of literary and rhetorical devices within a given written work. A distant reading, in contrast, involved "reading" as many works as possible in order to find patterns among literary and rhetorical devices and based on those patterns characterize a given genre, in the case of Moretti's article, the detective story of the 19th century. Counting occurrences of literary devices or other lexical features has a history prior to Moretti as most digital humanists groaningly acknowledge.[4] However, distant reading involved the application of tools beyond tallying and categorizing, and many of the tools are statistical. In *Macroanalyis,* Matthew Jockers (2013) explains the purpose of statistical tools, such as topic modeling in which a computer models a collection of written materials (a corpus) based on probabilities of collocated words within the collection. The model discovers themes based on how likely patterns of word clusters are within the collection, which then becomes labeled as a theme.

I am omitting a lot within the process in order to focus on the main concept: the themes identified using distant reading are supported through evidence on a larger scale (the corpus) in order to provide support or rebuttals to interpretations at a smaller scale (an individual item within the corpus). Ideally, the two readings happen in tandem to provide a richer understanding of the collected works – novels, advice columns, U.S. Civil War newspapers (see Cordell and Smith, 2017, *Viral Texts*), bibliographies of *College Composition and Communication* over a twenty-five year span (see Mueller, 2012, "Grasping Rhetoric and Composition by Its Long Tail"), research methods of 2,711 rhetoric doctoral dissertations (see Miller, 2014, "Mapping the Methods of Composition/Rhetoric Dissertations"), or any other large collection of materials, written or graphical.

[4] Father Roberto Busa created *Index Thomisticus* about the writing of Thomas Aquinas, which has earned Busa the title "the founding father of humanities computing" (Jockers, 2013, p. 1). Busa started the project in 1946 and developed it for roughly 34 years by creating and maintaining the computer punch cards to render the works of Thomas Aquinas into machine-readable material.

Much of the motivation for fusing distant and close readings is to encourage seeing materials in different relationships. Often, the scholarship is about literally seeing the material in a new way, such as Klein (2013) demonstrating “how a set of techniques that derive from the fields of computational linguistics and data visualization help render visible the archival silences implicit in our understanding of chattel slavery” (p. 665). By working with existing data sets and through emerging interpretive techniques, a richer understanding of topics is possible. Ramsay characterizes this interplay of interpretive techniques and data sets as the hermeneutics of screwing around: “Your ethical obligation is neither to read [every book published] nor to pretend that you have read them all, but to understand each path through the vast archive as an important moment in the world’s duration – as an invitation to community, relationship, and play” (p. 9). As welcoming and optimistic as this invitation might be, there are consequences to a seemingly ever-expanding access to information, interpretive tools, and transference to practical uses.

Jones (2019) points to the challenges of understanding how difficult perspective shifts can be, especially when the shifts involve a change in scale. “Environmental problems are at once too vast and too mundane. Scale is a central rhetorical concern for communicators interested in promoting ways to meaningfully inhabit places of change” (Jones, 2019, p. 81). The volume of data (too vast) can overwhelm when framed through long time scales, such as decades long projections about climate change. Or, it is difficult to understand how activity at one scale is influential at another scale. For example, “individual decisions, such as cranking up a car, are ‘statistically meaningless’ on the micro-level yet geologically impactful when scaled up to the level of species” (Jones, 2019, p. 81). The challenge becomes that explaining the interaction of various scales requires literacies of several methods. A narrative paired with a quantitative analysis is more effective than either one alone. Unfortunately, the pairing involves an inclusive view toward what is acceptable evidence, and often the knowledge and skills necessary to understand, let alone synthesize, evidence from different scales and often different disciplinary sources is difficult to develop, distribute, and maintain.

Similar challenges are familiar to archival scholars and archivists, especially those who work within the broadly defined group known as digital humanists. Drucker (2021) argues the challenges hinges on two important concepts, sustainability and complexity. Through a narrative of preparing *ArtistsBooksOnline*, which was a large archival project to digitize the notes and other written materials by artists, Drucker identifies several areas of concern that threaten sustainability: “infrastructure dependency, platform design, intellectual frameworks, community buy-in, and obsolescence of plug-in functionality” (p. 87). The aspiration of the project was to collect materials from artists, digitize them, and use the resulting archive as a means to expand the interpretive potential of art works and grant greater insights into patterns of artistic developments. The aspiration deflated as each of the threats to sustainability punctured the plan. “Filling these fields [in the forms for the database] was a challenge for many scholars and students, who complained it was ‘too hard’. Pushback from within the [artist] community demonstrated hostility of individuals largely ignorant of the basic concept of metadata and unwilling to experiment in its design” (Drucker, 2021, p. 87-88). Compounding the problems was that Drucker had recently changed jobs from University of Virginia to University of California, Los Angeles. Consequently, “[UVA staff] tired of care-taking [of the project], even though this involved little more than continuing to host the project files on a server” (Drucker, 2021, p. 88). Drucker admits to becoming tired of the project.

Further bolstering the argument that Drucker has made for the challenges of the projects, Ekiba et al. (2015) completed a comprehensive survey of data projects across a wide variety of disciplines to generate the report "Bigger Data, Bigger Dilemmas: A Critical Review." The report covers a range of topical areas from how to conceptualize what constitutes "big" of big data, how to reconcile competing methods for gathering data in the first place but then what does the storage resemble (garden or graveyard – both have a purpose), what to do regarding technological changes in terms of sustaining projects, and the often overlooked ethical dilemmas regarding who decides what can be collected and then what happens to the collected data. The interdisciplinary team provides a heuristic. "In order to go beyond dilemmas, one needs to understand their historical and conceptual origins, the dynamics of their development, the drivers of the dynamics, and the alternatives that they present" (Ekiba et al., 2015, p. 1540).

This exigence about dilemmas points to major issues for interdisciplinary approaches and projects. On the one hand, the new methods and tools increase what types of questions scholars and others might ask and thereby support the inception of new types of projects. On the other hand, the new methods and tools also increase the necessary knowledge and skills and administrative acumen necessary to realize addressing the new questions raised from the new types of projects. To explore a topic at several scales, it is necessary to exercise caution in the decisions about which tools to use, the reasons for using those tools, and the scope of what is possible. Drucker, for example, admits that one problem with *ArtistsBooksOnline* was the design incorporated largely the recommendations of the University of Virginia library technical specifications, which resulted in infrastructure dependency and platform design constraints. The intellectual framework shifted from artistic theory toward theories of database architecture. The artists had highly varied interest in what was happening with their materials, and "many made unreasonable demands *based on the assumption that I had a large staff and budget*" (Drucker, 2021, p. 88, emphasis added). The complexity of completing any piece of one part of the overall project grew faster than the resources – budget, staff, intellectual knowledge, technical skill – could address. Drucker uses the example as a cautionary tale about the uses of interdisciplinary methods. Gallagher et al. (2020) note one of the limitations of "big data" projects is that "structuring data in a standardized way so that software programs can analyze it is time consuming" (p. 166) and "large amounts of computing power were necessary" (p. 166).

Limitations such as how to curate data, how to clean and structure data, and how to analyze data are not intended to discourage scholarship that is data intensive. Rather, scholars and administrators need to be attentive to what might cause a project to experience complications and possible failure. Furthermore, the limitations point the necessity to ground a project that aspires to be accessible needs to be sensitive to the realities of performing the research. Writing program administrators risk spending more time as technical support than as writing program administrators. That is, writing program administrators ought to begin from knowing writing studies and writing pedagogy and allowing that knowledge to lead how tools from other disciplines, such as inferential statistics, shape a project. To understand the reasons for using inferential statistics, therefore, it is important to understand the reasons for selecting the Trans-Atlantic and Pacific Project as a variable of interest for a zoomable assessment.

2.4. The Trans-Atlantic and Pacific Project and Other International Collaborations

The Trans-Atlantic and Pacific Project (TAPP) originated in planning conversations in 1999 and then in 2000 the first collaboration between a class taught by Bruce Maylath, at the time teaching at the University of Wisconsin-Stout, and a class taught by Sonia Vandepitte, who taught at Mercator College of Translation and Interpretation, in Ghent, Belgium. Maylath taught a Technical Writing course, and Vandepitte taught an Essentials of Translation and Interpretation course. The students enrolled in Technical Writing authored instructions, then prepared the written material for translation. The students enrolled in Essentials of Translation and Interpretation translated the provided instructions. Through the processes of translation, the authors and translators negotiated terminology and other aspects of the material. This model of exchange is known as a bilateral collaboration, as it links two classes in two different countries. From this pairing of courses, the TAPP has expanded participation to include many universities, written composition courses including first-year writing (Verzella & Tommaso, 2014), writing-in-the-disciplines courses (Steinmann, Saduov, & Maylath, 2016), technical communication courses focused on international communication (Sorensen, Hammer, & Maylath, 2015). In addition to bilateral collaborations, several partnerships have more complex multilateral collaborations in which more than two courses participate in authoring and translating materials (Maylath et al., 2013). Some of the multilateral collaborations have included usability and user experience testing (Maylath et al. 2013). The collaborations did not always involve the English-speaking students as the sole authors of material for translation. For example, Maylath, King, & Arnó Macià, (2013) coordinated a collaboration in which engineering students in Spain and in the US co-authored instructions that were then conveyed to another group of translators. The Spanish and Catalan students, therefore, had to compose using a foreign language (English) and depend upon their co-authors in the US course to provide support and guidance. In addition to research interests related to linguistic and cultural topics, the activities conducted during a TAPP collaboration also offer unique opportunities to examine various media used as part of a project and student practices using the different media while they participate in a collaboration and complete the projects. Consequently, some TAPP scholarship has examined student digital literacies (Hammer & Maylath, 2014). However, the collaborations tend to foreground the importance of linguistic and cultural aspects within a project workflow.

To illustrate the project workflow, Mousten, Maylath, Vandepitte, & Humbley (2010) coined the phrase *text travel*. Upon reflecting on the phrase *text transfer*, Mousten et al. (2010) recognized *transfer* overly simplified what happens during collaborations that involved translation and localization.

> Since the Trans-Atlantic Project[5] involves the travel of texts from one culture to another, we have gradually replaced the term "text transfer" with "text travel." The latter draws increased attention to the process, rather than just source and target texts. Text travel, therefore, covers diverse processes in the Trans-Atlantic Project such as texts flowing in different directions at different times, subjected to diverse cultural and linguistic changes on the way (Mousten et al., 2010, p. 410)

The concept of text travel as a constellation of processes illustrates the complexity of the activities involved in even an ostensibly simple bilateral exchange in which one class writes materials and another class translates the materials. This complexity has been noted in other international collaborations. For example, Starke-Meyerring & Andrews (2010) designed a collaboration between one course at McGill University, in Montreal, and one course at the University of Delaware. The students of both courses self-reported higher competency in several categories based on a 0-5 scale. Several of the greatest increases were in "Analyzing the

[5] At the time, the Trans-Atlantic and Pacific Project had not expanded to include a trans-Pacific collaboration. A collaboration between North Dakota State University and the Beijing Foreign Studies University in 2013 first expanded the TAPP across the Pacific.

audience for a communication product" and "Select an appropriate technology for team communication"[6].

International collaborations have benefits in a variety of ways that are not readily quantified. For example, Bosley (2010) notes in "Do Fish Know They are Swimming in Water" that during a collaboration between an Introduction to Technical Writing course at University of North Carolina, Charlotte, and twenty graduate students of technical writing at Université Paris – Denis Diderot, "many of these [student] comments indicate a level of curiosity that I rarely see; they [the U.S. students] were genuinely intrigued by the idea of communicating with students outside of the United States" (p. 222). Curiosity is difficult, possibly impossible, to quantify using any instrument.[7] However, between the quantification through self-report by Starke-Meyerring & Andrew (2010) and the interviews and survey responses by Bosley (2010), at the course level, the benefits are abundantly apparent. The pair of studies reflect what insights are possible through mixed methods. Though international collaborations are increasingly common, Prior and Lunsford (2008) is one of a few sources to note that translations (often a major component of international collaborations) have rarely been mentioned within the history of U.S.

[6] Starke-Meyerring and Andrews (2010) do not provide the exact scores, only several bar graphs. Consequently, I am approximating which categories have "the greatest increases." I would also note that Starke-Meyerring and Andrews (2010) could have conducted inferential statistical tests with the collected data to provide more thorough analysis of the student self-reports. For example, the competencies are grouped into Genre Knowledge, Proficiency in Communication Processes and Strategies, and Intercultural Virtual Team Communication. Based on these groups, inferential statistics could identify if there are patterns among the self-reports, for example whether students who reported greater proficiency in Genre Knowledge also reported greater proficiency in the other two groups.

[7] Bosley (2010) relates the inspirational source of the chapter title: "Like fish who do not know they swim in water, our American students often do not realize that their culture is not the culture of the world. That they need help to recognize "the water" is best illustrated in this comment one American student made near the beginning of the project: 'It did not occur to me that our Paris partners would not have a holiday on November 25, Thanksgiving…Of course this makes sense, but it did take me a second to shift my thinking" (p. 224).

writing pedagogy. In the chapter “History of Reflection, Theory, and Research on Writing,” Prior and Lunsford (2008) conclude a section dedicated to translation studies by noting, “reflection and research on translation offer particularly rich illustrations of how complexly writing works within and across literate and social ecologies” (p. 87). As the scholarship on international collaborations makes evident, collaborations that include a translation component are a source of deep, rich insights into written composition. By applying inferential statistics to data about a writing program, I contribute to this understanding by analyzing if participation in an international collaboration, the Trans-Atlantic and Pacific Project, has a detectable difference at a more abstract level, the writing program, rather than more frequently analyzed levels of the course or individual student.

3. METHODS: KEEP IT SIMPLE STATISTICS

3.1. Introduction

The focus of this dissertation is on how a statistical inferential test complements an existing writing program's assessment design, thereby allowing writing program administrators (WPAs) to learn more about a program that they need to assess. Statistical inferential tests enable WPAs to ask and answer questions about the program beyond typical writing program assessment designs, such as writing a program narrative or tabulating portfolio scores, without extensive modification to the design. However, whether the questions-answers that the tests enable WPAs to obtain are meaningful remains open to debate. For this dissertation, I conducted an inferential statistical test to determine if participation in the Trans-Atlantic and Pacific Project (TAPP) had a detectable difference on portfolio scores. A writing section participating in TAPP has a partner section at another university. The two sections arrange for students to collaborate on a project. The sections participating from the upper division writing program would incorporate activities such as translation preparation into the pedagogical structure of the section. The TAPP partners would receive the student authored documents then provide feedback through in-document (such as comments or edits) and email correspondence. Beyond this incorporation of an exchange, participation in the TAPP does not require instructors to modify the existing course design. As a minimally invasive set of activities, therefore, I am interested to examine if the TAPP has a noticeable difference on the participating sections that is detectable at the program level. It would illustrate the idea of a small change having a noticeable difference, and consequently provide WPAs hope and headaches. Hope, because it indicates small activities can have tremendous impact; headaches, because, once again, it indicates small activities can have tremendous impact. I dub this pairing the nuances and nuisances of inferential statistical analysis.

3.2. Variable of Interest and Data Source

The variable of interest was portfolio scores for a specific writing program outcome, Integrating Knowledge and Ideas (IKI). For three academic years, I collected student portfolio data adhering to the existing end-of-semester writing program assessment. To satisfy the existing assessment design, a portfolio must contain at least three genres of writing, include at least fifteen pages of written content, and open with a one-page reflective letter by the student. The conditions were accurate of all portfolios that were assessed.

I divided the student portfolios of the upper division writing courses into two categories, TAPP-participating and Non-TAPP participating. Instructors elected to have sections participate in the TAPP during the semester when the students completed the portfolio contents. Instructors were not assigned to participate or not. Therefore, a possible confounding variable that escapes the current research design is whether instructors who elected to participate in the TAPP were distinctive in some capacity compared to those who did not participate in the TAPP. However, a succinct explanation is that the existing writing program assessment design delegates instructor performance to other mechanisms of the program review[8]. Program assessment narrowly

[8] Program review and program assessment are two phrases often used interchangeable. However, in writing program scholarship, the established understanding of the two phrases is that program review encompasses a cluster of activities ranging from determining if student placement mechanisms are effective, evaluating the clarity of program outcomes, and developing instructors whether full-time, part-time, or graduate assistants. Program review also contains program assessment. Program assessment is the specific set of activities such as sampling activities in courses (writing an essay exam or collecting a portfolio of student writing), judging the sampled material (pass-fail, scoring scales, etc.), and reporting the results of the activities. As steeped in writing assessment as I am, I still do not draw the distinctions easily. However, a distinction exists that is important, as important as the distinction that assessing a portfolio is not grading the student work again – which is a source of more questions and confusion.

addresses the question of whether the portfolio satisfies a program outcome. The assessment does not address instructor performance – or student performance for that matter. The content of the portfolio is a product of the labor of students and of instructors. However, the assessment seeks only to check the alignment between the portfolio and outcomes. The granularity of the distinction complicated every aspect of this dissertation and seems to exist at the crux of much writing studies scholarship: how to determine whether a student knows how to write, whether an instructor knows how to teach, or whether an administrator knows how to manage. Or, zooming out to the most global perspective, whether learning has occurred. A zoomable assessment takes existing instruments, such as portfolios and program outcomes and scoring techniques, and attempts to harmonize them through perspective shifts, looking at the data in new ways, notably inferential statistical tests.

For zoomable assessment to avoid burdens of maintenance and technical debts, the method needs ought to not disrupt an existing assessment design. The inclusion of an inferential test was to complement the design, not to redefine the entire writing program assessment (which can be beyond immediate control of a WPA). Beyond the inferential statistical test, the one additional activity was to categorize portfolios into TAPP and Non-TAPP. The existing assessment design used dynamic criteria mapping (DCM) (Broad, 2003). Dynamic criteria mapping involved a group of instructors identifying values then based on the identified values created outcomes, including IKI. The upper division writing program first instituted the outcomes derived through DCM roughly five years prior to the dissertation project, so the writing program outcomes had been established. However, during the study period, the writing program assessment explored one new outcome that was also rejected within the study period. The dissertation had no role in the determination of making then dropping the outcome.

For several years prior to the data collection period of this dissertation, the assessment addressed a pair of program outcomes, IKI and “Genre, Audiences, Purposes, and Situations” (GAPS). Also, several sub-outcomes for each program outcome had been part of the assessment design. For some semesters, both program outcomes were assessed. Then, for some semesters, only one program outcome was assessed. The change in focus of outcome introduced variability that I could not control. It was unclear why only one outcome or the other was assessed. However, to limit variability, though seven semesters of data were collected, I only used five semesters because two semesters of the assessment focused on different writing program outcomes (the piloted then rejected one and another existing program outcome). In this way, I limited variability by concentrating on a single outcome, IKI.

3.3. Statistical Test and Hypotheses

To determine whether participation in the TAPP resulted in a statistically significant difference of IKI scores, I conducted a *Welch’s t-test*. This test examines whether the mean scores (μ) of a particular variable (such as an IKI score) of two samples from a single population are significantly different. The test does not point to explanations if a significant difference exists. The use of an inferential test departed from the frequent reliance on only descriptive statistics or predictive tools frequently discussed in writing program assessment scholarship. For example, in *A Guide to College Writing Assessment,* O’Neill, Moore, and Huot (2009) describe the possibility of incorporating statistical tools for the purpose of student placement, noting regression analysis is preferable though acknowledged that the resulting coefficients (the “Rs” of regression) are difficult to interpret if the placement decision requires consensus among several decision makers (such as the course instructor, program administrator, and the student). White, Elliot, and Peckham (2015) also encourage the use of regression analysis when designing an assessment, including student placement but also for other purposes related to program review. However, more importantly, White et al. (2015) provided inspiration for the method of this dissertation. While White et al. prefer regression analysis, there is a note that “the workhorse of inferential measurement [is the t-test]” (p. 124), but they do not specify the type of t-test. White et al. do not discuss their rationale for choosing regression and delegating less attention to the “workhorse” of statistical analysis. My rationale for avoiding regression analysis is the utility of the regression resides primarily in *predicting* performance of a facet of a writing program (e.g., student success in first-year writing) rather than *inferring* about facets of a writing program (e.g., a pedagogical approach had a difference on mean portfolio scores). I did not want to predict if the TAPP *will* make a difference on portfolio scores, which is a different study. I wanted to infer if the TAPP *did* make a difference based on the patterns of portfolio scores. It is a subtle distinction but an important one. [9]

Furthermore, a second reason to choose a t-test rather than conduct a regression analysis was I want a zoomable assessment to be accessible. T-test is more accessible than regression analysis because the former involves less knowledge of statistics and experimental design. For

[9] The debate about the merits of regression analysis applied to human activities is (way) beyond the scope of this dissertation. However, I direct interested readers to the following materials for a discussion of the *application* of regression analysis: *Data and Goliath* Bruce Schneier (2015); *Judgment Under Uncertainty* (1974) edited by Daniel Kahneman, Paul Slovic, and Amos Tvrsky; *The Mismeasure of Man* (1980) by Stephen Gould; *The Model Thinker* (2018) by Scott Page; and *The Seven Pillars of Statistical Wisdom* (2016) by Stephen Stigler, who is an excellent historian of statistical thinking in general (*Statistics on the Table* (2002) and *The History of Statistics* (1990)). As an example of a recurring theme among the readings, “Regression,” a chapter in *The Seven Pillars of Statistical Wisdom,* Stigler (2016) discusses how regression analysis helped Francis Galton elaborate on evolutionary biology but also enabled Galton’s racist interpretation of selective breeding to eliminate ‘negative traits’ in human populations, which motivated eugenics proponents. In summary, regression is powerful, but the power requires more caution than usually exercised when people apply it. As I note, regression analysis has attracted the attention of writing studies scholars, but as of 2021, no useful scholarship has emerged, only “calls” for using the method.

the tool to be accessible, it should require minimal knowledge of statistics, accessible data sources, and no specialized software. The motivation for creating an accessible tool was to increase the likelihood that inferential statistics could become a complementary component of an assessment design. As has often been noted, sophisticated projects tend to become abandoned when original implementors are no longer involved or no longer serving in a role (such as WPA). Meloncon and St. Amant (2019) examined five years of research methods used to conduct technical and professional communication research. One aim was to characterize the research happening in technical and professional communication, but another aim was to identify traits of a sustainable research project. “If an approach is explicitly used repeatedly, there is an inherent and implicit if not explicit value argument being made” (Meloncon & St. Amant, 2019, p. 151). The least frequently considered parameter for research was cost. Cost not only in terms of money but also in terms of time to conduct the research in all phases – data collection, data preparation, data analysis, data interpretation, and then actions based on interpreted results. Methods have maintenance burdens (how to store information and transfer it when needed) and technical debts (what skills and knowledge are necessary to use the methods – software proficiencies, data analysis knowledge, and so forth).

As noted, for this dissertation project, the application of inferential statistics depends upon the existing method of creating a program outcome using dynamic criteria mapping (Broad, 2003) and procedure for scoring end-of-semester portfolios (O’Neill, Moore, and Huot, 2009). The data set created using the existing assessment design presently considers one data point, the IKI score of the entire sample of portfolios. This data point is the highest level of observation. To zoom into the data set, I used IKI scores for courses and IKI scores for TAPP and Non-TAPP sections. These scores draw the observation into greater details of the program, similar to zooming into a map to view features of specific cities rather than a viewing only the border of a state. It is a simple move in scale, but one that has significant potential and minimal resource requirements. Statistics enables the move by providing another perspective on a writing program, and a zoomable assessment retains the insights from writing studies to inform the application of statistics.

In this regard, a zoomable assessment builds upon the success of research-based and technical assessment designs. "Research-based assessment requires that the community of teachers, students and administrators come together to articulate a set of research questions about student performance" (Huot, 2002, p. 178), and technical assessment draws upon the expertise of crafting apparatus or instruments to determine how to ask questions. Research-based privileges broad participation in assessment design, and technical privileges expertise in the designing of assessment. Both assessment designs have merits. The purpose was to expand upon what the designs can provide for a WPA by demonstrating the versatility possible through conducting one statistical test. The application of inferential statistics aligns with the theory of writing assessment that in *Coming to Terms* Patricia Lynne (2004) characterizes as meaningful and ethical assessment. According to Lynne, an assessment is meaningful when the assessment process "draws attention to the object of assessment [portfolios] and queries of this sort tie evaluation to the situation in which literacy takes place" (2004, p. 117). Lynne defines ethical assessments as ones that "organize and provide principles for understanding the conduct of the procedures for evaluation" (2004, p. 118). That is, the assessment process is accessible and readily understood by many, or preferably all, involved or interested individuals. A zoomable assessment is accessible by maintaining the major activities of an existing assessment, and a zoomable assessment is readily understood by more individuals by providing a purpose to numbers aside from including the numbers in a report. A zoomable assessment makes numbers tell stories by allowing WPAs to ask questions of data sets.

To begin the testing process, I performed the assessment according to its design. For five semesters, I collected end-of-semester portfolios after the instructors assigned scores (following the process described in 3.3 below) then calculated the mean IKI scores of all portfolios, of each course, and of TAPP and Non-TAPP sections. In order to minimize possible scoring bias, I did not inform portfolio scorers that participation in the TAPP was the variable of interest. They followed the procedure for the existing assessment. They provided holistic scores for IKI.

Based on the mean scores, I conducted the t-test. A reason for selecting Welch's t-test rather than the typically selected Student t-test was that Welch's assumes that the samples created from a population of interest will have unequal variance with regard to the variable of interest. That is, Welch's t-test recognizes that much of the variability is beyond the control of the experimental design. In comparison, the Student assumes the population has an equal variance. In *Willful Ignorance* Herbert Weisberg (2014) argues that the underlying assumption of equal variance when researching a human activity, such as writing, even when a very large sample is available, is a misguided application of statistical inference. Weisberg dubs such misguided applications as examples of the titular willful ignorance. In *Statistics Done Wrong,* Alex Reinhart (2015) agrees with Weisberg, noting that most misapplications of statistics originate in experimental design, such as presuming equal variance in certain human activities. The pattern is apparent regarding assuming variance will be equal, but some writing program assessment scholarship does not check this assumption before using many statistical methods of any type, predictive or inferential (White et al., 2015).

A Welch’s t-test asks the question of whether two samples from a population have a similar mean (μ) for the variable of interest, which in the current study is the same mean student portfolio score. If the two samples have a similar mean, a person may confidently conclude that the two samples have no difference based on the variable of interest. The two samples would be deemed equivalent to one another in respect to the variable of interest. The one caution is to note that failure to reject a null hypothesis (no difference exists) is restricted to the variable of interest. Differences might still exist but require testing other variables to determine whether the differences are significant.

This dissertation involved two sampling techniques. The first sampling technique involved dividing the entire collection of portfolio data into a group of TAPP-participating courses and a group of Non-TAPP-participating courses. The sampling technique is quasi-random because students did not choose between joining a TAPP-participating course or not. They would not know in advance of joining a course’s section whether the section was participating in the TAPP. Though the students do not choose whether to join a course or not, I note again that the instructors did choose to participate in TAPP. A second sampling technique involved sorting the collected portfolio data by course. After sorting by course, for each course the data was divided into TAPP and Non-TAPP portfolios.

For all samples, the test procedure remains the same. If the samples do not have the same mean, a person may confidently conclude that the two samples are different when considering the variable of interest, and the difference is not a consequence of randomness but a pattern. However, a person cannot determine causality or correlation, so the underlying explanation for why the variable has a difference in means requires examining possible distinctions encapsulated within the variable of interest.

Based on the two sampling techniques, I created two sets of hypotheses. For the first set of hypotheses, I focused on the first sample, addressing the question, for the entire writing program, does the mean IKI score for the TAPP portfolios differ significantly from the mean IKI score for the Non-TAPP portfolios?

Based on this question, the hypotheses are,

H_0: $\mu_{TAPP} = \mu_{nonTAPP}$

H_a: $\mu_{TAPP} \neq \mu_{nonTAPP}$

For the second sampling technique, I created hypotheses based on course and participation in TAPP. Among the courses taught during the study period, two courses had large enough samples to satisfy test requirements: English 321 (Writing in the Technical Professions) and English 325 (Writing in the Health Professions). The other courses had insufficient participation in TAPP for meaningful comparison, often no sections participating or only one. For both English 321 and English 325, the question is whether TAPP there was a difference in mean portfolio score: does the mean IKI score for English 321 TAPP participating portfolios differ significantly from the mean IKI score for English 321 Non-TAPP participating portfolios? And, does the mean IKI score for English 325 TAPP participating portfolios differ significantly from the mean IKI score for English 325 Non-TAPP participating portfolios? The hypotheses set for each course resembled the hypothesis for the first sample. The two hypothesis tests for the second sampling technique are,

For English 321 H_0: $\mu_{321TAPP} = \mu_{321NonTAPP}$

H_a: $\mu_{321TAPP} \neq \mu_{321NonTAPP}$

For English 325 H_0: $\mu_{325TAPP} = \mu_{325TAPP}$

H_a: $\mu_{325TAPP} \neq \mu_{325TAPP}$

Where the designation of *a* (such as μ_{321a}) denotes a TAPP-participating mean IKI score, and *b* designates a non-TAPP-participating mean IKI score. Similar to the main hypothesis, the test for the course does not determine the reasons for any identified differences, only whether participation yielded a difference in portfolio scores.

3.4. Portfolio Scoring Process

The scoring method for the student writing portfolios used a version of multi-trait holistic scoring designed for a specific outcome. The outcome of interest was the Integrating Knowledge and Ideas (IKI). The outcome reads as follows: “Students will be able to integrate knowledge and ideas in a coherent and meaningful manner.” The scoring process started with instructors collecting a random sample of portfolios from the courses that they taught. After selecting the portfolios, instructors of a course exchanged the collected portfolios, so that instructors of English 320 scored portfolios of English 320 courses.

The portfolio selection process involved the writing program director providing five numbers, between 1 and 22, to instructors to select student portfolios. The provided numbers directed instructors to the class enrollment roster to find corresponding student with the roster position whose portfolio was added to the population. If a section did not have a student for a provided number, the instructor looped the count; for example, if the provided number was 22 but the course had 20 students enrolled, the instructor would select the portfolio of the student listed as 2 on the roster. If the looped count overlapped with a provided number, the instructor selected the next portfolio; continuing with the previous example, if 2 was already a one of the provided numbers, the instructor would select the portfolio of student listed as 3 on the course roster. Though this randomized sampling does not follow the convention of typical random sampling for experimental research, the purpose is to create a convenience sample not a representative sample.[10]

The holistic scoring procedure used a 1-5 scale, following a design similar to the Likert scale commonly used for survey research. Prior to scoring portfolios, the instructors participated in a scoring calibration activity. The calibration involves all participating assessors using a portfolio as a shared point of reference. The assessors assigned a score to the reference portfolio, then a discussion followed regarding reasons for assigning a particular score. The purpose of calibration, in particular the discussion, was to encourage assessors to shift from a grading and feedback mentality toward an assessing mentality. It is worth reminding readers that the scoring adheres to *assessing* as a mode of judgment rather than *grading* or *responding* (Tchudi, 1997). After calibration, the instructors scored ten portfolios for their course (five portfolios from their course and five portfolios from another instructor who taught the same course). Any score differences greater than 1 required a third scorer to provide the final score. The interrater agreement functions as a proximity indication of the success of the calibration activity, suggesting consensus about the scoring process. While this interrater agreement threshold does

[10] A ‘true’ random sampling would involve pooling, for example, all English 320 portfolios into one group, then using that pooled list to select the sample. Following this sampling method, all portfolios of a sample could be from one section. While an appropriate practice for checking performance quality of machine parts, this sampling technique is undesirable for a writing program for several reasons, most of them rhetorical. The sampling method risks the sample being comprised of portfolios entirely from one section or mostly from one section, which would *feel* more like an assessment of that one section and, quite possibly, the instructor, rather than an assessment of how English 320 portfolios performed in respect to the outcome. Furthermore, only a few sections of courses participate in Trans-Atlantic and Pacific Project activities, so the sampling method risks completely omitting these sections from the sample or overrepresenting participation by selecting mostly from the TAPP participating sections. As a sampling technique, the effort to create a ‘true’ random sample for writing program assessment befuddles me aside from a desire to appear ‘objective’ or ‘rigorous’ which I argue ultimately misrepresents the program and provides no useful information for program description or improvement. I am not interested in pretending to be an ‘objective’ researcher using ‘true’ random sampling when such a method is probably more harmful than insightful.

not meet a proposed stricter standard (re-scoring *any* difference, White et al., 2015), it does conform to expectations for a smaller scoring scale (Huot, 2002; Elbow & Yancey, 1994). Of greater interest for the existing design was the inconsistency of documenting the need for a third reader. Consequently, I could not provide interrater scores for all courses. I note that best practices for scoring should follow the conditions available for a WPA (O'Neill, Moore, & Huot, 2009; Strickland, 2001), consequently, the real question should be whether the interrater score warrants closer documentation than the current design permits or if it has been decided that only frequent need for third readers warrants documenting. At any rate, the scoring process does foreground that neither the students nor the instructors are the subject of the assessment; rather, the program itself via the outcome is the subject of the assessment.

3.4.1. Personal Participation

The sample included portfolios from courses that I taught. The interrater agreement functioned as a check for the scoring that I completed for portfolios from my own course and from the course of scoring partners. The arrangement for writing assessors to serve in several roles within an assessment design occurs frequently (Donahue, "What is WPA Research," 2013). This multiplicity of roles was accurate even prior to the creation of the role of WPA (*Historical Studies of Writing Program Administration,* L'Eplattenier & Mastrangelo, 2004). To address possible biases in scoring, the assessment procedure includes an interrater agreement component. The scores of paired assessors are compared, and if the scores differ by more than 1, a third reader assesses the portfolio. The third reader provides the decisive score. The interrater agreement indicates whether the scoring reflects a plausible consensus about whether a portfolio satisfies a course outcome. For the portfolio scores that I provided, the interrater agreement was

0, meaning no third readers were necessary, so the scores that I assigned differed by no more than 1 or were the same scores as the scorer paired with me during assessment.

3.4.2. Tracking Course Participation

In addition to the scoring of portfolios, I tracked which portfolios were from courses participating in the Trans-Atlantic and Pacific Project (TAPP) and which ones were not. In order to avoid influencing the scoring process, the scorers did not know that this information was part of the assessment. It was not a major concern, but it was worth avoiding a complicating factor during a labor-intensive process such as scoring portfolios. That is, it avoided addressing questions such as *do I need to score a TAPP portfolio differently* or *do I even have TAPP portfolios* and allowed scorers to focus on the assessment process.

Though tracking participation in the TAPP, I was unable to track specific project types or trace specific partnerships, which often changed each semester. Interestingly, the frequent re-configuration of the TAPP partnerships served as a randomizing element. That is, any measured difference would be attributable to participation rather than a specific collaboration between instructors. Regarding the project types, the collaborations are bilateral projects. A bilateral project is an exchange between two classes, the classes offered in different countries. The students in one class write a primary text of a specific genre; the students in the other class translate the primary text into a target language and provide other localization as needed.

The writing-translation arrangement was the rudimentary form of a bilateral project (Maylath et al., 2013). However, some collaborations exchanged texts once, whereas other collaborations exchanged texts on several occasions. The frequency of text exchange often depended on the students from each collaborating class. Instructors tend to require only one exchange, but some students decided to exchange texts several times. Consequently, one variable of interest that eluded the design of the current study was frequency of exchange. Often the frequency depends upon time constraints and student motivation(s) to exchange; however, I assumed that at least one exchange of texts occurred for each TAPP-participating student portfolio. Exchange frequency and other variables of interest for research on TAPP itself will be discussed further in the limitations and areas for future research. For the scope of the current study, the sole variable of participating or not was sufficient to begin analysis.

3.5. Testing Setup

The collected portfolios scores were recorded in an Excel spreadsheet. The director of upper division writing checked the transcription of scores from paper record to the spreadsheet as part of reporting responsibilities and as a quality assurance step. All mean values were calculated in Excel, and all t-tests were performed using functionality built into Excel. The rationale for using Excel was ease of access. Initially, the data analysis was conducted using R; however, I ceased using R in order to satisfy the criteria for zoomable to be successful it must be accessible as previously described. Though easy to learn then use, the use of R would increase the difficulty of implementing a zoomable assessment (e.g., how to prepare an Excel file to be ported into an R environment and how to manipulate data within R). While the skills to use R might be easily acquired, one frequently noted concern about writing program assessment has been a lack of sustainability in collection, analysis, interpretation, and reporting of program review. Though I wrote a lot of scripts in R to (re)conduct tests initially conducted using Excel, I wish to emphasize that no special statistical software was necessary to perform a zoomable assessment. Writing program administrators have to perform enough tasks, and I already ask that they learn about inferential statistics.

3.5.1. Assumptions for Statistical Tests and Interpretations

Several assumptions are necessary for Welch's t-test. One assumption is that the sampled population has a normal distribution with regard to the variable of interest. While Welch's t-test removes the need for samples to have an equal variance, by retaining the assumption of a normal distribution, the test supposes that a "true mean" value exists. That is, for portfolio scores, the true mean portfolio score will emerge after enough portfolios are scored for the variable of interest. This assumption influenced which courses could be part of the tests at the course level. Only those courses with enough TAPP-participating sections could be included for statistical testing. The assumption supposes any activity (including writing), after enough observations are available, will eventually converge toward a distribution that clusters most of the activity around "average" with few examples exceling or lagging from that large cluster. The debate regards how many observations constitutes "enough" to be confident that the distribution examined is not due to random circumstances but indicative of an underlying pattern. In addition to only large samples, the calibration component of the portfolio scoring process addresses this assumption to some degree. Calibration, therefore, shifts focus from grading or feedback to assessing, and provides constraints on the scorers.

The interpretation of the test results was possible only by consulting the existing scholarship about writing studies and writing programs. Statistical tests identified differences or similarities but explanations for those results resides beyond what numeric data can provide.

4. RESULTS: NUANCES AND NUISANCES

4.1. Introduction

Portfolios provide a rich source of data for a writing program, and the purpose of a zoomable assessment is to illustrate how inferential statistics can improve the writing program assessment design by providing more insights into the collected data beyond descriptive statistics. This chapter presents the descriptive statistics of the portfolio scores and the results of the inferential tests conducted to determine whether the participation in Trans-Atlantic and Pacific Project (TAPP) resulted in difference in mean portfolio scores. Five semesters of portfolio scores are the source of data (n=338). The portfolios are from all courses in the upper division writing program. However, only two courses had enough sections participating in the TAPP to satisfy assumptions for the statistical tests, English 321 Writing in the Technical Professions and English 358 Writing in the Health Professions. The portfolio scoring process used a multi-trait holistic scoring of a program outcome created using dynamic criteria mapping, which is detailed in Chapter 3. The descriptive statistics and test results revealed how a single data set provides nuances and nuisances when interpreting data about a writing program. Depending on the level, I can offer several responses to the question about what might explain the differences in mean portfolio scores.

From the broadest level, the TAPP Integrating Knowledge and Ideas (IKI) scores were greater than the Non-TAPP IKI scores, which suggests the portfolios performed better. However, according to statistical tests, this difference was statistically insignificant. Consequently, the first hypothesis, the null hypothesis (H_0: $\mu_{TAPP} = \mu_{nonTAPP}$), cannot be rejected. This failure to reject is a nuisance for a WPA because the descriptive statistics suggest the TAPP scores are greater than the Non-TAPP scores but the result of the test indicate that the TAPP scores are not greater. The descriptive statistics are a single episode, but inferential statistics are considering what would happen if I continued to collect scores for TAPP and Non-TAPP. Based on the available data, inferential statistics suggest that if I continued to collect scores, eventually TAPP and Non-TAPP would converge even closer to being the same.

Therefore, I sought nuances in the other hypotheses. Following the zooming metaphor, I examined the next level of test results for the samples by course to find more two patterns. For one course, English 321, TAPP IKI scores were significantly greater than Non-TAPP IKI scores, so the null hypothesis is rejected (H_0: $\mu_{321a} = \mu_{321b}$). For another course, English 325, TAPP IKI scores were not significantly different from Non-TAPP IKI scores, so the null hypothesis could not be rejected (H_0: $\mu_{325a} = \mu_{325b}$). These results resolve all three hypotheses posed to address the main research question about whether participation in TAPP resulted in a difference in mean IKI scores in the upper division writing program. In the subsequent sections, I elaborate on the inherent nuances and nuisances within the sample collected to address the research question.

4.2. Descriptive Statistics of Portfolio Data

A portfolio score reflects whether the scorers interpret the portfolio as satisfying the program outcome of Integrating Knowledge and Ideas (IKI). The score does not reflect student performance or instructor performance. This distinction is important because *grading* and *evaluating* are judgment activities aimed to determine student performance in a course, whereas *assessment* is a judgment activity intended to determine how the program is performing relative to the program outcomes. I devoted more time to elaborating on the distinction of various judgment activities in Chapter 2, but it is a distinction worth foregrounding before examining the collected portfolio data. At this level, the assessment focuses on the program and whether the sample portfolios satisfy the targeted outcome. However, from this level, several inferences are possible that inform how other facets of the program could be assessed, such as a pedagogical activity like TAPP collaborations.

When I examine the data zoomed back to the highest level, all portfolios considered as one group, the portfolios had a mean IKI score of 3.28 (±0.86) based on a 1-5 score (Figure 1). The results suggest the scorers assessed the portfolios to be adequate in terms of satisfying the program outcome. Prior to the use of inferential statistics, the assessment would stop at that data point of a mean IKI score for the entire sample of portfolios.

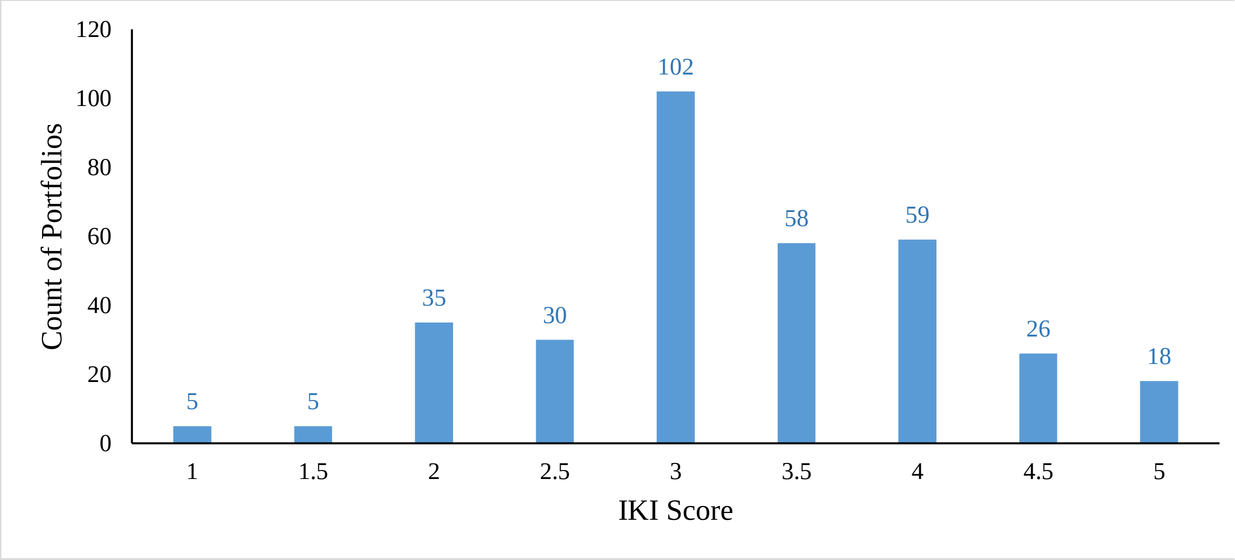

Figure 1 – Portfolio Scores for All Writing Sections Fall 2012 – Spring 2015

The zoomable assessment by design requires further separating the portfolios into categories, courses. After disaggregating the IKI scores by course, I identified a similar pattern of satisfactory performance among all the courses of the writing program. Courses that had a small sample of portfolios tended to follow similar distributions to courses that had larger samples, indicating scorers assigned to assess specific courses found the portfolios satisfied the outcome.

In Table 1, the summary of descriptive statistics and specific distributions of scores indicate normal distributions for each course, and the box plots offer further visual support (Figure 2).

Table 1 – Frequency of Portfolio IKI Scores by Course

			IKI Scores								
Course	Mean	St. Dev.	1	1.5	2	2.5	3	3.5	4	4.5	5
ENGL 320	3.03	0.83	0	2	19	10	30	12	11	5	3
ENGL 321	2.98	1.00	5	3	14	12	20	8	16	3	4
ENGL 322	3.75	0.86	0	0	0	0	5	0	2	1	2
ENGL 323	3.30	0.54	0	0	0	1	2	3	0	1	0
ENGL 324	3.35	0.63	0	0	1	2	14	4	6	3	0
ENGL 325	3.67	0.71	0	0	1	2	17	13	13	8	5
ENGL 357	3.85	0.53	0	0	0	0	1	4	2	3	0
ENGL 358	3.59	0.67	0	0	0	3	11	13	9	2	3
All Courses	3.28	0.86	5	5	35	30	102	58	59	26	18

However, when I selected courses for statistical tests, a few courses had only one section of portfolios (five or fewer portfolios) because the course was not taught every semester. Also, a few courses did not have any TAPP participating sections during the study period, or not enough sections to fulfill assumptions for a meaningful statistical test. Consequently, six courses (English 320, English 322, English 323, English 324, English 357, and English 358) do not have large enough samples to be part of statistical testing to determine whether the TAPP had a significant difference on mean portfolio scores. The main research question, therefore, narrowed in focus to English 321 and English 325, which had large enough samples and both TAPP and non-TAPP sections. Regardless of participation in TAPP, among the courses that have large enough sample sizes (English 320, 321, 324, 325, and 358), no single course was assessed to indicate significant differences in portfolio scores. After comparing mean portfolio scores, course itself did not indicate of significant difference within the data set (p = 0.154).

The distribution of scores by course indicates that no single course outperforms the other courses. That is, English 320 did not perform better than English 321, and so forth. I checked this comparison to eliminate course as a confounding variable to explain any identified differences (Figure 2).

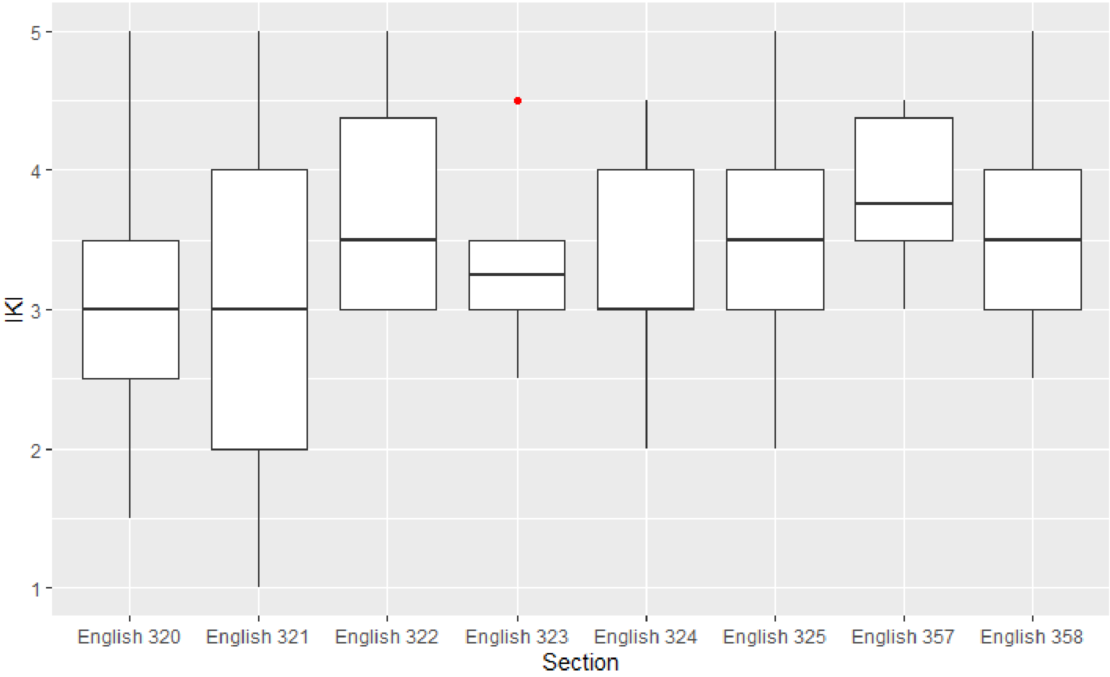

Figure 2 – Portfolio Scores by Writing Course
Counts by Course: English 320 Business and Professional Writing (n=92), English 321 Writing in the Technical Professions (n=85), English 322 Writing and Creative Process (n=10), English 323 Creative Writing (n=10), English 324 Writing in the Sciences (n=30), English 325 Writing in the Health Professions (n=59), English 357 Visual Culture and Language (n=10), and English 358 Writing in the Humanities and Social Sciences (n=42)
NOTE: The median of English 324 is the lower boundary of the 'box'.

The box plots have the median as a solid line and the interquartile range (IQR) is the 'box', marking the lower and upper quartiles of the data (i.e., roughly 50% of portfolios have a score falling within the IQR). The lines extending from the top and bottom of the box are known as whiskers, which extend to the maximum and minimum portfolio score. The red dot for

English 323 denotes an outlier, which would be meaningful, however, English 323 has only seven portfolios so the sample is too small. Therefore, the noted outlier might not be an outlier, and only the collection of more English 323 portfolios could help determine if the distribution is accurate. Again, prior to inferential statistics, the assessment might stop at these new data points of mean IKI scores for each course. However, a zoomable assessment, guided by the research question, pushes for more insight into the program based on a pedagogical activity, participation in a TAPP collaboration.

4.3. Zooming into the Data to Compare TAPP and Non-TAPP

As a group, TAPP portfolios had a mean IKI score of 3.78 (±0.62, n=88), and the Non-TAPP portfolios had a mean IKI score of 3.11 (±0.87, n=250). While the TAPP portfolios had a greater mean, it was not a significantly different amount (T=-7.75, p=0.0837). Therefore, when I consider the main research question regarding TAPP and Non-TAPP portfolios, the two groups are not significantly different, which is a failure to reject the null hypothesis (Failure to reject H_0: $\mu_{TAPP} = \mu_{NonTAPP}$). TAPP portfolios performed in a similar manner to Non-TAPP portfolios (Figure 3).

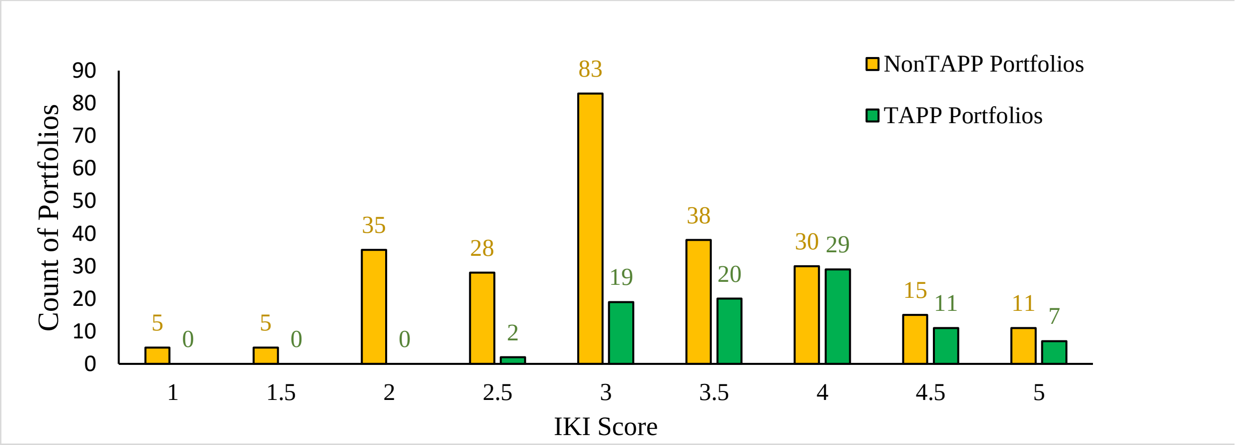

Figure 3 – Comparison of TAPP and Non-TAPP Portfolio Scores

However, the courses that had greater participation in TAPP warranted further examination because participation in TAPP was not evenly distributed among courses. Two courses, English 321 and English 325, had more portfolios that participated in TAPP exchanges. The aggregate portfolios for English 321 Writing in the Technical Professions as a group had a mean score of 2.98 (±1.00, n=85). The TAPP portfolios within this group had a mean of 3.89 (±0.61, n=34), and the Non-TAPP portfolios had a mean of 2.36 (±0.70, n=51). The t-test indicated that this difference was significant ($p = 2.2 \times 10^{-16}$), so I can reject the null hypothesis (Reject H_0: $\mu_{321TAPP} = \mu_{321NonTAPP}$). This suggests the TAPP portfolios did perform better satisfying the IKI outcome.

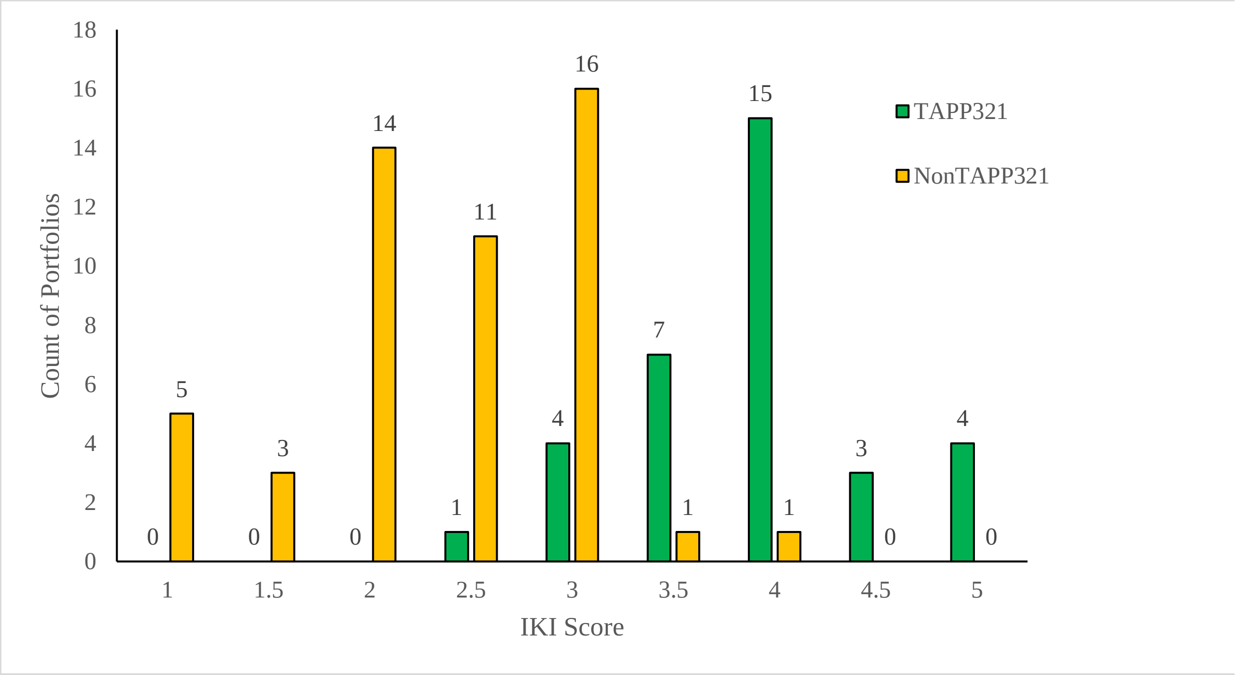

Figure 4 – Comparison of TAPP and Non-TAPP for English 321 Portfolio Scores

The significant difference between TAPP and Non-TAPP in English 321 demonstrates the nuances that statistical tests can reveal. However, English 325 portfolios revealed a different pattern that proves a nuisance when interpreting the data for that course. English 325 also had a similar number of TAPP and Non-TAPP portfolios, so I conducted a t-test on the sample (Figure

5). As a group, the English 325 portfolios had a mean score of 3.67 (±0.71, n=59). The TAPP portfolios had a mean score of 3.80 (±0.66, n=25), and Non-TAPP portfolios had a mean score of 3.57 (±0.74, n=34). Though the TAPP portfolios had a greater mean score, it was not significant (p. = 0.2219), therefore I cannot reject the null hypothesis (Failure to reject H_0: $\mu_{TAPP325} = \mu_{NonTAPP325}$). English 325 was the highest performing course that had more than 20 portfolios. English 357 was the highest performing of any course, but it did not have a large sample size because the course often only had one section per semester and no courses one semester. The high performance of English 325 portfolios raised questions about how to understand the *assessing* judgment activity as not being about student performance and instructor performance.

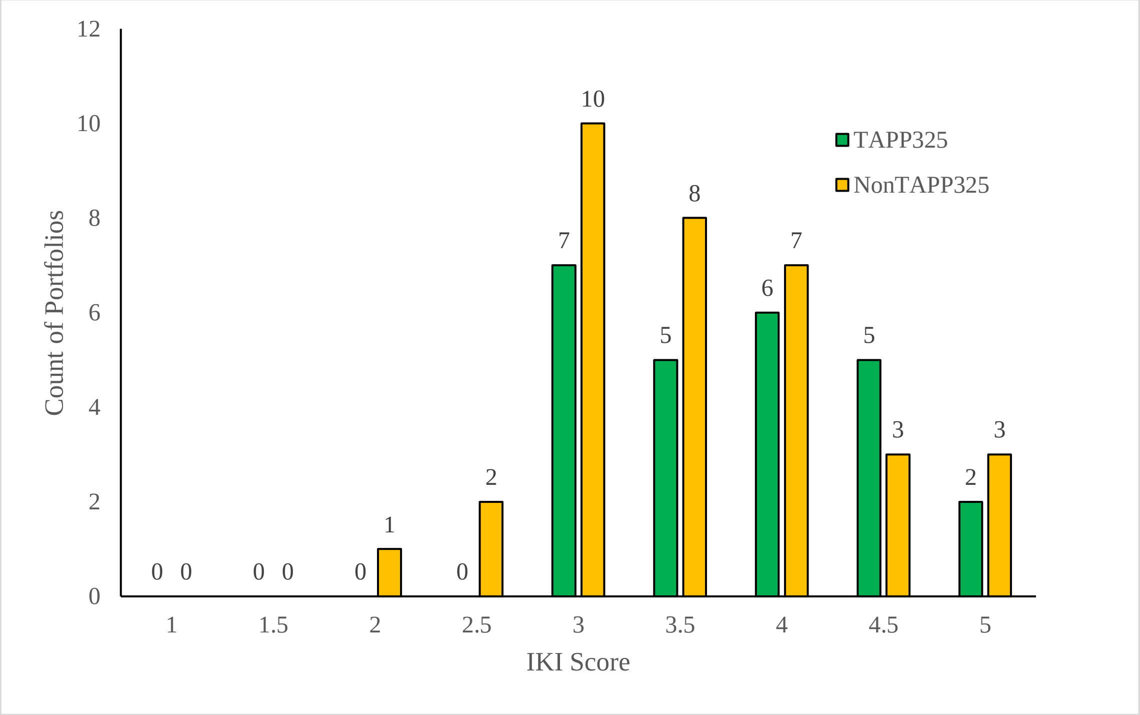

Figure 5 – Comparison of TAPP and Non-TAPP for English 325 Portfolio Scores

In summary, the portfolios satisfy the programmatic outcome of Integrating Knowledge and Ideas. As a demonstration of zoomable assessment, the collected data reveal that a writing program administrator (WPA) can use a single data set, one statistical tool, and several visualizations to gain insight into a program at several levels. At the highest level, the entire upper division writing program, the sole question is whether the portfolios satisfy the programmatic outcome.

The question addresses the highest level by taking into account the entire program. At that level, it is difficult to provide much insight into the program other than whether the portfolios satisfied the programmatic outcome, and I argue it is an artificially clean assessment, serving administrative purposes but not necessarily pedagogical ones. By zooming into the data, the jaggedness of the data becomes apparent. To zoom into the next level, I consider the samples of TAPP and non-TAPP in order to assess performance based on a pedagogical activity. At this level, no significant difference was found between the TAPP and non-TAPP portfolio scores. After I examined the participation in TAPP, course was the next level. Among course, no significant difference was found among the courses. English 325 had the highest portfolio scores based on mean score, and English 320 had the greatest variance in portfolio scores. After zooming to another level, I examined TAPP participation within the individual courses, and I found a significant difference between TAPP and non-TAPP courses for English 321 but found no significant difference for English 325.

The variety of performance at these various levels illustrates the importance of knowing how to ‘zoom’ within a data set in search of nuances and nuisances. A single data set possesses more richness than might initially be apparent by considering only descriptive statistics. The process of testing does not depend upon sophisticated statistical tools or require much modification to data collection procedures. The sole modification was to note which sections participated in the TAPP. The zooming does indicate how even an ostensibly simple data set contains rich insights for a WPA. The rich insights enable a WPA to ask a greater variety of questions about the program and then consider follow-up activities for the program, such as professional development and pedagogical suggestions. Following the argument of Strickland (2011), most WPA work involves “unconscious” negotiation between two types of training: (1) the official training of a WPA based on graduate course work and scholarly activities in composition, literary scholarship, education, or rhetoric, and (2) unofficial training of a WPA while fulfilling the expectations of administering a program. The results suggest that a zoomable assessment frames data collection and interpretation against the balancing of those two training sources. The official training can guide the type of question asked, such as whether a TAPP collaboration results in a difference in mean IKI scores. The ‘unofficial training,’ however, becomes the source for how to interpret the outcomes of inferential statistical tests and apply the interpretations into the maintenance and improvement of the writing program, which I will address in the conclusion of the dissertation.

5. CONCLUSION: FINDING THE CURRENT BOUNDS FOR A ZOOMABLE ASSESSMENT

5.1. Introduction

This dissertation examined whether participation in the Trans-Atlantic and Pacific Project (TAPP) had a significant difference in mean portfolio scores. To evaluate whether participation had a difference in mean portfolio scores, I collected scores from 338 student portfolios then conducted an inferential statistical test to compare mean values of two groups of portfolios, TAPP and Non-TAPP. The examination is a first application of a zoomable assessment. A zoomable assessment is an ensemble of approaches, including inferential statistics, to enable a writing program administrator (WPA) more questions about the program by shifting the perspectives available within a data set.

Writing program assessment tends to concentrate on the highest level, whether the object of assessment (such as student portfolios) offers evidence about the performance of the program (for example, satisfying a program outcome). On the one hand, analysis at that level is useful for demonstrating that the program is functioning in terms that someone outside the writing program might understand. Fulfillment of an outcome is a recognized threshold for realizing curricular goals. On the other hand, the high-level analysis tends to occlude addressing what the high-level data mean for lower levels – instructors of sections, coordinators for courses, and so forth – and consequently, does not provide a WPA with much to offer for explanations or even speculations about program performance. A response that the portfolios earned a score of 3.14 does not offer much without extensive interpretation or additional analysis, which a zoomable assessment offers. Arguably, most active and former WPAs realize assessment designs offer a narrow view of a writing program because assessment scholarship seems to have stalled at multi-trait holistic scoring and portfolios as the preferred unit of analysis. This assessment design might be optimal for most program assessment, but it has limitations that a zoomable assessment can address. For example, a WPA might have interests beyond such a zoomed-out perspective as satisfying a program outcome, however, they might assume the entire assessment design requires modification in order to examine any aspect of the program other than satisfying a program outcome. That level is the intended purpose of the assessment design. This dissertation has sought to demonstrate that an existing design provides more insights without extensive modification and the application of a readily accessible statistical test.

In this chapter, I will discuss the findings provided in the previous chapter and their implications. The results indicate that the TAPP did not have a difference on mean portfolio scores at the highest level, the entire writing program, and had mixed results after zooming into another level, courses of the program. For one course, English 325 Writing in the Health Professions, participation in the TAPP did not seem to have a difference on mean portfolio scores. For another course, English 321 Writing in the Technical professions, participation in the TAPP seemed to have a difference on mean portfolio scores. Based on these identified patterns, a zoomable assessment provides WPAs a model for examining data to learn about a writing program.

As the aphorism attributed to statistician George Box declares, “All models are wrong, but some are useful.” The usefulness of a zoomable assessment emerges from interpretation of results accompanied by delineating current limitations of interpretation. Or, as O’Neill (2015) asserts, “The argument should clarify that the benefits of interpretation and use of the assessment results outweigh the costs” (p. 160). Inferential statistical analysis allows for more questions without more overhead, such as additional software or scoring processes. The sole burden is that, when asking more questions, the WPA also has to wrangle with more responses to those questions. According to the logic in the assertion of O'Neill, the responses require grounding in available scholarship about writing, writing instruction, and writing programs in order to justify spending time to collect data, conduct tests, and interpret the results. Regarding this project, I want to explore what the results mean within the classroom and what inferential statistics could mean for assessment design, including commentary on dynamic-criteria mapping. Phrased another way, what conversations emerged after the integration of an inferential statistical test into the assessment design? How might those conversations facilitate further questions and applications?

5.2. Three Perspectives from a Single Data Set

One reason I suspect inferential statistics are not more widely applied for analysis is that the analysis tends toward subdued results. By subdued results, I mean that the patterns tend to pose more questions rather than propose solutions. Inferential statistics does not identify best practices. Instead, the results from inferential statistical tests indicate whether patterns seem to be non-random. If a pattern is identified as non-random, the interpretative work begins. In the context of the dissertation, the three patterns emerged from the collected data set and the results of the inferential statistical tests. One pattern is that participation in TAPP seems to have a difference in mean portfolio scores for English 321, but the reason for the difference remains unknown. A second pattern is that participation in TAPP did not seem to have a difference in mean portfolio scores for English 325, but the distribution of IKI scores for the course indicates something is happening within the course to produce relatively high scores, whether or not a section participated in TAPP. The third pattern, at the highest level, is that pedagogical activities seem obfuscated by the existing assessment design. That is, I argue that, without zooming, the existing program assessment does not create meaningful data sets for the program. A WPA cannot *draw inferences* about the program based solely on the values for the assessed outcomes. However, I think what often escapes discussion of writing program assessments is the importance of ritual, which I identify as an important component of dynamic criteria mapping. While assessment is one piece of program review (i.e., does the program fulfill the desired outcomes), the activities conducted in performing the assessment build confidence for the WPA that in total the assessment design will produce a meaningful program review. Numbers or narrative independently will not produce a meaningful program review.

5.2.1. Interpreting Results for English 321

Though the results indicated that as a group the TAPP-participating portfolios were not significantly different from the Non-TAPP-participating portfolios, the portfolios for English 321 were significantly different. Several possible explanations exist for the difference. The explanations are all grounded in activities either unique to a TAPP collaboration or an instructional point of greater emphasis due to a section participating in a TAPP collaboration. An example of an activity that is unique to a TAPP collaboration is preparation of a translation brief. A translation brief is a specialized version of a peer review (Appendix). The specialty is to prepare a document for translation from a source language into a target language. The preparation involves attention to linguistic features (such as idiomatic expressions and specialized terminology) and cultural features (such as interpretation of graphical elements or accessibility).

In the translation brief, I note an emphasis not on textual production but on textual consumption. The translation brief reflects a concentrated emphasis to anticipate the needs of a variety of audiences. In this way, the translation brief encourages instructors and students to foreground reader reception. Furthermore, a TAPP collaboration has the benefit of an exchange between producers (students creating a document for a group of readers) and consumers (students translating a document for another group of readers). This type of peer review delivers authentic feedback. If the translators do not understand how to work with the materials – linguistic, graphical, and even genre in some instances – then the report their difficulties to the students producing the materials who attempt to resolve the problems raised by the translators.

This observation extends beyond TAPP-related scholarship to other international collaborations. For example, in *Cross-Cultural Technology Design*, Huatong Sun (2012) notes the advantages of the culturally localized user experience (CLUE) framework in her case studies of Chinese and American students preparing for careers in technical fields. In one case, by tracing student activities, Sun found patterns in temporal consumption of texts. American students wanted to "be heard" and therefore tended to send messages and speak more frequently across collaborative settings. In comparison, Chinese students conducted what one study participant, Mei, dubbed "idioms solitaire." In idioms solitaire, "one [project contributor] would stop [and wait] until others want to join" (Sun, 2012, p. 176). That is, production proceeds only as others join and respond to one another, much like building stacks in the card game of solitaire, move by move, message by message. Sun noted that textual production happened as each participant contributed rather than waiting for each participant to unleash an entire finished text. If one contributor does not cease adding material, the other contributors presume that the active contributor does not want them to participate. Similarly, a contributor might interpret another participant joining not as a sign to stop, but to increase production only to become frustrated when the participant stops contributing – presuming that they had interrupted. The collaboration is a complex negotiation of cultural expectations and requires attentiveness to a unique awareness of textual consumption during the written composition processes.

Of course, I am speculating based on what I know about a TAPP collaboration and what writing studies and translation studies have established in previous scholarship (see Chapter 2). The result for English 321 supports the scholarship and narrows the type of subsequent research questions to ask about a written composition course intended for engineers. I note the idiom solitaire because it pertains to one question that still intrigues me about TAPP collaborations, and I find nothing in writing studies scholarship: does length of time period for feedback between collaborators produce a difference regarding students learning about written composition? While students in a TAPP collaboration often bemoan long periods of no interaction with collaborators, the process of learning patience with irresolution (not receiving feedback immediately) is important. I have found evidence for this in mathematics education (see Dan Meyer 2009). However, I was unable to locate any research in writing studies dedicated specifically to the question of *what is a beneficial rate of feedback*. Patience is a well-documented characteristic of teaching writing. In “Teach Writing as a Process Not a Product,” Donald Murray (1972) observed, “To be a teacher of process such as [I, Murray, am suggesting] takes qualities too few of us have, but which most of us can develop. We have to be quiet, to listen, to respond. […] We are the reader, the recipient. We have to be patient and wait, and wait, and wait” (p. 5). Patience is difficult to incorporate into a hectic semester schedule, especially given that many writing courses have inherited a coverage model of syllabus from literary studies. A coverage model presumes a known amount of material is necessary to include for a course to be deemed successful. In literature, a coverage model presumes certain canonical authors and their works are necessary reading in order for a student to be considered knowledgeable and prepared to conduct scholarship. David Smit (2008) argues the coverage model in writing studies is a consequence of a post-Dartmouth Conference of 1966 proposal by James Kinneavy in *A Theory of Discourse* (1971) for a "discourse [writing] grouped in terms of four purposes or aims: expressive, referential, persuasive, and literary discourse" (p. 188).[11] A writing course ought to cover all four purposes. Admittedly, time is always a limited resource, so attempts to select what material to cover within a set timeframe are necessary. While I succumb to coverage-like approaches to structuring my own sections, upon reflection I notice that participating in a TAPP collaboration allows for patience – wait, and wait, and wait.

Murray offers useful principles about teaching written composition, and the TAPP collaborations seem to support the idea of developing patience. However, I do not know if the rate of feedback through the translation brief and comments on documents is a strong explanatory factor. As a future application of inferential statistics, I would want to design a research project that tested whether longer feedback periods resulted in a detectable difference in student learning based on portfolio scores but perhaps additional variables at a level zoomed closer to the section-level of the program. Pragmatically and operationally, the loop of feedback is bound by constraints, such as the course has a schedule and the students and instructors have

[11] The "Anglo-American Seminar in the Teaching of English" (i.e., the Dartmouth Conference) of 1966 is an important influence on the design of writing instruction. For excellent and thorough archival research on the Dartmouth Conference, Annette Vee of the University of Pittsburgh has collected and curated the presentations, correspondence preparing the organization, subsequent publications, and more at the "Dartmouth '66 Seminar Exhibit," which is housed in the WAC Clearinghouse. The archive is a dissertation project for someone interested in the history of writing instruction. The Conference also was a so-called peace accord between the National Council of Teachers of English (NCTE) and Modern Language Association (MLA). Amid much heated correspondence, Americans and Canadians (Canadian Council of Teachers of English) wanted the conference to be a symbolic passing of the Anglophone world from England to America. These trans-Atlantic interactions piqued my interest, and the archive is saturated with research potential about writing instruction from the 1970s to present: https://wac.colostate.edu/resources/research/dartmouth

other commitments and obligations. Feedback is a judgment activity outside of the scope of this dissertation, but the result from the statistical analysis of English 321 suggests that writing studies ought to consider the temporal as much as the content of feedback. Does the delay of feedback actually provide a type of quality? Regarding the content of feedback, I prefer to direct the interpretation toward English 325.

5.2.2. Interpreting Results for English 325

The IKI scores for English 325 indicate the course consistently scored high during the sampling period. The scores provide an opportunity for additional use of inferential statistics in order to investigate possible reasons for the high scores. Given that portfolios from TAPP and Non-TAPP sections scored high, one question is whether the students of each group had subsequently high performance elsewhere, such as a senior project. Admittedly, I could pose a similar question of the students in the TAPP-participating English 321 sections. That is, students participating in a TAPP collaboration scored higher but it remains undetermined what if anything they retained in courses or projects after completing English 321. However, the English 325 results raise questions about how the courses produce high-scoring portfolios. One option is a longitudinal study of students prior to enrolling in English 325 and after completing English 325.

From an assessment perspective, longitudinal data is difficult to obtain and interpret for a variety of reasons. Primarily, longitudinal data often focus on individual students or programs as case studies (Smith, Girdharry, & Gallagher, 2021; Blair, 2017; Inoue, 2015; Ross and Arnett, 2013; Dixon and Moxley, 2012). The case studies treat production as the main interest. Smith, Girdharry, and Gallagher (2021) followed twenty student writers to examine how the students transferred textual production practices into subsequent courses, proposing students learn through a series of instances that they characterize as integration: “A theory of writing develops [for students] unevenly and often over long periods of time. Often, transfer episodes (positive or negative) reveal themselves in the fullness of time to be embedded within larger processes of integration" (Smith, Girdharry, & Gallagher, 2021, p. 20). While these studies focus upon textual production and qualitative coding of written reflections by students, I think an interesting opportunity is present when I consider the result for English 325 and uses for inferential statistics. In *(Re)Articulating Writing Assessment for Teaching and Learning,* Brian Huot (2002) claims that most writing assessment at the course level focuses on student textual production (essay tests, portfolios with reflective letters, written student self-assessments based on course outcomes), and not on student textual consumption (reading comprehension, quality of peer review, research methods).

> "These practices [of privileging production over consumption] ultimately deny that linguistic, rhetorical and literate capabilities can only be developed within the context of discovering and making meaning with the written word. We [writing studies scholars] have yet to create in any substantive way a pedagogy that links the teaching and assessing of writing" (Huot, 2002, p. 61).

A major issue is how to assess textual consumption. Among available instruments (e.g., multiple-choice exams or written summaries of sources), nothing points toward use of textual content by students when they produce their own textual content.

Huot's conclusion returns my thoughts to the exchange I outlined in Chapter 1 in which Doug Hesse and Joseph Teller expressed differing ideas about how to design a written composition course. In particular, Teller wanted a separate course dedicated to reading (consumption) of texts, noting it was a part of learning to improve as a writer that had been integrated into writing, thereby arguably diminishing its importance by removing explicit instruction in reading. Inferential statistics, a zoomable assessment, can pose questions about textual consumption that evade current assessments. In examining TAPP collaborations among English 325 sections, I may have inadvertently identified the start of an assessment design that emphasizes through explicit activities textual consumption (translation preparation and reviews with authentic readership). However, I hesitate to push the idea much further, aside from noting that the course outcome, Integrating Knowledge and Ideas, lends itself to examining whether the English 325 students sustain levels of textual production and, inferably, textual consumption in subsequent courses.

In addition, the English 325 results raise questions about participating in a TAPP collaboration based on the apparent lack of difference in scores. By including activities that directly relate to textual consumption, I would want to investigate the concept of collateral learning. John Dewey (1938) started to formalize collateral learning in *Experience & Education*:

> Perhaps the greatest of all pedagogical fallacies is the notion that a person learns only the particular thing [they are] studying at the time. Collateral learning in the way of formation of enduring attitudes, of likes and dislikes, may be and often is much more important than the spelling lesson or lesson in geography or history that is learned. (p. 48)

Collateral learning has subsequently become conceptualized as a “hidden curriculum,” in which ways of thinking and writing for a specific discipline are implicitly taught and, because these ways are only implicitly taught, never considered for assessment (see Paul Prior (1998) *Writing/Disciplinarity: A Sociohistoric Account of Literate Activity in the Academe*). The concept of a hidden curriculum also suggests that many attitudes toward written composition require dislodging because students (and instructors) have amassed an enduring attitude often of dislike toward writing. Writing scholars and instructors are well aware of the dislikes that students, instructors, and even administrators associate with writing; it is a topic notably addressed by Cheryl Ball and Drew Loewe in the open-access edited collection *Bad Ideas about Writing* (2017). In a TAPP collaboration, I think students approach the processes of written composition in a refreshingly new way. That is, a TAPP project encourages participants to contemplate language in a way that is unique relative to many pedagogical activities. The complaint associated with collateral learning is that it defies measurement aside from questionnaires about rating enjoyment, which is not a reliable indicator of learning or fulfillment of a course outcome.

To consider a type of collateral learning, Webber (2017) examined machine-scoring of essays on learning written composition. Weber draws upon John Dewey's pragmatism to propose the concept of artfulness critique. An artful critique is one that diminishes or avoids "the gulf between experts and publics 'being bridged not by the intellectuals but by the investors and engineers hired by captains of industry'" (Webber, 2017, p. 139). Webber quotes Dewey in defining an artful critique as part of the pushback against machine-scoring essays. Machine-scored essays risk the essay writers becoming capable test-takers but not capable writers. Instead of teaching to the test, it becomes a case of teaching to the machine (see Watters *Teaching Machines* for a history of the phenomenon in several areas but especially math education). In contrast, writers who must interact with audiences – such as collaborators for a TAPP project – learn the importance of processes in writing, including the role of reader reception.

The results obtained for English 325 provide a testing ground. Both the TAPP and the Non-TAPP sections of English 325 have high performing portfolios. It is worth pursuing whether the students maintain high performance in a subsequent writing-intensive course. What is the distribution of scores in a senior capstone? One immediate issue is that this proposed follow-up test requires zooming closer toward individual student performance, which could require creating a more elaborate model of a writing program than the current zoomable assessment offers.

On the topic of a more elaborate model of a writing program, the question about outcome performance after a course or before enrolling in a course will allow for more research on a vertical writing curriculum. In concept, a vertical writing curriculum involves a sequence of courses, typically a first- or second-year written composition course and then another upper division written composition course. A third and final course is often part of the curriculum, but it often does not share the same outcome, except perhaps implicitly, such as "improve communication in various forms." The existing models use predictive statistics to ask if student performance in first-year can account for variability in performance in upper division or senior projects. In contrast, inferential statistical tests would focus on questions such as should students wait until they earn a specific number of credits before enrolling in upper division courses? How long? Or, should the upper division course be available immediately after first-year courses so students might consolidate knowledge? By examining fulfillment based on groups, much the way I grouped portfolios into TAPP and Non-TAPP, a WPA could learn about how to structure the curriculum. At present, I argue that curricular structure is based on a few studies (in part because few universities have a vertical writing curriculum) and unexamined suppositions about learning. For example, apparently, Ross and Arnett (2012), in a proposal for how to design a graduate-level research methods course, is one of a few examples to direct attention to the importance of consumption as much as production.

The assessment of a vertical curriculum is ideal for a zoomable assessment. A vertical curriculum has opportunities to trace student development from first-year courses into subsequent courses, and all of the courses are writing-intensive. However, I recognize that the dearth of studies is probably the logistical and operational undertaking of longitudinal research requires resources not readily available to most WPAs (see *Composition in the Age of Austerity, ed.* Welch and Scott). However, by applying inferential statistics, I am optimistic that such long-term projects could be divided into manageable smaller projects and more focused by allowing statistical testing to guide resource allocation. For example, based on the analysis of TAPP and Non-TAPP, I would know to focus on English 325 and conduct more conversations with the related interests – instructors of the course, instructors of subsequent courses whether in the department or in another department. By zooming into a feature, the assessment allows for the overall program review to become an important ritual.

5.3. The Rituals of Data Collection and Interpretation

Within writing program assessment scholarship, a frequent theme is that assessment requires more than an individual. Of course, the task of programmatic assessment might become the responsibility of one person, the WPA. However, the program has students and instructors, curricular requirements, outcomes (university or department). Therefore program review is enmeshed in a thick coating of activities. Therefore, regardless of how assessment happens, the conceptualization of the writing program is of complex flows of information. The complexity is apparent in the conceptual metaphors about writing and writing programs. The three frequently used conceptual metaphors are ecologies (Branson, et al., 2017; Inoue, 2015; Reiff, Bawarshi, Ballif, & Weisser, 2015; Ryan, 2012), networks (Rice, 2011), and assemblages (Yancey & McElroy, 2017). All three metaphors imagine a program as a series of relationships, though imagine the nature of the relationships in different ways. However, a shared aspect among all the conceptual metaphors is establishing procedures and processes, which I wish to re-imagine as rituals because rituals involve meaning making. Arguably, the most important rituals are data collection and interpretation because those activities aid to some degree everyone in the ecology, network, or assemblage to ascribe purpose and meaning to the relationships.

The role of inferential statistics within the ritual appears to be re-enchantment. Descriptive statistics provide an image, but they are static instances that cannot inform practices. As previously noted, a mean for IKI scores on its own does not help instructors determine if an activity helps to fulfill an outcome. This observation is consistent with claims about finding ways to use data rather than merely collect it. For example, Barker (2012) completed a survey of the Council for Programs in Technical and Scientific Communication regarding program assessment and the purpose of outcomes. The outcomes were deemed appropriate, but a major concern was communication about program performance in satisfying the outcomes. Inferential statistics join the ensemble of a zoomable assessment and aid me in ascribing meaning to collected data rather than have descriptive statistics to "fill the file" (White, Elliot, and Peckham, 2015) and possibly also avoids the concept of numeric data being "foiled against a teaching grounded in humanism" (Yancey, 1999, p. 495). Inference puts description into motion, taking otherwise inert data points and infuses them with interpretative potential.

Similarly, a program narrative shapes identity and provides goals and language to convey what a program values, but conversations about whether or not the program is attaining those values remains elusive. Broad provided the framework of dynamic criteria mapping (DCM) as a means of articulating values, but the framework is intended to be sustained through conversations, which might also necessitate altering the program if evidence is not available that the values are infused into the outcome and observable in the labor of the involved participants (instructors, students, and administrators). In the introduction to *Organic Writing Assessment*, a collection of case studies about dynamic criteria mapping, Broad notes, “Along the way, these DCM explorers worried about whether the adaptions and compromises they made were ‘legitimate’ in relation to DCM praxis. My response to this concern brings us back to the beginning of this process, to the beginning of my earlier book [*What We Really Value]*, and to Piercy’s poem ‘To Be of Use.’” (p.11). I think WPAs who use dynamic criteria mapping forget what *dynamic* entails and instead fret over *criteria* (outcomes) and *mapping* (methods). The focus upon TAPP within an upper division writing program demonstrates that assessment can be responsive to organizational circumstances, and inferential statistics provides the means, the dynamism, to conceive of ways to determine if the program is delivering on its values.

To achieve a zoomable assessment, a manageable modification to an existing design is the only necessity. By manageable, I connect the concept of zoomable to Donna Strickland’s (2011) call for operational reasoning in administering writing programs. Strickland argues that operational reasoning happens through *tweaking* the various elements of a writing program. When describing efforts from her own experiences, Strickland notes the necessity for conversations, with people directly in the program and with people indirectly (other departments and even non-university entities such as K-12 educators and employers), in order “to begin to *tweak* the portfolio system, to experiment with more context-rich, small-scale portfolio assessment, one that left room for plenty of teacherly autonomy” (p. 120). Inferential statistics represent a method to tweak in order to learn about a writing program by providing richer insights, and the directed nature of the tweak is small-scale. I decided to zoom into the upper-division portion of the writing program because it had an existing diversity of courses. The variety of courses provides a pre-existing grouping, which eased collecting the material. The decision to include TAPP as a variable of interest stemmed from noting that the activities of a

TAPP collaboration felt unique. When to zoom (or even to tweak, referring to Strickland's concept) is a complicated question without participation in a program (see Salvo & Ren, 2007, "Participatory Assessment: Negotiating Engagement in a Technical Communication Program").

The richness of the insights stems from changing the scale of an assessment without interrupting pedagogical practices. Much as Strickland (2011) was responding to James Berlin's idea that research in pedagogy reflected "the search for the one best way [of teaching], the normalizing quest, [that] resonates with traditional systematic management aims" (p. 106), I think inferential statistics provides a response to traditional writing program administration aims, which are constrained to predictive statistics (regression models to anticipate student performance) and descriptive statistics (reporting only mean scores to provide reports for administration purposes). Through a shift to include an inferential statistical test to inquire about participation in TAPP, I was able to identify which level requires additional attention, and what questions the instructors and the WPA might ask about their sections or the course more broadly.

5.4. Bounds for Zooming and Inferential Statistics: Future Directions

In 1997, Stephen Tchudi wrote a summary of the findings by the Committee on Alternatives to Grading Student Writing, a committee created by the National Council of Teachers of English. The committee had the charge to research best practices for writing instructors to provide judgments of students and their writing. Tchudi borrowed a concept, *degrees of freedom,* to frame the main finding. "We [the Committee] think it is useful to conceive of the problem [of approaches to grading student writing] by adapting a concept from math and science: 'degrees of freedom'. […] Changing parameters or restrictions often opens up new areas of freedom, but just as often results in the loss of other directions of movement" (Tchudi, 1997, p. xii). The consequence of relying on any one type of judgment (grades, written feedback, in-person discussions) would be a narrowed view of the activities involved in writing. Each judgment activity had a role in supporting students in learning. Grades, *per se*, were not the problem. The problem was instructors depended upon grades to perform a purpose for which the grading system was not intended: help students self-evaluate their understanding of important concepts, whether in writing or in chemistry or any other subject matter. Grades have an origin in the idea of sorting by social class which has morphed into a variety of standards (Inoue, 2015); however, the use of grades has evolved. While grades can help to guide student learning, Tchudi argued against relying solely on them. That is, alternatives to grading did not necessarily mean eliminate grades but find complements to them, ones that had degrees of freedom for instructors to support students as needed.

By applying inferential statistics as a complementary component to an existing assessment design, I aimed to accomplish a similar insight for judgments about programmatic performance that Tchudi provided for section level judgments (instructor-student interaction). Toward that end, I think it is also important to recognize the limitations of what inferential statistics can accomplish, and to accomplish that recognition, I will borrow a phrase: *bounds*. Inferential statistics have boundaries for which they are not suited to resolve issues. One of the significant bounds is guidance during interpretation of results.

Regarding bounds, it is worth recalling that any interpretation ought to adhere to the limits of what available evidence permits. I proposed a zoomable assessment because I perceived that the existing numeric data did not facilitate much use of the existing assessment design. The design produced a number, a mean score for an outcome, but did not attempt to extract more meaning from that number. Inferential statistics provided a way to expand the bounds of the existing assessment by generating more data points, basically more mean scores for more specific parts of the writing program then using these data points to conduct statistical tests to determine if there were meaningful patterns associated with those selected parts of the writing program. The interpretation of the results of the statistical tests remained situated within the scholarship of writing studies and of writing program administration. I do not know whether an upper bound, the farthest expansion of meaningful interpretation can be reached. I propose that data, statistics, and evidence only tear loose from meaning when the processes and procedures to gather them are divorced from the people who need data, statistics, and evidence to improve their circumstances. And a dilemma of assess or be assessed, attributed to Edward White, looms large.

In the 2017 CCCC Chair address, Linda Adler-Kassner outlined the emergence of the education intelligence complex, which is conceptually modelled after the military industrial complex. Adler-Kassner argued a series of periodic "crises" including *A Nation at Risk* in 1983, *Ready or Not* in 2004, and *Putting Students First* in 2016 has enabled a thriving group of companies to package education into readily measured, quantified, and reportable data points. By packaging education in this manner, the metaphor is that education shifts from being a "maze" to a "chute" (Adler-Kassner, 2017, p. 325). To accomplish the feat of disentangling the loops of learning into a linear process, the companies resort to predicative analytics. As part of a rebuttal to this approach, Adler-Kassner cites Simon Buckingham Shum: "Data points on a graph are tiny portholes onto a rich human world, [but they] do not do justice to the complexity of real people, and the rich forms that learning take" (p. 326-327). The metaphor of data points as portholes is appealing because it recognizes the view is restricted to the dimensions of the viewing port, the bounds. With inferential statistics and the nuances and nuisances that they reveal, a zoomable assessment is a way of encouraging more portholes rather than trusting that a prescribed or assigned view is appropriate. Based on the use of inferential statistics to determine if TAPP participation resulted in a difference of mean portfolio score, I am confident that inferential statistics, a zoomable assessment, offer WPAs ways to open more portholes, consequently seeing more of the rich forms of learning happening within a writing program.

REFERENCES

Adler-Kasner, L. (2017). Because writing is never just writing. *College Composition and Communication, 69*(2), 317-340.

Ball, C. E. & Loewe, D. M. (2017). *Bad ideas about writing*. West Virginia University Libraries Digital Publishing Institute.

Barker, T. (2012). Program assessment: The role of outcomes. *Programmatic Perspectives, 4*(2), 183-210.

Bordelon, S., Wright, E.A., & Halloran, S.M. (2012). From rhetoric to rhetorics: An interim report on the history of American writing instruction to 1900. In J.J. Murphy (Ed.), *A Short History of Writing Instruction: From Ancient Greece to Contemporary America* (pp. 209-231). Routledge.

Branson, T.S., Sanchez, J.C., Robbins, S. R., & Wehlburg, C.M. (2017). Collaborative ecologies of emergent assessment: Challenges and benefits linked to a writing based institutional partnership. *College Composition and Communication, 69*(2), 287-316.

Brereton, (1995). *The origins of composition studies in the American college, 1875-1925.* University of Pittsburgh Press.

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

Broad, B. (Ed.). (2009). *Organic writing assessment: Dynamic criteria mapping in action*. Utah State University Press.

Bosley, D.S. (2010). Do fish know they are swimming in water? In J. Allen & N. H. Margaret (Eds.), *Assessment in technical and professional communication* (pp. 221-224). Routledge.

Chenoweth, N. A., Hayes, J. R., Gripp, P., Littleton, E. B., Steinberg, E. R., & Van Every, D. A. (1999). Are our courses working?: Measuring student learning. *Written Communication, 16*(1), 29–50.

Condon, W. & Rutz, C. (2012). A taxonomy of writing across the curriculum programs: Evolving to serve broader agendas. *College Composition and Communication, 64*(2), 357-382.

Connors, R. (1996). *The abolition debate in composition: A short history*. Southern Illinois University Press.

Cordell, R. (2013) “Taken possession of”: The reprinting and reauthorship of Hawthorne’s “Celestial Railroad” in the antebellum religious press. *Digital Humanities Quarterly, 7*(1), 1-21.

Cordell, R. & Smith, D.A. (2022). The viral texts project. https://viraltexts.org/

Cox M, Galin, J.R., & Melzer, D. (2018). *Sustainable WAC: A whole systems approach to launching and developing writing across the curriculum programs.* NCTE.

Creswell, J.W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches.* SAGE.

Dewey, J. (1938). *Education and experience*. Macmillan.

Diogenes, M. & A.A. Lunsford. (2006). Toward delivering new definitions of writing. *Delivering College Composition: The Fifth Canon*. 141-154.

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing*, *18*(4), 241–256.

Drucker, J. (2021). Sustainability and complexity: Knowledge and authority in digital humanities. *Digital Scholarship in the Humanities 36*(2), 86-94.

Ekiba, H. Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S., & Sugimoto, C.R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523-1545.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing*, *1*(1), 91–107.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Fleckenstein, K., Spinnuzi, C., Rickly, R., & Papper, C. (2008). The importance of harmony: An ecological metaphor for writing research. *College Composition and Communication*, *60*(2), 388–419.

Gallagher, J.R., Chen, Y.Y., Wagner, K., Wang, X., Zong, J.Y., & Kong, A. L. (2020). Peering into the internet abyss: Using big data audience analysis to understand online comments. *Technical Communication Quarterly, 29*(2), 155-173.

Gere, A. R., Curzan, A., Hammond, J.W., Hughes, S., Li, R., Moos, A., Smith, K., Van Zanen, K., Wheeler, K. L., & Zanders, C. J. (2021). Communal Justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication, 72*(3), 384-412.

Gold, D., Hobbs, C.L., & Berlin, J.A. (2012). Writing instruction in school and college English: The twentieth century and the new millennium. In J.J. Murphy (Ed.), *A Short History of Writing Instruction: From Ancient Greece to Contemporary America* (pp. 232-272). Routledge.

Hammer, S. & Maylath, B. (2014). Real time and social media in trans-Atlantic writing/translation and translation/editing projects. In M. Limbu & B Gurung (Eds.), *Emerging Pedagogies in the Networked Knowledge Society: Practices Integrating Social Media and Globalization* (pp. 144-161). IGI Global.

Hamp-Lyons, L. (2016). Farewell to holistic scoring? *Assessing Writing, 27*, 1-5.

Hamp-Lyons, L. (2011). Writing assessment: Shifting issues, new tools, enduring questions. *Assessing Writing, 16*(1), 3–5.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*(1), 5-16.

Haswell, R. H. (2009). Teaching of writing in higher education. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 331-346). Lawrence Erlbaum Associates.

Haswell, R. H. (2005). Ncte/cccc's recent war on scholarship. *Written Communication*, *22*(2), 198–223.

Haswell, R. H., & Wyche-Smith, S. (1994). Adventuring into writing assessment. *College Composition and Communication*, *45*(2), 220-

Hayes, J. R., & Hatch, J. A. (1999). Issues in measuring reliability: Correlation versus percentage of agreement. *Written Communication*, *16*(3), 354–367.

Horton, (2018). Trans-scalar challenge of ecology. *Interdisciplinary Studies in Literature and Environment, 26*(1), 5-26.

Huot, B. (2002). *Re-articulating writing assessment for teaching and learning.* Utah State University Press.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication*, *41*(2), 201.

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*(2), 237–263.

Huot, B. & Schendel, E. (1999). Reflecting on assessment: Validity inquiry as ethical inquiry. *Journal of Teaching Writing, 17*(1), 37-55.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. The WAC Clearinghouse.

Inoue, A.B. (2009). The technology of writing assessment and racial validity. In C.S. Schreiner (Ed.), *Handbook of Research on Assessment Technologies, Methods, and Applications in Higher Education* (pp. 97-120). IGI Global.

Jockers, M. L. (2013). *Macroanalysis: Digital methods and literary history.* University of Illinois Press.

Jones, M. P. (2019). (Re)placing the rhetoric of scale: Ecoliteracy, networked writing, and Memorial mapping. In S. I. Dobrin & S. Morey (Eds.), *Mediating Nature: The Role of Technology in Ecological Literacy* (pp. 79-95). Routledge

Kinneavy, J. L. (1980). *A theory of discourse: The aims of discourse.* Norton.

Klein, (2013). Image of absence: Archival silence, data visualization, and James Hemings. *American Literature, 85*(4), 661-688.

Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing, 31*, 84–97.

Lang, S., & Baehr, C. (2021). Data mining: A hybrid methodology for complex and dynamic research. *College Composition and Communication*, *64*(1), 172–194.

Lea, M. R. & Street, B.V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172.

Maylath, B., King, T., Arno Macia, E. (2013). Linking engineering students in Spain and technical writing students in the US as coauthors: The challenges and outcomes of subject-matter experts and language specialists collaborating internationally. *Connexions: International Professional Communication, 1*(2), 150-185.

Maylath, B., Vandepitte, S., Minacori, P., Isohella, S., Mousten, B., & Humbley, J. (2013). Managing complexity: A technical communication/translation case study in multilateral international collaboration. *Technical Communication Quarterly, 22*, 67-84.

Meloncon, L. & St. Amant, K. (2019). Empirical research in technical and professional communication: A 5-year examination of research methods and a call for research sustainability. *Journal of Technical Writing and Communication, 49*(2), 128-155.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2), 5-11.

Meyer, D. (2009). Impatience with irresolution. *Dy/Dan.*

Miller, B. (2014). Mapping the methods of composition/rhetoric dissertations: A "landscape plotted and pieced." *College Composition and Communication, 66*(1), 145-176.

Moretti, F. (2000). The slaughterhouse of literature. *Modern Language Quarterly, 61*(1), 207-227.

Mousten, B., Maylath, B., Vandepitte, S., & Humbley, J. (2010). Learning localization through trans-Atlantic collaboration: Bridging the gap between professions. *IEEE-Transactions on Professional Communication, 53,* 401-411.

Mueller, D. (2012). Grasping rhetoric and composition by its long tail: What graphs can tell us about the field's changing shape. *College Composition and Communication, 64*(1), 195-223.

Murphy, S. (1994). Portfolios and curriculum reform: Patterns in practice. *Assessing Writing*, *1*(2), 175–206.

Murphy, S. & Yancey, K.B. (2009). Construct and consequence: Validity in writing assessment. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 365-385). Lawrence Erlbaum Associates.

Murray, D. (1972). Teaching writing as a process not a product. In V. Villanueva & K.L. Arola (Eds.) *Cross-Talk in Comp Theory* (pp. 3-6). NCTE.

North, S. M. (1987). *The making of knowledge in composition: Portrait of an emerging field*. Boynton/Cook Publishers.

O'Neill, P., Moore, C., & Huot, B. A. (2009). *A guide to college writing assessment*. Utah State University Press.

Penny, J. A., & Johnson, R. L. (2011). The accuracy of performance task scores after resolution of rater disagreement: A Monte Carlo study. *Assessing Writing*, *16*(4), 221–236.

Poe, M., Elliot, N., Cogan Jr., J. A., & Nurudeen Jr., T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication 65*(4), 588-611.

Prior, P. (1998). *Writing/disciplinarity: A sociohistoric account of literate activity in the academy*. Routledge.

Prior, P., & Lunsford, K. (2009). History of reflection, theory, and research on writing. In C. Bazerman (Ed.), *Handbook of Research on Writing: History Society, School, Individual, Text* (pp. 81–96). Lawrence Erlbaum Associates.

Raucci, J. (2021). A replication agenda for composition studies. *College Composition and Communication*, *72*(3), 440–461.

Reiff, M.J., Bawarshi, A., Ballif, M., & Weisser, C. (2015). Writing program ecologies: An introduction. In M.J. Reiff, A. Bawarshi, M. Ballif, & C. Weisser (Eds.), *Ecologies of writing programs: Program profiles in context* (pp. 3-18). Parlor Press.

Rice, J. (2011). Networked assessment. *Computers and Composition*, *28*(1), 28–39.

Ross, D.G. & Arnett, E.J. (2012). To do is to learn: The value of hands-on research in an introductory research methods course. *Programmatic Perspectives*, *5*(2), 214-242.

Ryan, K. J. (2012). Thinking ecologically: Rhetorical ecological feminist agency and writing program administration. *WPA: Writing Program Administration*. *36*(1), 74-94.

Scott, T., & Brannon, L. (2013). Democracy, struggle, and the praxis of assessment. *College Composition and Communication*, *65*(2), 273–298.

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, *17*(2), 81–91.

Slomp, D. H. (2019). Complexity, consequence, and frames: A quarter century of research in Assessing Writing. *Assessing Writing*, *42*, 100424.

Smith, K.G., Girdharry, K., & Gallagher, C.W. (2021). Writing Transfer, Integration, and the Need for the Long View. *College Composition and Communication*, *73*(1), 4-26.

Sorensen, K., Hammer, S., & Maylath, B. (2015). Synchronous and asynchronous online international collaboration: The trans-Atlantic & Pacific project. *Connexions*, *3*(1), 153-177.

Spinuzzi, C. (2015). Toward a typology of activities: Understanding internal contradictions in multiperspectival activities. *Journal of Business and Technical Communication*, *29*(1), 3–35.

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

Strickland, D. (2001). Taking dictation: The emergence of writing programs and the cultural contradictions of composition teaching. *College English*, *63*(4), 457.

Stark-Meyerring, D. & Andrews, D.C. (2010). Assessment in an intercultural virtual team project. In J. Allen & N. H. Margaret (Eds.), *Assessment in technical and professional communication* (pp. 197-220). Routledge.

Steinmann, H., Saduov, R., & Maylath, B. (2016). Learning across boarders: A teaching case connecting writing students internationally. *Konin Language Studies, 3*, 271-287.

Tchudi, S. (1997). Degrees of freedom in assessment, evaluation, and grading. In S. Tchudi (Ed.), *Alternatives to grading student writing* (pp ix-xvii). NTCE.

Verzella, M. & Tommaso, L. (2014). Learning to write for an international audience through cross-cultural collaboration and text-negotiation. *Changing English: Studies in Culture and Education, 21*(4), 310-321.

Wardle, E., & Roozen, K. (2012). Addressing the complexity of writing development: Toward an ecological model of assessment. *Assessing Writing*, *17*(2), 106–119.

Watters, A. (2021). *Teaching machines: The history of personalized learning*. The MIT Press.

Webber, J. (2017). Toward an artful critique of reform: Responding to standards, assessment, and machine scoring. *College Composition and Communication*, *69*(1), 118-145.

Welch, N. & Scott, T. (2016) Introduction: Composition in the age of austerity. In N. Welch and T. Scott (Eds.), *Composition in the Age of Austerity*. (pp. 3-20).

White, E. M. (1994). Issues and problems in writing assessment. *Assessing Writing*, *1*(1), 11–27.

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

White, E. M., Lutz, W., & Kamusikiri, S. (Eds.). (1996). *Assessment of writing: Politics, policies, practices*. Modern Language Association of America.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483.

Yancey, K. B. & McElroy, S.J. (2017). Assembling composition: An introduction. In K.B. Yancey & S.J. McElroy (Eds.), *Assembling composition* (pp. 3-26)

Yu, H. (2012). Intercultural competence in technical communication: A working definition and review of assessment methods. *Technical Communication Quarterly*, *21*(2), 168–186.

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from *Assessing Writing* (2000–2018). *Assessing Writing*, *42*, 100421.

APPENDIX. EXAMPLE TRANSLATION BRIEF

Localisation in the classroom

Translation brief

To give your partner(s) in Europe a better overview of the purpose of the translation, please fill in the following form, and you send it to your translator(s). In the Cc: line, please include your instructor's address and that of Dr. Birthe Mousten bmo@asb.dk (when working with Danish students) OR Dr. Sonia Vandepitte sonia.vandepitte@hogent.be (when working with Belgian students) OR Dr. Federica Scarpa fscarpa@univ.trieste.it (when working with Italian students).

Fill in your name (or initials)	
Who is the intended reader of the translated text?	
Is this reader different from the one who you originally intended the text for, and if so, how?	
Who is the reader of the translated text going to see as the sender of the text?	
Is the message (this is what I want to tell you) of the text going to be exactly the same in a translated version. If not, what changes do you envision?	
What medium is the text intended for (magazine, brochure, instruction insert with a product, other things?)	
Is there anything in the 'code' (pictures, wording, level of formality, tone) of the text that you would like the translator/localizer to pay special attention to? If so, what?	
Are there references in the text (to materials, institutions, other things) that you see as problematic for non-American people? If so, what?	
In what situation do you think the reader will actually read this text?	
Questions to translator/localizer	
Other comments	