Spring 2018

Beyond motivation: Differences in score meaning between assessment conditions

Nikole Gregg

Beyond Motivation: Differences in Score Meaning between Assessment Conditions

Nikole Gregg

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Master of Arts

Department of Graduate Psychology

May 2018

FACULTY COMMITTEE:

Committee Chair: John Hathcoat

Committee Members/ Readers:

Allison Ames

Dena Pastor

Acknowledgements

I would first like to thank my advisor, John Hathcoat. You have brought thoughtful insights into my work. Your constructive feedback and encouragements have strengthened me as a writer, presenter, and researcher. Thank you for teaching me what it looks like to be a scholar.

I would also like to thank my two committee members, Allison Ames and Dena Pastor. You have both helped me understand the nuances of data analysis and research than I thought possible in such a short amount of time. Your wisdom, patience, and feedback throughout this process has been invaluable to my learning. Thank you.

An additional thank you to the students in CARS and Psychological Sciences program who have supported and encouraged me throughout this project. I am grateful particularly to my cohort: Shane, Chi, Tom, and Andrea. Thank you for the encouraging words, and the advice as I trudged through difficult pieces of this project. I also want to thank Madison, Liz, and Andrea, for not only helping me through this project, but life as well. I also want to thank Aaron, who made himself available for advice whenever I needed it, no matter how busy you may have been at the time. Thank you.

Of course, thank you to my friends who checked in on me when timelines were tight. Thank you Carlin and Hailey, for always supporting me through multiple stages of my life, this one being no different. Your patience and understanding throughout this project is evidence of how lucky I am to be your friend.

Finally, I must thank my family. Mom and Dad, thank you for showing me what hard work looks like. Thank you for being selfless, kind, and patient with me as I have

undergone the most challenging, but rewarding time of my life thus far. I could not have been here, doing what I do, without knowing you both have my back.

Table of Contents

List of Tables

List of Figures

Abstract

Written communication is a skill necessary for not only the success of undergraduate students, but for post-graduates in the workplace. Furthermore, according to employers the writing skills of post-graduates tend to be below expectations. Therefore, the assessment of such skills within higher education is in high demand. Written communication assessments tend to be administered in one of two conditions: 1) course embedded and 2) a low-stakes, non-embedded condition. The current study investigated possible construct-irrelevant variance in writing assessment scores by using data from a mid-sized public university in the Mid-Atlantic region of the United States. Specifically, 157 student products were scored using the Association of American Colleges and Universities' Written Communication rubric by Multi-State Collaborative trained raters. A final sample size of 57 student products were in the non-embedded assessment condition and 107 student products were in the embedded assessment condition. Differential item functioning analyses were conducted using a Rasch Rating Scale model and an Ordinal Regression wherein Verbal SAT was used an external criterion of ability. Said differently, this study investigated whether students of the same proficiency had different probabilities of receiving particular written communication scores. After controlling for motivation, the results provide evidence of possible differential item functioning for Content Development as well as Genre and Disciplinary Conventions. Students of the same ability tend to obtain higher written communication scores in the non-embedded assessment condition. These results raise concerns about the presence of construct-irrelevant variance aside from motivation. Future research should investigate faculty feedback, allotted time, and task structure as possible sources of construct-

irrelevant variance when using low-stakes, non-embedded assessments of written communication.

CHAPTER 1

Introduction

Written communication is a key skill in everyday life for both the employee and the student. The college student writes numerous essays, reports, and research papers during their academic career. After graduation, employers expect these students to communicate through written text for varying purposes and across different forms (e.g. emails, creation of websites, analytic reports, etc.). There is a high demand and consistent need to write coherently within higher education and post-graduation (Sparks, Song, Brantley, & Liu, 2014). In order to be successful in the workplace, it is essential students develop effective writing skills in college (Casner-Lotto & Barrington, 2006; CWPA, 2011). Therefore, it is no surprise that higher education institutions prioritize written communication skills across their curriculum (AAC&U, 2007; Markle, Brenneman, Jackson, Burrus, & Robbins, 2013).

Despite a relative consensus among stakeholders of higher education about the importance of written communication, there are numerous controversies about the best way to assess this skill. For example, there are discrepancies between written communication theory and the instruments used to assess this outcome (O'Neill & Murphy, 2012). Faculty who teach written communication also tend to believe that timed writing tests do not reflect the kind of writing that is valued within the classroom (Calfee & Miller, 2007; O'Neill & Murphy, 2012). Moreover, students who take such exams in low-stakes testing, on average, tend to perform lower than students who have consequences for poor performance (DeMars, 2000). This has led some to argue that course-embedded assessments, defined as an assessment implemented within a course or curriculum, may alleviate many of these concerns (Coates & Seifert, 2011).

The purpose of my study is to investigate this claim. More specifically, this study investigated whether students who are matched on ability have different probabilities of receiving a particular score on a written communication rubric across a course-embedded and a non-course-embedded, low stakes assessment. If students of the same estimated ability have a lower probability of obtaining a particular score in a low stakes context then it would support the position that writing may be better assessed by sampling "authentic" student work. However, a failure to find such differences would suggest that the type assessment used to evaluate student writing ability may not be as important as some researchers claim.

To introduce these issues, first I will address validity, the two validity threats, and their corresponding issues. This is followed by an examination of how the assessment context may influence the presence of such validity threats. Finally, I will introduce the concept of differential item functioning as a strategy for investigating validity threats across two assessment contexts (i.e. non-embedded assessment and embedded assessment).

Written Communication: Validity and Construct-Irrelevant Variance

According to Messick (1995) validity is, "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment" (p. 741). Validity is not a property of a test (Standards, 2014). In other words, an assessment itself is not valid or invalid. According to Cronbach (1971) the interpretation and meaning of the scores from an assessment should be valid. Yet the

meaning of scores may change depending on the assessment context. In other words, the meaning of written communication score may change across assessment contexts.

Construct-Irrelevant Variance

The investigation of whether score meanings hold, "across settings or contexts is a persistent and perennial empirical question" (Messick, 1995, p. 741). One possible reason assessment scores do not hold the same meaning across assessment situations is due to the introduction of construct-irrelevant variance. According to Messick (1995), an assessment contains construct-irrelevant variance if the scores of the assessment contain systematic variance not pertinent to the construct of interest. For example, consider a prompt asking students to analyze a historical battle in the Civil War as if their writing would be in an established academic journal. They are asked to use coherent, organized language, and appropriate sources. This prompt also asked the student to consider the audience of their writing, and the context of the time period. Intuitively, the prompt seems to get at certain elements important to writing such as audience awareness and organization of thought.

Yet the performance of the student is also dependent on their Civil War knowledge thus making it inappropriate in certain assessment contexts such as a college admission exam. The variability in student performance within the written communication assessment scores due to the differing levels of Civil War knowledge may be considered construct-irrelevant variance if it is ancillary knowledge. In other words, construct-irrelevant variance would be present when using this prompt if we wanted to assess general written communication skill or their level of skill irrespective of their knowledge of the Civil War.

Construct Underrepresentation

Construct-irrelevant variance is one of two main threats to validity. The second threat to validity is construct underrepresentation. According to Messick (1995), an assessment contains construct underrepresentation when, "the assessment is too narrow and fails to include important dimensions or facets of the construct" (p. 742). For example, according to written communication frameworks (AAC&U, 2009; Adelman et al., 2011; CWPA, 2011; Sparks et al., 2014), elements of written communication include: genre, forms, audience awareness, context, and purpose, the writing process, and various linguistic elements (e.g. syntax and grammar). Yet written communication assessments differ in the extent to which each of these elements is measured (AAC&U, 2009; ACT, 2015; ETS, 2010). Therefore, many assessments underrepresent written communication to some extent. Construct underrepresentation may be more problematic in one type of assessment context compared to another.

Assessment Context and Validity Threats

In terms of the assessment context, the current study investigates scores across a high-stakes assessment and low-stakes assessment. In particular, the stakes of a test are the personal consequences associated with the examinee's performance. For example, high-stakes assessment situations refer to when the students taking the assessment have some personal consequence for their performance such as a grade on an assignment (Barry, Horst, Finney, Brown, & Kopp, 2010). In low-stakes assessment situations there are little to no personal consequences associated with an examinee's performance on a test.

In the following sections I review problems with low-stakes assessment such as lower motivation influencing average assessment scores, and the misalignment between timed writing assessments and the writing process element of written communication. I relate these problems within the low-stakes assessment context to the two aforementioned threats to validity (i.e. construct-irrelevant variance and construct underrepresentation). Then, I introduce a possible solution to low-stakes written communication assessment in higher education. Finally, I provide information about a national initiative which uses course-embedded assessments. The assessment and data from this initiative were used in the current study.

Low-Stakes Assessment: Motivation and Other Considerations.

Researchers are concerned that students do not exert their best effort during low-stakes assessments because little consequences are associated with their performance (Banta, 2008). If this is the case, then it is likely that students' written communication assessment scores may reflect –at least in part-- differences in levels of effort rather than variation in writing ability (Barry et al., 2010). Yet the purpose of the written communication assessment is to evidence the writing ability of students.

Because examinee effort is lower in low-stakes context, average test scores tend to be lower in low-stakes contexts compared to high-stakes contexts. For example, multiple studies found lower student performance and motivation in low-stakes assessments versus high-stakes assessments (e.g. DeMars, 2000; Liu, Bridgeman, & Adler, 2012; Sundre & Kitsantas, 2004; Sungur, 2007; Wise & DeMars, 2005; Wolf & Smith, 1995; Wolf, Smith, & DiPaolo, 1996). Elaboration of a few of these studies is found below and in the subsequent sections on motivation.

Specifically, a study by Liu, Bridgeman, and Adler (2012) investigated student performance across three motivational conditions (i.e. control condition, personal condition, and institutional condition) which differed in their student examinee instructions and two assessment-types (i.e. constructed-response and multiple choice). The control condition was the low-stakes condition where the students' assessment scores were not shared with anyone but the research team. The personal condition stated the students' scores could be shared with their employers and faculty, and the institutional condition stated students' scores would be used for research purposes and averaged across all other students but would not be shared with the research team.

Specifically, Liu and colleagues (2012) found lower performance of students in their low-stakes condition (i.e. control condition) compared to high-stakes conditions (i.e. personal and institutional conditions). In addition to lower performance in the low-stakes condition, there was a greater difference in student performance in the constructed-response condition between low and high-stakes conditions compared to student performance in the multiple-choice condition (Sundre & Kitsantas, 2004). A previous study by DeMars (2000) found similar results where students under high-stakes situations performed better than in the low-stakes situation, but the differences between high-stakes and low-stakes performance was larger for constructed-response items in the assessment compared to the selected-response items. According to DeMars (2000), this may be due to the increased cognitive demand of constructed-response compared to selected-response. Therefore, scores from a constructed-response assessments that have a higher cognitive demand than selected-response assessments, may be more affected by low-stakes assessment conditions.

Strategies to Fix the Motivation Issue: Motivation Filtering. Many researchers agree that low motivation and effort are problematic for assessments implemented in a low-stakes context. Yet this assessment context is the only way for many higher education institutions to gain access to student time to assess institutional learning outcomes. Therefore, many researchers provide information on different strategies to increase the student perception of stakes in testing. Some common practices for increasing the stakes of tests for students include: having scores contribute to student course grades (Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996), providing extra monetary compensation for higher performance (Braun, Kirsch, & Yamamoto, 2011; Duckworth, Quinn, Lynam, Loeber, & Stouthamer-Loeber, 2011; O'Neil, Sugrue, & Baker, 1996; Taylor & White, 1981), and providing feedback after the test (Baumert & Demmrich, 2001; Wise, 2004).

Some of these strategies may be more useful than others. Specifically, Duckworth et al. (2011) evidenced an increase in test scores by an average of .64 standard deviations with monetary incentives. In addition, Wolf and Smith (1995) and Wolf, Smith, and DiPaulo (1996) found about a 1.5 standard deviation increase in student performance between a condition of no grade consequence and a condition with a grade consequence. Yet researchers such as Baumert and Demmrich (2001), Finney, Sundre, Swain, and Williams (2016), and Wise (2004) found providing feedback does not substantially increase student motivation or test.

In addition to these previous methods, some researchers use the strategy of motivation filtering to address some validity issues of low student motivation across low-stakes assessments. According to Wise, Wise, and Bhola (2006), "the logic underlying

motivation filtering is that the data from those giving low effort are untrustworthy, and by deleting this data, the remaining data will better represent the proficiency levels of the target group of students" (p. 66). There are two broad methods of motivation filtering. One method is by collecting information on student motivation during the assessment session using a self-report measure. For example, Sundre and Wise (2003) administered a scale termed the Student Opinion Scale (SOS) to students right after the completion of their assessment.

When analyzing the data Sundre and Wise (2003) first rank ordered motivation scores from low to high and then deleted the student performances with the lowest motivation scores. In this study by Sundre and Wise (2003) along with another study by Wise and DeMars (2005), found that the average student performance increased by at least one-quarter of a standard deviation after motivation filtering. Yet a limitation of this strategy is there needs to be accessibility to collect motivation information. Specifically, it may not be feasible to collect self-report motivation information during a testing administration from examinees. In addition, before this motivation filtering technique can be performed two criteria must be met. There must be a meaningful correlation between test performance and test motivation, and test motivation should be unrelated to ability (Wise, Wise, & Bhola, 2006).

These assumptions and the need to administer a motivation scale are limitations to this motivation filtering approach. In addition, it is unclear the extent to which examinees respond truthfully when asked about their effort and motivation of an assessment they have just taken (Wise & Kong, 2005). Self-report measures assume that students

seriously respond to the self-report motivation scale whether they did or did not take their assessment seriously.

An alternative, second method to motivation filtering is through the use of item response-time. According to Wise and Kong (2005), the method of response time effort (*RTE*) filtering, "is based on the hypothesis that when administered an item, unmotivated examinees will answer too quickly" (p. 163). Within this method, a procedure is necessary which differentiates rapid-guessing behavior from what is termed solution behavior on each item. In this procedure a threshold is established for the response time on each item which represents a boundary between a rapid-guessing response and a solution response (Wise, Pastor, & Kong, 2009). In other words, this procedure requires one to identify the minimal amount of time that is needed to meaningfully respond to an item. Once these thresholds are established, responses are classified as solution behavior (1) or rapid guesses (0), with the proportion of solution behaviors across items serving as RTE. As in motivation filtering with self-report scales, a cutoff is used with RTE such that examinees below this RTE cutoff are eliminated from the data.

Though RTE seems promising as a technique to identify rapid responders (Wise & Kong, 2005), it has one important limitation relevant to the current study. The RTE method works well with selected-response, dichotomous items, but is not applicable to constructed-response assessments or performance assessments. Specifically in performance assessment, RTE is not necessarily indicative of motivation (Steedle, 2014). In other words, it is difficult to find an appropriate threshold to differentiate solution behavior from unmotivated responding with performance assessments. Therefore, the concept of rapid-responding is ambiguous when using constructed-response formats.

Other Considerations: Timed-Writing. There are other issues within non-course-embedded assessments. One of these issues is timed-writing. For example, research conducted by Yancey, Fishman, Gresham, Neal, and Taylor (2005) indicated that college composition teachers expect complex, in-depth, and well developed writing – writing that is very different from the kind of writing students are able to produce in an abbreviated time frame. Furthermore, Calfee and Miller (2007) in their article on the *Best Practices in Writing Assessment* state that, "it is important to remember that writing is a performance task that requires substantial effort, motivation, persistence, strategic planning, and skill, as well as knowledge about your topic" (p. 268). In addition, according to O'Neill and Murphy (2012), the assessments that do not represent written communication as a process do not accurately and fully represent written communication as a construct. In other words, timed writing assessments underrepresent written communication.

Yet low-stakes assessment of writing may be more problematic than producing scores than alternative assessment strategies, such as course-embedded assessments. Specifically, timed low-stakes written communication assessment may not have scores indicative of the written communication elements it aims to measure. For example, Brown (2010) explains this problem by referencing the examination essay:

> Contrary to good writing practice, the examination essay is a first-draft piece of writing; it has not been read by a peer, no feedback has been given and no external tools for editing or proofing were allowed. The essay examination results in a first-draft expression, which probably does not represent fairly or accurately the full range of a student's writing ability or event thinking (p. 227).

Therefore, the scores may not be indicative of writing ability an assessment intends to assess.

Course-Embedded Assessments: Strengths and Weaknesses

Recall that lower student motivation issues relate to construct-irrelevant variance in assessment scores. A possible solution to this motivation issue in low-stakes assessment can be the use course-embedded assessments, where the evaluation of student ability and student learning occurs by sampling student products created within the classroom. In addition to possibly combating a source of construct-irrelevant variance, course-embedded assessments may also add some positive outcomes to the assessment process of student learning. For example, faculty are more likely to directly relate assessment scores to student learning (Banta & Blaich, 2010). Furthermore, embedded assessments increase faculty involvement in the assessment process, allow for possible faculty development, and create an alignment with coursework and curriculum. In the following sections I further explain these possible positive consequences of using course-embedded assessments. In addition, I contrast these probable positive consequences of course-embedded assessment with some challenges of course-embedded assessments. Then, I describe a national example of a course-embedded assessment initiative, along with some of its strengths and challenges.

Strengths of Embedded-Assessment. In general, the practice of embedded assessment may increase the stakes for students, and therefore may combat lower motivation often seen in low-stakes assessment. In other words, student assessment within a classroom often coincides with student consequences (e.g. a grade) and therefore there is a likelihood of increased motivation to exert more effort in performance in embedded-assessment compared to low-stakes assessments. In addition to the increased stakes for students, professionals within higher education argue embedded-assessments

are more ‘authentic’ compared to assessments implemented in high-stakes assessment situations (Rhodes, 2010).

According to Gulikers, Batiaens, and Kirschner (2004), “an authentic task is a problem task that confronts students with activities that are also carried out in professional practice” (p. 71). Furthermore, the authentic task requires students to use the skills necessary to perform the same task in a professional or outside-the-classroom setting (Van Merreienboer, 1997). For example, students are required to write post-graduation for various career related activities (e.g. report writing, constructing emails, etc.). Therefore, constructed-response, not selected-response written communication assessments are believed by many to be more authentic assessments.

The authenticity of an assessment task increases the likelihood of faculty involvement in the assessment of student learning (Banta & Blaich, 2010). Specifically, “good assessments have to have ‘face validity’ for faculty, who should be able to see how the information gathered during assessment will help them in the classroom” (Banta & Blaich, 2010, p. 24). In other words, the increased authenticity of an embedded assessment task provides faculty an opportunity to connect assessment results to their curriculum.

This connection between assessment results and faculty teaching is particularly difficult with low-stakes assessments which are likely to be implemented outside the classroom in an increasingly standardized setting. For instance, in many instances of non-course-embedded assessments, faculty do not know how to relate the results to their classroom, and consequently do not know how to improve the student learning experience from the assessment results (Banta & Blaich, 2010; Blaich & Wise, 2011).

This is particularly concerning since assessment results are key to improving the student learning experience (Fulcher, Good, Coleman, & Smith, 2014).

In addition, according to a survey conducted by the National Institute on Learning Outcomes Assessment (NILOA), two-thirds of chief academic officers at higher education institutions state faculty engagement is a key element necessary for the advancement of assessment practice at their institution (Kuh, Jankowski, Ikenberry, & Kinzie, 2014). The involvement of faculty in the assessment process increases the likelihood that the assessment will be used in a formative manner. For example, if faculty can relate results of an embedded written communication assessment to their particular curriculum, the results are more likely to be used to make some sort of change to the curriculum to better student learning in the future.

In summation, course-embedded assessments may combat some of the motivation challenges of low-stakes assessment. In addition, another positive consequence of course-embedded assessment is the perceived authenticity of the assessment task, which increases the likelihood that the results are meaningful to faculty. Consequently, faculty may be more likely to use assessment results to evidence possible student learning improvement. Lastly, such assessments are better aligned with the kind of writing that is valued by written communication experts than the timed writing tasks that are typical of non-embedded assessments (Calfee & Miller, 2007; CWPA, 2011; O'Neill & Murphy, 2012).

Challenges of Embedded Assessment. Though embedded assessments may have some strengths, they also have particular challenges. For example, in order for the embedded assessments to be aligned with the curriculum at a particular institution, it is

often necessary for the assessments to be created by the faculty themselves. Therefore, in addition to other duties, faculty must also make time to create an embedded assessment that can be assessed using a common rubric. This may be particularly difficult if such alignment is perceived to limit faculty autonomy, though many have indicated that such situations have resulted in faculty development opportunities (Banta & Blaich, 2010).

A second challenge of embedded assessments is their lack of standardization. According to Hathcoat, Penn, Barnes, and Comer (2016), standardization is, "simply defined as administering the same, or theoretically interchangeable, tasks to a group of students" (p. 895). Institutions and other organizations involved with higher education typically use standardization as a way to provide some level of comparability across groups (e.g. institutions, classrooms, student demographics).

Regarding embedded assessments in written communication, *within* a classroom there usually is some level of standardization. For example, there is usually a common task for all students to complete as part of their coursework. Yet standardization does not always occur *between* classrooms sampled to examine institutional outcomes. For example, course assignments and tasks for written communication may be combined across a particular program or institution, and then rated on a common rubric. Each task may vary in difficulty, extent to which feedback was provided prior to scoring, and alignment with a common rubric. These course differences may contribute to construct-irrelevant variance in the assessment scores.

Overall, course-embedded assessments have both strengths and challenges. Course-embedded assessments are hard to implement across multiple sections and courses at an institution. Therefore, gathering meaningful results from a large enough

sample of students may be difficult. Furthermore, scoring course-embedded assessments may be resource intensive in addition to other faculty responsibilities. Yet faculty may be more likely to relate assessment results to their classroom when using an embedded-assessment process. In addition, these course-embedded assessments may combat low student motivation in low-stakes assessments, non-course-embedded assessments (Rhodes, 2010).

The Multi-State Collaborative: A National Example of Embedded Assessment

An example of a national model for embedded assessment of student learning outcomes assessment is the Multi-State Collaborative (MSC) to Advance Quality Student Learning Initiative. Specifically, the MSC uses the written communication, quantitative literacy, and critical thinking VALUE rubrics to assess student learning achievement from actual work students produce as a result of their formal instructional curriculum (AAC&U, 2017). The implementation of these rubrics within the MSC initiative is nation-wide with over 100 institutions submitting student work.

Once institutions submit student work, raters score the assignments using the AAC&U VALUE rubrics. These rubrics were created by AAC&U in 2009 with the intention they would be implemented as course-embedded assessment tools. In reference to the comparison of embedded assessments and low-stakes assessment, AAC&U (2007), states that, “using the work that students produce through assignments mitigates many of the issues of motivation” (p. 60). Therefore, the intended purpose of the VALUE rubrics is to assess student learning within the classroom and curriculum, where students are more likely to be motivated and therefore more likely to put forward effort in their

performance. In other words, the issue of low effort in low-stakes assessment is lessened by the MSC and AAC&U VALUE rubric approach.

The Current Study

Yet what if the VALUE Written Communication rubric was implemented in a low-stakes assessment context? Would the scores from the low-stakes context have a different meaning than writing scores from an embedded-assessment context? From a validity standpoint, it is problematic if scores from the same assessment have different meanings dependent on the assessment context (Messick, 1995). If this occurs, scores from one of the assessment contexts may have some level of construct-irrelevant variance. In other words, some property involving the assessment context not attributable to the construct of interest is influencing variability in student scores.

The current study assesses potential construct-irrelevant variance between two assessment situations (i.e. low-stakes and course-embedded assessments) through differential item functioning (DIF) analyses. According to Zumbo (1999) DIF occurs when, "examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure" (p. 12). In other words, DIF results when an individual in one group has the same ability level as an individual in another group, yet they have different probabilities of receiving a specific score on an item.

In the context of the Written Communication VALUE rubric, students of a particular ability in one testing situation (e.g. low-stakes assessment) may have a different probability of receiving a specific score in the other assessment situation (e.g. embedded assessment) after filtering for low motivation. If this occurs, the difficulty of

the Written Communication VALUE rubric is different for students of the same ability depending on their assessment situation. This occurrence suggests that something else may be contributing to the students' assessment score other than their writing ability. In addition, this would imply that the assessment scores from differing contexts should not be compared. Yet if there is no evidence of DIF, then the scores across assessment contexts may not be substantially different in meaning.

CHAPTER 2

Literature Review

Within the literature review I first describe the demand for evidence of student learning, and the implications of such learning to stakeholders (e.g. employers and policy-makers). In addition, I describe expected competencies of higher education graduates, with particular focus on the written communication competency. I define written communication, describe how it is assessed within higher education, and provide validity considerations for written communication assessment. Finally, I describe how the current study plans to assess the validity consideration of construct-irrelevant variance in a nationally implemented written communication rubric.

A Demand for Evidence of Student Learning

For decades pressure to provide evidence of student learning has not only come from policy-stakeholders such as the Department of Education, but also from employers and from within higher education itself. Policy-makers' interest in evidence of student learning was demonstrated in 1986 when Virginia legislature demanded all public colleges and universities to assess student learning (Miller, 2012). Two decades later, the federal government called for evidence of student learning through a report released by the Commission on the Future of Higher Education (U.S. Department of Education, 2006).

The then U.S. Secretary of Education Margaret Spellings, along with a committee of 19 members, released *A Test of Leadership: Charting the Future of U.S. Higher Education*. In what is known as the Spelling's Report, the Department of Education warned of the increase in international competition in higher education, the low rates of access to post-secondary education, and the lack of answers to questions pertaining to the

knowledge, skills, and abilities obtained by college graduates. Among other recommendations, this report called for postsecondary education institutions to measure and report meaningful student learning outcomes.

In addition to the demands for evidence of student learning from the federal government, employers also put pressure on higher education. In the modern economic world, employers are, "asking employees to take on more responsibility and to use a broader set of skills than in the past" (AAC&U, 2011). Along with this demand on employees, employers are also calling for assessments to evaluate whether graduates can apply the knowledge they learn in their post-secondary education to real-world challenges. This demand is founded upon the gap between what skills employers desire and the skills of graduates from higher education institutions (Markle, Brenneman, Jackson, Burrus, & Robbins, 2013). For example, reports from Deloitte & The Manufacturing Institute (2011) and ManpowerGroup (2012) reported difficulties finding sufficiently skilled workers appropriate for the 21st century workplace.

Another study by the Conference Board, Corporate Voices for Working Families, Partnership for 21st Century Skills, and the Society for Human Resource Management found that according to employers, there is a need for more applied skills such as oral and written communication, teamwork, and professionalism (Casner-Lotto & Barrington, 2006). More specifically, over 25% of employers perceived college graduates as deficient in leadership and written communication. Similar results came from Hart Research Associates' (2010), who sampled over 300 executives of varying organizational sectors and found that only 28% of respondents thought higher education did a good job of preparing graduates for the workplace.

Additional to the policymakers and dissatisfied employers, the third source of pressure for such evidence comes from within higher education itself. In 2006, the National Association of State Universities and Land-Grant Colleges and the American Association of State Colleges and Universities created the Voluntary System of Accountability Program (VSA) for Public Universities and Colleges (McPherson & Schulenberg, 2006). The VSA gathers evidence across three student learning outcomes: critical thinking, analytical reasoning, and written communication in an effort to not only increase transparency but also to gather similar information across institutions.

Since 2006, the VSA created new methods for gathering evidence across student learning outcomes. Initially the VSA limited the instruments institutions used to report student learning to three nationally normed measures: the ETS Proficiency Profile, Collegiate Learning Assessment, and the Collegiate Assessment of Academic Proficiency (Liu, 2011). In 2015, The VSA Board adopted the National Institute for Learning Outcomes Assessment (NILOA) Transparency Framework as the reporting method of student learning outcomes assessment across higher education institutions. This new framework does not restrict student learning outcomes reporting to specific instruments. Currently, the Voluntary System of Accountability Program (VSA), along with the National Institute for Learning Outcomes Assessment (NILOA) and the Association of American Colleges & Universities (AAC&U) created a reward system to institutions using the Transparency Framework to intentionally integrate learning outcomes assessment across campus (VSA, 2017).

The Association for American Colleges & Universities (AAC&U) contributed their own response to the demand of student learning evidence independent of the VSA

response. In 2005, AAC&U created the Liberal Education and America's Promise (LEAP) initiative. This initiative took into account employer voices as it sought to, "recalibrate college learning to the needs of the new global economy" (AAC&U, 2007, p. vii). The VSA and AAC&U LEAP initiative are examples of nationwide movements within higher education to evidence student learning. Together higher education, employers, and policy-makers put pressure toward answering, "What are students learning?"

In addition, as college tuition substantially increases, various stakeholders want to know the knowledge, skills, and abilities of graduates. What are students learning? How is this attainment through the higher education experience going to increase their likelihood of being financially stable and successful? These are substantial questions given that the annual cost of attending a four-year institution in 1978 was $7,181 and in 2010 the annual cost was at $16,140 (College Board, 2010). Therefore, with the constant, substantially increasing costs of obtaining a post-secondary degree, "students, parents, and public policy makers seek to understand how public colleges and universities operate and whether they have done a satisfactory job preparing students for the challenges in the 21st century" (Liu, 2011).

What Important Competencies should Students Learn?

What skills should students learn in order to rise to the challenges of the 21st century? In other words, what skills and abilities are important for the success of a college graduate? Most employers, policy-makers, and other stakeholders within higher education describe 21st century skills as critical thinking, communication, and teamwork amongst others (AAC&U, 2011; Casner-Lotto & Barrington, 2006; Klein et al., 2009;

Markle et al., 2013; Sparks et al., 2014). More specifically, educational researchers have described a broad range of skills as being important to the success of 21st century college graduates. These skills have been primarily discussed by three organizations, all of which identified written communication as an essential learning outcome. These organizations include the Educational Testing Service (ETS), the Assessment of Higher Education Learning Outcomes (AHELO), and the Association for American Colleges & Universities (AAC&U).

The Educational Testing Service (Markle et al., 2013) compiled a review of higher education frameworks of student learning outcomes across 4-year higher education institutions. The purposes of the ETS project (2013) was:

1. To gather and review outcomes frameworks relevant to higher education, considering the social, educational, and occupational perspectives.
2. Determine the commonalities among these frameworks.
3. Identify assessments that Educational Testing Service (ETS) has developed in each of these domains and the extent to which the assessments align with the definitions presented here (p. 3).

The frameworks included were the Framework for Higher Education Qualifications (QAA-FHEQ), European Higher Education Area Competencies, Liberal Education and America's Promise (LEAP), Frameworks for Learning and Development Outcomes (CAS), the Degree Qualifications Profile (DQP), The Assessment & Teaching of 21st Century Skills (ATC21S), ETA Competency Model Clearinghouse's General Competency Model Framework (USDOL-ETA). These nine frameworks shared seven common domains of student learning important for graduates to be successful in the 21st century. These critical domains included, "creativity, critical thinking, teamwork, effective communication, digital & information literacy, citizenship, and various life

skills such as time management, goal setting, adaptation, and flexibility" (Markle et al., 2013, p. 13).

Another framework not included in the Markle et al. (2013) project is the student learning outcomes of the Assessment of Higher Education Learning Outcomes (AHELO). Within their feasibility study to develop international measures of learning outcomes in higher education, they stated, "Learning outcomes are indeed key to a meaningful education, and focusing on learning outcomes is essential to inform diagnosis and improve teaching processes and student learning" (Tremblay, Lalancette, & Roseveare, 2012, p. 9). In their framework for student learning on an international, global level, AHELO included written communication, problem-solving, analytical reasoning, and critical thinking as generic skills important for students to obtain before graduation (Tremblay, Lalancette, & Roseveare, 2012). These outcomes represent not only what is accepted as important for graduates of the United States, but of graduates internationally.

In addition to AHELO, the Association for American Colleges & Universities' (AAC&U) LEAP initiative acknowledged the importance of a global, international perspective on skills and abilities of graduates. In the College Learning for the New Global Century (2007) report, AAC&U's Leap Initiative stated that, "Today it is clear that the United States—and individual Americans—will be challenged to engage in unprecedented ways with the global community, collaboratively and competitively." (p. 15). AAC&U identified 16 learning outcomes believed to be necessary in order for students to contribute to such a global community. These 16 essential learning outcomes include: Integrative learning, global learning, foundations and skills for lifelong learning, ethical reasoning, intercultural knowledge and competence, civic engagement – local and

global, problem solving, teamwork, information literacy, quantitative literacy, reading, oral communication, written communication, creative thinking, critical thinking, and inquiry and analysis (AAC&U, 2007). Out of these 16 essential learning outcomes, 99% of academic upper administration staff across 433 educational institutions rated written communication as one of the most important intellectual skills for students to obtain when they graduate from an institution of higher education (AAC&U, 2011).

Amongst other skills and abilities, "written communication is considered one of the most critical competencies for academic and career success, as evident in surveys of stakeholders from higher education and the workforce" (Sparks et al., p. 1). More specifically, the Educational Testing Service (Markle et al., 2013) conducted interviews with provosts and vice presidents from more than 200 institutions, finding that written communication was the most frequently mentioned competency considered for academic and career success. In addition, employers have also reported that written communication as an important competency of graduates. For example, 93% of employers believe that written communication is an important skill of employees with a college degree (Casner-Lotto & Barrington, 2006).

Written Communication as an Essential Competency of Graduates

Students, higher education institutions, and employers agree written communication is an essential skill for graduates, yet there are discrepancies between the importance of writing as a skill and the actual writing performance of graduates (Sparks et al., 2014). These discrepancies provide evidence of the need for useful assessments of written communication across higher education. Useful assessments should be able to

inform programmatic, curricular, and instructional decisions needed in order to change the ability (e.g. improve the quality) of student writing.

Professional organizations such as Conference on College Composition and Communication (2009) and National Council of Teachers of English – National Writing Project (2010) assume the primary aim for written communication assessment is to improve teaching and learning. Assessment scholars acknowledge writing assessments' power, "to shape curricula, define values, influence pedagogy, and affect students' educational experiences and self-perception" (O'Neill & Murphy, 2012, p. 413). Yet according to Hillocks (2002), for such formative use of assessment results to be useful, the writing assessment must produce scores capable of indicating how well a student can actually write in various contexts.

Therefore, assessments must adequately represent written communication as a construct. If not, scores of the assessment may not adequately indicate writing ability. These scores may only represent a small portion of the full construct of written communication. In that case, the assessment is underrepresenting written communication as a construct and therefore the scores of the assessment do not represent a comprehensive view of written communication.

Scores of a written communication assessment may also be indicative of some other quality or ability of the student. For example, a writing assessment with a prompt about United States history may not produce scores solely indicative of written communication. Knowledge of U.S. History may also be represented in the scores. Whether this is an issue or not depends on the purpose of the assessment. If an individual

desires knowledge of writing ability of students and writing ability alone, this assessment is not appropriate.

These complications represent validity issues in written communication assessment. Elaboration of these validity issues pertaining to communication assessment are in the following sub-sections. Yet before addressing the question of how higher education assesses written communication, written communication must first be defined since definitions are crucial for understanding the validity issues surrounding assessment practices.

Frameworks for Delineating Written Communication

The following section parses out various definitions of written communication across higher education frameworks. Each of these frameworks include various written communication elements, which together create the written communication construct. Therefore after this section, the reader should understand the main frameworks contributing to written communication construct, and the specific elements which encompass written communication.

Frameworks for Written Communication

In an attempt to conceptualize the broadness of written communication Sparks et al. (2014) reviewed nine frameworks of written communication, all of which have slight differences in their definition of written communication. From these frameworks, Sparks and colleagues identified key elements of written communication commonly described throughout the nine frameworks. Only three of the nine frameworks in the Sparks et al. (2014) paper pertained to written communication within the United States Higher

Education system. Therefore, the following sections focus on three of the nine written communication frameworks.

Framework for Success in Postsecondary Writing: CWPA, NCTE, & NWP. Together the Council of Writing Program Administrators (CWPA), the National Council of Teachers of English (NCTE), and the National Writing Project (NWP) created the *Framework for Success in Postsecondary Writing* (2011). College faculty and high school writing teachers across the nation wrote and reviewed the framework, all of which agreed that the ability to write well is a basic skill necessary for success within and beyond college. The underlying premise of the document expresses that, "teaching writing and learning to write are central to education and to the development of a literate citizen" (p. 2). *The Framework for Success in Postsecondary Writing* also states that the development of good writing occurs through experiences and encounters of different contexts, tasks, audiences, and purposes. Furthermore, this framework emphasizes the need for students to compose written materials across a variety of texts (e.g. nonfiction, informational, imaginative, printed, visual, and spatial) to further understand such concepts as audience, purpose, context, and genre.

Degree Qualifications Profile. The Degree Qualifications Profile (DQP), created by the Lumina Foundation (Adelman et al., 2011), describe their communicative fluency expectations of students similarly to the writing expectations of the *Framework for Success in Postsecondary Writing* (2011). According to this framework, writing occurs across different audiences, and purposes, using multiple expressive modes and forms (e.g. digital strategies and platforms). The DQP has increasing expectations of student success across differing levels of education for the associate degree, bachelor's degree,

and master's degree. As students progress through higher education, so do the expectations for the student in terms of the level of cognition and tasks a student should be able to do or complete. For example, at the associate level, a student should, "develop and present cogent, coherent and substantially error-free writing for communication to general and specific audiences" (Adelman et al., 2011, p. 18). A higher level of writing is expected of at the bachelor's level where students should be able to, "construct sustained, coherent arguments, narratives or explications of issues, problems, or technical issues and processes, in writing and at least one other medium, to general and specific audiences" (p. 18). For an in-depth inquiry of the specific expectations of student communication ability across education levels, see the Degree Qualifications Profile report by the Lumina Foundation (Adelman et al., 2011).

AAC&U LEAP Initiative. According to the Association for American Colleges and Universities (2009) written communication, "involves learning to work in many genres and styles. It can involve working with many different writing technologies and mixing texts, data, and images" (para 2). More specifically, written communication according to the LEAP initiative encompasses five dimensions: context and purpose for writing, content development, genre and disciplinary conventions, sources and evidence, and control of syntax and mechanics (AAC&U, 2009). The AAC&U VALUE rubric for written communication includes descriptions of these dimensions, as well as behavioral anchors within the rubric itself as to what is expected of students across differing levels of ability for each element. The Written Communication VALUE rubric and supporting definitions of each dimension are located in Appendix A.

Components of Written Communication

Recall that the above frameworks include important elements of written communication. These elements include forms, genre, context, purpose, audience, language conventions, use of sources, and the writing process, and are compiled and mapped to the three aforementioned frameworks in Table 1 (AAC&U, 2009; Adelman et al., 2011; CWPA, 2011; O'Neill & Murphy, 2012; Sparks et al., 2014). Not all frameworks include all components, but most elements are mentioned across multiple frameworks. The exception to this statement is the writing process component of written communication. Other elements such as textual features (semantics, word usage, and syntax), genre, context, purpose, and audience awareness are more frequent across written communication frameworks. The following sections describe these individual elements in more detail in an attempt to disentangle their similarities and differences.

Forms. The skill of handling different forms of writing is one of the most common elements of written communication frameworks (Sparks et al., 2014). The LEAP initiative (AAC&U, 2009), DQP (Adelman et al., 2011), and *Framework for Success in Postsecondary Writing* (2011) also view forms as an important aspect of writing (See Table 1). The LEAP initiative describes forms of writing as "working with many different writing technologies, and mixing texts, data, and images" (Rhodes, 2010, p. 1); whereas the DQP (2011) similarly define forms as the use of "multiple expressive modes and formulations, including digital strategies and platforms" (Adelman, 2011, p. 18). *The Framework for the Success in Postsecondary Writing* (2011) includes the forms component within their 'Composing in Multiple Environments' element, referring to students', "ability to create writing using everything from traditional pen and paper to

electronic technologies (p. 10). In sum, forms of writing can be broadly characterized as the integration of different technologies, data, and images to support comprehension and the complexity of written material (Binkley, Erstad, Herman, Raizen, Ripley, & Rumble, 2010).

Genre, Context, Purpose, and Audience Awareness. Genre, context, purpose, and audience awareness are additional characteristics of written communication identified by the DQP (Adelman et al., 2011), LEAP initiative (AAC&U, 2009), and *Framework for Success in Postsecondary Writing* (2011). Due to the connectedness of these elements, genre, context, purpose, and audience awareness are collectively defined and explained within this section. For example, different genres relate to different purposes of writing, and different audiences relate to different purposes and contexts of a writing task. The following section expands on the interconnectedness of these terms while simultaneously defining these terms as independent elements of written communication. However, the definitions were not always provided for these terms by Written Communication frameworks and instead they relied on simple examples to illustrate each concept.

Genre is defined by AAC&U (2009) as the "formal and informal rules for particular kinds of texts and/or media that guide formatting, organization, and stylistic choices" (para 12). The DQP and the *Framework for Success in Postsecondary Writing* (2011) both describe genre through the use of examples. The *Framework for Success in Postsecondary Writing* (2011) includes nonfiction and imaginative writing as examples whereas and the Degree Qualification Profile additionally adds an action plan to the

genre category. In addition, examples from AAC&U (2009) include: lab reports, academic papers, poetry, webpages, or personal essays.

In many respects, each genre relates to different purposes of the writing task and context. For example, writing an opinion essay on the reasons for social disparity across ethnicities in the United States differs in purpose from a comparative essay about the different approaches to the treatment of colon cancer. Finally, each genre tends to be associated with different contexts of writing. For example, situations (i.e. contexts) that call for an opinion essay differ from situation that calls for an analytic report. In other words, writing should be appropriate for the purposes of the writing task (Sparks et al., 2014).

All three aforementioned frameworks include context and align it closely with purpose of writing. For example, the Association of American Colleges & Universities LEAP initiative define context and purpose of writing together stating:

> The context of writing is the situation surrounding a text: Who is reading it? Who is writing it? Under what circumstances will the text be shared or circulated? What social or political factors might affect how the text is composed or interpreted? The purpose of writing is the writer's intended effect on an audience. Writers might want to persuade or inform; they might want to report or summarize information; they might want to work through complexity or confusion; they might want to argue with other writers, or connect with other writers; they might to convey urgency or amuse, they might write for themselves or an assignment or to remember (AAC&U, 2009, para 9).

The other frameworks (i.e. DQP and the Framework for Success in Postsecondary Writing) do not include a set definition of context and purpose, but mention the need for students to experience writing across different contexts and purposes (Adelman, 2011; CWPA, 2011).

In general, genre, context, and purpose relate to the writing task, and more specifically what is being asked of the student (e.g. in the description of the assignment). In other words, the terms genre, context, and purpose reference differences in the writing task provided to students.

Differences in a writing task also determine the appropriate audience for the writer, the audience being defined as the intended reader or consumer of the written product (O'Neill & Murphy, 2012). For example, a research proposal for a conference presentation in biochemical engineering has a different audience than an imaginative fictional story of a woman working toward a CEO position in a majority-male business. These different written products (e.g. research proposal and fiction story) also differ in context (e.g. under what circumstances will the text be shared or circulated) and general purpose (e.g. to inform versus to entertain).

Though the terms genre, audience, purpose, and context relate and depend on one another, written communication theory and frameworks differentiate between these elements. Therefore, these elements are viewed independently, with an acknowledgement of their interconnectedness.

Language Conventions. Overall, language conventions generally refer to the grammar, spelling, word-choice, and syntactic conventions of language. Specifically, the DQP specifically expects students to produce fluent text that is "substantially error-free" (Adelman et al., 2011, p. 14). *The Framework for Success in Postsecondary Writing* (2011) includes language conventions such as knowledge of vocabulary and stylistic conventions in their framework. Lastly, the Association for American Colleges and Universities (AAC&U) LEAP initiative describes language conventions as disciplinary

conventions, including aspects such as formal and informal writing rules and use of active and passive voice (AAC&U, 2009). Recall that in general, language conventions also extends to various textual features such as syntactic use, grammar, and mechanics (O'Neill & Murphy, 2012; Sparks et al., 2014). Additional aspects of language conventions such as use of active or passive voice and other stylistic conventions tend to be framework specific.

Use of Sources. The use of sources is a dimension of the AAC&U LEAP Written Communication VALUE rubric (See Table 1). According to the LEAP initiative, sources are, "texts (written, oral, behavioral, visual, or other) that writers draw on as they work for a variety of purposes – to extend, argue with, develop, define, or shape their ideas" (AAC&U, 2009). High quality writing according to AAC&U incorporates sources of high quality (e.g. credible and relevant), used to develop appropriate ideas and prose, with appropriateness being dependent on the genre and purpose of writing.

The LEAP initiative framework (Rhodes, 2010) focuses on the relevance, quality, and credibility of sources. The DQP (Adelman et al., 2011) expands the use of sources from relevance and quality, to student use of non-English language sources at the bachelor's and master's level. Recall that both the LEAP and DQP initiatives direct attention to the types of sources incorporated in writing (e.g. quality, relevance, non-English language sources). In addition, *The Framework for Success in Postsecondary Writing* (2011) focuses on the application of sources more than the types of sources used in written communication products. In other words, according to this framework, sources should be appropriate in terms of relevance to the purpose and context of the written product or task. Therefore, when including *The Framework for Success in Postsecondary*

Writing (2011), the use of sources element of written communication not only includes the type of sources within a writing product, but also the relevance and appropriate application of the sources included in the written product.

In addition, the presence of sources as an element of written communication is dependent on the task itself. For example, what sources would one use when engaged in imaginative writing? In other words, there may be situations in which the writing task would not indicate a need for the use of sources. Therefore, the use of sources as a written communication element may depend on the writing task and purpose of writing. Yet despite this consideration, for the purpose of consensus across written communication frameworks, use of sources maintains itself as a key element to written communication.

Writing as a Process. *The Framework for Success in Postsecondary Writing (2011)* heavily emphasizes written communication as a process. Aspects of the writing process include, "invention, research, drafting, sharing with others, revising in response to reviews, and editing" (p. 8). In addition, many college composition teachers expect well-developed writing, only allowable by an extensive process of research, drafts, feedback, and revising (O'Neill & Murphy, 2012; Yancey, Fishman, Gresham, Neal, & Taylor, 2005). Yet many other frameworks, including the LEAP initiative and the Degree Proficiency Profile, do not include the writing process as a key component of written communication (See Table 1). Said differently, there is a discrepancy between writing as practiced in the classroom and the way in which theoretical frameworks have conceptualized written communication.

O'Neill and Murphy (2012) in their *Postsecondary Writing Assessment* chapter in the Handbook on Measurement, Assessment, and Evaluation in Higher Education highlight the disjunction between writing instruction and written communication theory. Writing instruction tends to involve and emphasize processes of writing (e.g. research, drafts, feedback, and revising), but most frameworks of written communication (e.g. Sparks et al., 2014) do not incorporate the writing process as a key component of written communication. Sparks and colleagues (2014) comment on this disjunction pertaining to the writing process as a component of written communication when stating, "these strategies and processes are a critical aspect of writing at the college level and, this, should be included in any comprehensive definition of written communication" (p. 8). In other words, many frameworks used to delineate, and hence operationalize written communication in higher education, have neglected process as a component of writing.

Validity Considerations: Higher Education Assessment

The Standards for Educational and Psychological Testing (2014) define validity as, "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test" (p. 14). Construct underrepresentation and construct-irrelevant variance are the two prominent threats to validity. Construct underrepresentation refers to, "the degree to which a test fails to capture important aspects of the construct" (Standards, 2014, p. 12). The second threat to validity, construct-irrelevant variance, refers to, "the degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (Standards, 2014, p. 12). Messick (1995) stated construct underrepresentation occurs when an assessment is too narrow and fails to include important aspects of the construct. Alternatively, construct-irrelevant variance occurs

when an assessment is too broad and contains excess, systematic variance associated with constructs, methods, or response-processes irrelevant to the interpreted construct.

Both construct underrepresentation and construct-irrelevant variance threaten written communication assessments. The following section reviews how task characteristics and assessment situations may exacerbate construct underrepresentation and construct-irrelevant variance when assessing written communication. In addition, the following sections include specific written communication assessments administered throughout higher education to illustrate, through the use of examples, the threats of validity in written communication assessment.

Construct Underrepresentation

The task structure of an assessment influences the level or amount of construct underrepresentation. In other words, the task structure or the task characteristics of an assessment gives evidence as to what a student should demonstrate and be able to accomplish. If the task structure or characteristics do not allow for a demonstration or ask for specific elements of written communication, the assessment underrepresents the written communication construct. Assessment scores therefore cannot be interpreted as fully representing all written communication components. In order to identify whether an assessment underrepresents the written communication construct, one must identify the written communication elements included in the particular assessment of interest. The following section describes assessments of written communication, their coverage of written communication components, and implications to underrepresentation of written communication elements across each assessment.

Assessment Coverage of Written Communication Components. These assessments are used nationally across many institutions. Each section below includes the purpose of the assessment, as well as the elements of written communication which are stated to be present and which are missing. Furthermore, most assessments described below include both a multiple-choice and constructed-response section of the test. Though both sections (i.e. multiple-choice and constructed-response) of an assessment share some elements of written communication, some elements of written communication may be stated to be present for one section and not the other section. See Table 2 to visualize how the written communication elements map to the following assessments.

Collegiate Assessment of Academic Proficiency (CAAP). According to ACT (2015), "CAAP tests are used by both two- and four-year postsecondary institutions to measure the academic progress of students and to help determine the educational development of individual students" (p. 1). The Collegiate Assessment of Academic Proficiency has five main uses:

1. Document achievement of selected general education objectives.
2. Indicate change from one educational level to another – "value added."
3. Compare local performance with that of other populations.
4. Establish requirements for eligibility to enter the junior year.
5. Establish other eligibility requirements (ACT, 2015, p. 2).

These uses generalize across five different domains: critical thinking, science, reading, writing skills, essay writing, and mathematics. More specifically, the writing assessment is in a selected-response format (i.e. multiple-choice) and a constructed-response format (i.e. essay).

Overall, both formats together comprehensively cover all of the written communication elements except for the writing process. Specifically, the selected-

response, multiple-choice format stated coverage of the genre, context, purpose, audience awareness, use of sources, forms, and language conventions. In contrast, the constructed-response stated coverage across less components, covering specifically the elements of audience awareness, use of sources, and language conventions (see Table 2).

Selected-Response/Multiple Choice. The selected-response, multiple choice assessment for written communication contains 72 items with a time allotment of 40 minutes. In addition, this assessment has two main components: use and mechanics and rhetorical skills. There is an overall score for this assessment, as well as subscale scores to represent proficiency in use and mechanics as well as rhetorical skills. Specifically, the use and mechanics subscale aligns with the language conventions written communication element. The rhetorical skills subscale aligns with the genre, context, purpose, audience awareness, use of sources, and forms written communication elements.

Constructed-Response/Essay. The constructed response essay contains two 20 minute writing tasks. The writing task of the constructed response portion of the CAAP assessment intends to elicit responses that include evidence of students':

1. Formulating an assertion about a given issue
2. Supporting that assertion with evidence appropriate to the issue, position taken, and a given audience
3. Organizing and connecting major ideas
4. Expressing those ideas in clear, effective language (ACT, 2015, p. 8).

Scores for the essay portion of the CAAP assessment range from 1 to 6 in increments of .25. Multiple raters score each essay using a holistic rubric, therefore there is only one overall score for performance on this portion of the assessment.

ETS Proficiency Profile. Similar to the Collegiate Assessment of Academic Proficiency (ACT, 2015), Educational Testing Service's Proficiency Profile assesses

writing in both selected-response (i.e. multiple-choice) and constructed-response (i.e. essay) formats. The Proficiency Profile, formerly known as the Measure of Academic Proficiency and Progress, evaluates four skill areas: reading, writing, mathematics, and critical thinking (Roohr, Liu, & Liu, 2017). Overall, the ETS Proficiency Profile writing assessment covers substantially less written communication elements than the CAAP assessment. When combined, both sections of the Proficiency Profile assessment only cover the use of sources and language conventions components (see Table 2). Specifically, the selected-response assessment covers the language conventions component, and the constructed-response section of the ETS assessment covers the language conventions component and the use of sources component of written communication.

Selected-Response/Multiple Choice. There are two versions of this test, an abbreviated form and a standard form. The abbreviated form is a 40-minute assessment, while the standard form is a two-hour assessment. In the standard form, the writing section (along with all the other sections) contains twenty-seven multiple-choice items. In the abbreviated form, there are nine multiple-choice questions in the writing section. The writing questions aim to measure students' ability to:

1. Recognize the most grammatically correct revision of a clause, sentence or group of sentences
2. Organize units of language for coherence and rhetorical effect
3. Recognize and reword figurative language
4. Organize elements of writing into larger units of meaning (ETS, 2010, p. 4).

For this particular section of the ETS Proficiency Profile, a student receives an overall score across the four skills (critical thinking, reading, writing, and mathematics), and a subscore for each of the skills (ETS, 2010). In other words, students receive an overall

writing score for the writing section of the ETS Proficiency Profile multiple-choice section.

Constructed-Response/Essay. The constructed-response assessment portion of the Proficiency Profile (ETS, 2010) evaluates, "how organized, clear, and effective a response is to the prompt as well as the quality of reasons and evidence provided for their position on the topic" (p. 8). The prompts present a claim about a specific topic, and examinees are asked to construct a clear and organized response that takes a position on the topic/issue. Respondents have 30 minutes to complete the task. More specifically, this assessment aims to measure students' ability to:

1. articulate complex ideas clearly and effectively
2. state a position on a claim and provide supporting evidence
3. support ideas with relevant reasons and examples
4. sustain a well-focused, coherent discussion control the elements of standard written English (ETS Proficiency Profile, para 1, 2017).

E-rater scoring engine scores Student performance on the essay section of the ETS Proficiency Profile. According to ETS, the "the e-rater® engine scores essays by extracting a set of features representing important aspects of writing quality from each essay" (ETS Proficiency Profile, para 1, 2017). Students receive a holistic score ranging from 0 to 6 based on this automated scoring procedure.

Collegiate Learning Assessment (CLA) Performance Task. Similar to the ETS Proficiency Profile constructed-response section, the CLA Performance tasks expects students to use of source and appropriate language conventions (Council for Aid to Education, 2017b). The reader should see Table 2 for the map of these written communication components to the CLA Performance Task.

Along with written communication, the Performance Task also assesses critical thinking skills. Specifically, the three elements of the CLA Performance Task are analysis and problem solving, writing effectiveness, and writing mechanics. The writing effectiveness and writing mechanics dimensions are described below:

1. Writing effectiveness: constructing organized and logically cohesive arguments. Strengthening the writer's position by providing elaboration on facts or ideas (e.g. explaining how evidence bears on the problem, providing examples, and emphasizing especially convincing evidence)
2. Writing mechanics: Demonstrating facility with the conventions of standard written English (agreement, tense, capitalization, punctuation, and spelling) and control of the English language, including syntax (sentence structure) and diction (word choice and usage). (Council for Aid to Education, 2017b)

There is a 60-minute time-limit for this assessment, and within those 60-minutes students respond to a real-world situation where they are asked to identify an issue from a real-world problem or conflict, and provide a solution to the problem (Council for Aid to Education, 2017a). Specifically, student responses are scored across the three dimensions of skills: analysis and problem solving, writing effectiveness, and writing mechanics (Zahner & James, 2015). Each subscore ranges from 1 to 6, where multiple trained scorers rate each students' performance task assessment using an analytic rubric.

AAC&U VALUE Rubric. The AAC&U VALUE (Valid Assessment of Learning in Undergraduate Education) Rubric for written communication is in alignment with the AAC&U framework on written communication. The AAC&U established rubrics for sixteen competencies found essential for graduates of the higher education system. One of these competencies is written communication and is defined as, "the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many writing technologies, and missing texts, data, and images. Written communication abilities develop through iterative

experiences across the curriculum" (AAC&U, 2009, para 2). The rubric has six criteria including: content development, context of and purpose for writing, disciplinary conventions, evidence, genre conventions, and sources (see Appendix A for definitions of these criteria).

According to the descriptions of the rubric's criteria along with the embedded language across scoring dimensions, elements such as genre, context, purpose, audience awareness, modes and forms, as well as language conventions are present with the use of the VALUE rubric for written communication (See Table 2). As with all other previously mentioned assessments, the VALUE rubric does not include processes in the written communication construct definition or its elements (Sparks et al., 2014).

It is important to note unlike the other constructed response assessments such as the Collegiate Learning Assessment (CLA), ETS Profile Proficiency, and Collegiate Assessment of Academic Proficiency (CAAP) assessments, there is not a set task or set of tasks students are required to respond to with the VALUE Written Communication rubric. The rubric is an assessment evaluating important elements of written communication across many institutions of differing assignments and tasks. Therefore, time-limits and task descriptions cannot be made for this assessment measure similar to the aforementioned assessments.

Specifically the Multi-State Collaborative (MSC) initiative, in collaboration with the Association of State higher Education Executive Officers (SHEEO) and the Association of American Colleges & Universities (AAC&U), collects student products (e.g. essays) assessing written communication (among two other competencies) across higher education institutions in thirteen states (MSC, 2017). The MSC does not specify a

specific assignment or assessment prompt for institutions to use when collecting student performance data on written communication. After the collection of student products, they are rated by trained raters using the AAC&U Written Communication rubric, where raters provide a score for each rubric element individually on a scale from 0 to 4. In addition, many student products are rated by more than one-rater to assess inter-rater reliability. Scores are then sent back to the institution, in hopes of the results being useful for curricular, programmatic, or pedagogical changes.

Consequences of Underrepresentation. Recall that construct underrepresentation refers to, "the degree to which a test fails to capture important aspects of the construct" (Standards, p. 12). When construct underrepresentation occurs, the interpretations of assessment scores should take into consideration missing construct components within the assessment. For example, the *Framework for Success in Postsecondary Writing* (2011) emphasized the writing process as an important component of written communication. Sparks and colleagues (2014) seconded this statement, re-emphasizing the importance of the writing process as part of the written communication construct. Yet, the assessments previously discussed do not encompass the writing process (see Table 2). Therefore, all of these assessments underrepresent the complete written communication construct, and the resulting interpretation of scores from these assessments cannot represent the writing process component of written communication.

In terms of particular assessments, the ETS Proficiency Profile assessment (2010) does not evaluate student performance for most of the written communication components. Both the selected-response section and the constructed-response section of

the Proficiency Profile assess only the language conventions component and the use of sources component. In addition, the CLA assessment measures only these two elements as well. Therefore, scores on these assessments cannot represent student's ability across all other written communication elements (i.e. the writing process, forms, genre, context, purpose, and audience awareness).

In summation, the Proficiency Profile assessments and the CLA assessment underrepresent the written communication construct. Therefore, interpretation of scores from the CLA and Proficiency Profile do not generalize beyond the language conventions and use of sources components. Yet the other two assessments (i.e. CAAP and the AAC&U VALUE rubric) cover more written communication components, and therefore scores on these assessments can represent an increasingly holistic view of student written communication ability.

In other words, scores from the Collegiate Assessment of Academic Proficiency (CAAP) and the AAC&U VALUE rubric better represent the broad construct of written communication in comparison to the Proficiency Profile and CLA assessments (see Table 2). Both the CAAP assessment and AAC&U VALUE rubrics represent all elements of written communication except for the writing process component. Therefore, scores from these assessments represent all elements of written communication other than student ability to perform the writing process.

Finally, a recurrent theme across assessments such as CAAP, CLA, and the ETS Proficiency Profile is their one-occasion, timed implementation. In other words, it is common for these assessments to be administered at one-time point. In addition, all three assessments include a timed-component. These features of the CAAP, CLA, and ETS

Proficiency Profile assessments contribute to the missing writing process component of written communication within these assessments. It is difficult to assess a writing process during a timed, one-occasion administration of an assessment. Therefore, in order to capture this particular component, an assessment evaluating a student's ability to follow through a writing process should occur over more than one time-point.

Construct-Irrelevant Variance

Construct-irrelevant variance refers to, "the degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (Standards, 2014, p. 12). Both construct underrepresentation and construct-irrelevant variance are threats to validity. In addition, recall that according to *The Standards for Educational and Psychological Testing* (2014) validity refers to the, "the degree to which evidence and theory support the interpretations of test scores for proposed uses of test" (p. 14). Construct-irrelevant variance limits the interpretations of assessment scores as indicators of solely written communication. Instead, the written communication assessment scores could indicate student ability or knowledge of other constructs, or represent a method effect. The following sections investigate two possible contributors to construct-irrelevant variance in written communication assessment scores: assessment task structure and stakes in testing.

Task Structure. Task structure relates to the aspects of the writing prompt or stem of a written communication multiple-choice question. Specifically, the following information pertains to writing prompts for constructed-response assessments. Along with specific characteristics of the writing prompt (e.g. topic, length of prompt, specificity of prompt, and linguistic level of prompt), task structure also refers to the time

allotted to write a response to a prompt. In general, these aspects of task structure can influence student performance. If this occurs, student assessment scores may be influenced by unwanted aspects of the task structure and not necessarily student writing ability. Therefore, differences in student performance across different writing task structures can indicate possible construct-irrelevant variance in the assessment scores.

For example, the topic of the prompt can influence student performance, especially across specific student groups. Lim (2010) found that different writing prompts can have substantial differences in their perceived difficulty. Specifically, the topic of the writing prompt significantly contributed to differences in perceived prompt difficulty for ESL students. This evidence supports that different prompts may influence students of the same writing ability to perform differentially dependent on their prior knowledge about particular topics. Therefore, this evidence supports the possibility that writing ability scores may depend on the topic of the prompt, and not reflect general student ability in written communication.

Furthermore, similar to Lim (2010), Cho, Rijmen, and Novak (2013) found differences in ESL student writing scores due to a difference in writing task difficulty across two variables: distinctness of ideas within the prompt and difficulty of ideas in the passage. In addition, Abedi and Lord (2001) evidenced a decrease in performance gap between ESL and non-ESL students after reducing the language complexity of their written communication multiple-choice tests. Therefore, task complexity as well as linguistic complexity within a writing task can influence student performance on a writing assessment independent of true writing ability. In other words, task complexity,

which is not a component of written communication (see Table 1) can affect student performance on written communication assessments.

Time also is an important factor to consider in terms of construct-irrelevant variance in assessment scores. According to O'Neill and Murphy (2012), "when time is a serious factor for most of the test population, or for particular groups within that population, a test's validity is diminished" (p. 411). For example, multiple studies indicate increased time for writing can increase student performance among ESL students (Cho, 2003; Polio, Fleck, & Ledger, 1998). This occurrence indicates that student performance depends on the time given to complete an assessment, which may not be indicative of student writing ability in other contexts. This phenomena of increased writing performance due to an increased time on task generalizes to other groups of students as well. Powers and Fowles (1996) found higher performance of students given a 60-minute time period on the GRE essay test, while students given a 40-minute time period on the GRE essay test performed significantly worse. Socio-economic status is also a factor for timed essay tests and performance. For example, Simmons (1992) found that students from the poorest schools were disadvantaged from a timed test.

In general, time of a task contributes to the performance of students. Therefore, time allotment for a written communication assessment may contribute to student performance, not necessarily written communication ability within other contexts. In other words, the same student could perform substantially different across different topics, across different prompts of differing linguistic complexity, and across different time allotments for the task. If this is the case, the assessment score is not only indicative

of written communication ability. The irrelevant variance within the assessment scores interferes with the allowable interpretation of student writing ability.

In addition, these issues may be present in both course-embedded and non-course-embedded assessments. However, an essential difference between these contexts pertains to an opportunity to learn. Presumably, course-embedded assessments tend to occur after students have an opportunity to learn the material, an assumption that may be less tenable in non- course-embedded contexts.

Stakes in Testing. The task structure describes properties of the assessment. Specifically, task structure influences both construct underrepresentation and construct-irrelevant variance. In contrast, stakes in testing does not reference a property of the assessment prompt or task. Instead, stakes in testing refers to the consequences associated with performance in the assessment or testing situation. For the purpose of this study, the stakes of testing references whether students receive consequences for their performance.

Specifically, assessments are low-stakes when students do not have personal consequences for their performance. In contrast, assessments are high-stakes when students have personal consequences for their performance (Barry & Finney, 2009; DeMars, 2000; Wise & DeMars, 2005). In general, students who do not have consequences for their performance tend to put less effort toward the assessment compared to students with consequences for their performance (Liu, Bridgeman, & Adler, 2012; Wolfe, Smith, & DiPaulo, 1996).

If students do not put try on an assessment then scores may (in part) reflect differences in effort levels rather than variation in writing ability (Barry et al., 2010). This is indicated by a tendency to find that students perform worse, on average, in low-

stakes testing situations than in high-stakes testing conditions (Burke, 1991; Liu, Bridgeman, & Adler, 2012; Taylor & White, 1981; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996). In addition, the constructed-response format of many written communication assessments exacerbates the issue of student effort in low-stakes contexts. Research indicates item response format may influence motivation and, thus, performance (DeMars, 2000). Sundre (1999) and DeMars (2000) found that students put forth less effort on constructed response items (e.g. essay) than selected response items (e.g. multiple choice) in a low-stakes testing environment.

Studies by Wolf and Smith (1995) and Wolf, Smith, and DiPaulo (1996) found similar motivation effects across testing formats (i.e. low- and high-stakes). Constructed responses take more effort and motivation than a testing format with 'pre-made' response options (Liu, Bridgeman, & Adler, 2012). Therefore, appropriate interpretations of results may be more problematic for constructed response assessments (e.g. essay) compared to selected response assessments (e.g. multiple choice) in low-stakes assessment contexts.

Course-embedded assessments are an option when considering how to combat the issue of lower motivation and consequently lower performance in low-stakes testing compared to high-stakes testing (Rhodes, 2012). Course-embedded assessments evaluate student ability on a particular outcome by sampling student work that is part of their educational curriculum. In such a course-embedded assessment, students have consequences for their performance (e.g., a grade in the class). Therefore, students may be more likely to be motivated to perform with increased effort in course-embedded assessments when compared to a low-stakes testing situation (Liu, Bridgeman, & Adler, 2012).

Implications of Validity Threats

As described above, there are numerous issues in written communication assessment. These issues broadly refer to construct underrepresentation and construct-irrelevant variance. According to *The Framework for Success in Postsecondary Education* (2011), DQP (Adelman et al., 2012), and the AAC&U LEAP initiative (2009), written communication is composed of several elements: forms, genre, context, purpose, audience awareness, language conventions, use of sources, and the writing process. However, none of the assessments reviewed represent each of these aspects of writing. Though these issues are important, the current study focuses on construct-irrelevant variance as a threat to validity. More specifically, this study addresses the effect of different assessment situations and the inferences of student writing ability across both situations.

Institutions of higher education tend to assess student learning across two contexts: 1) low-stakes and 2) course-embedded. Each of these may have particular advantages and disadvantages. However, research indicates that assessment scores tend to differ across each context (Liu, Bridgeman, & Adler, 2012; Sundre, 1999; Wolf, Smith, & DiPaulo, 1996). This leads to the question of whether students with the same latent writing ability have different probabilities of receiving particular scores on a writing assessment across each situation. In other words, are students with the same writing ability more likely to score higher in a particular assessment situation (e.g. low-stakes) compared to another assessment situation (e.g. embedded assessments)? If this is the case, aspects of the assessment context may be a source of construct-irrelevant variance.

Differential item functioning (DIF), can be used to investigate such issues. DIF occurs when, "items on a test or psychological measure are multidimensional, measuring constructs or abilities or dimensions in addition to the primary dimension the assessment tool was designed to measure and two or more groups differ in their underlying distributions on these additional abilities" (Walker, 2011, p. 365). In other words, DIF occurs when an individual in one group has the same ability level as an individual in another group, yet these individuals have differing probabilities of receiving a specific score on an item.

DIF can also be described as the occurrence of different probabilities of getting an item on a test correct across two groups, after conditioning on ability (Angoff, 1993). For example, students with the same estimated ability level may have differing probabilities of scoring a particular score on a written communication rubric across specific assessment contexts (e.g. low-stakes and high-stakes). This would indicate DIF and provide evidence of possible construct-irrelevant variance in the assessment scores.

The following sections contain information about the history, terminology, assessment, and types of DIF. Then I discuss IRT-related models used in the current study to identify DIF. First, I review IRT models in general, then I move on to explain the Rasch Model, a 1 PL IRT-related model used in the current study. I then describe this model for dichotomous and polytomous items.

Investigating Construct-Irrelevant Variance: Differential Item Functioning

Originally, measurement practitioners conceived of DIF as being due to some characteristic of the test item that may not be relevant to the underlying ability of interest. Researchers now argue the occurrence of DIF can go beyond item characteristics to

incorporate testing situations (Zumbo, 1999). In the case of written communication assessment, students of the same ability across both testing situations (i.e. low-stakes assessment and course-embedded assessments) may have differing probabilities of getting a certain score on a rubric. If this is the case, scores may represent written communication ability and unwanted sources of systematic variance. This may occur due to some factor pertaining to the assessment conditions, such as impromptu writing in a constrained amount of time.

Defining Terms

It is important to note that DIF described above is not the same as when low-stakes and course-embedded groups of students perform differently, on average, on a written communication assessment. This occurrence is known as *adverse impact*, and is expected across low-stakes and high-stakes testing situations (Barry & Finney, 2009; DeMars, 2000; Wise & DeMars, 2005). Item impact is similar to adverse impact, but instead pertains to the item level and not the overall assessment score. In other words, item impact occurs when, on average, there are differences in individuals' scores across two groups score on a particular item (Ackerman, 1992).

Therefore, item impact occurs when there are differences across groups on the performance of a specific item. Yet this does not necessarily occur when there is DIF. Item impact and adverse impact look at average differences without taking the underlying ability of individuals in each group into account. Item impact and adverse impact can be present when DIF is present, but neither are necessary nor sufficient for DIF to occur.

The purpose of DIF assessment is to identify *item bias* which occurs when, "examinees of one group are less likely to answer an item correctly (or endorse an item)

than examinees of another group because of some characteristic of the test item or testing situation that is not relevant to the test purpose" (Zumbo, 1999, p. 12). If item bias is present, there is an indication that construct-irrelevant variance contributes to DIF. In other words, the occurrence of one group of examinees having a higher probability of getting an item right compared to another group of examinees after controlling for ability, would indicate that the scores represent something other than the trait of interest. Yet statistical evidence of DIF is not sufficient to determine item bias (Penfield & Lam, 2000). In other words, statistical detection of DIF is necessary, but not sufficient for item bias.

Evidence of statistical DIF leads into an investigation of item bias. The practice of evaluating item bias typically uses content-related procedures to identify non-target constructs responsible for differences in item performance independent of ability (Penfield & Lam, 2000). Though statistical DIF is not sufficient to conclude item bias, the following information presented in the current study solely pertains to statistical DIF as a beginning step toward determining possible item bias in the future.

Assessing Differential Item Functioning

In terms of statistical DIF, researchers typically follow certain steps. In addition to steps, researchers also consider differing models and approaches to conduct a DIF analysis. Specifically, researchers must make decisions about which groups will be compared against one another, the particular matching variable used for the analysis, along with other decisions pertaining to choosing a particular DIF assessment approach and model. In the following subsections, information is provided on these steps typically conducted in a DIF analysis.

Selecting Groups. The first step when conducting a DIF analysis is to select the groups of interest. In other words, what groups are we comparing? According to Myers, Wolfe, Feltz and Penfield (2006), this decision should be supported by theory and/or previous research. In addition, another consideration when selecting groups is determining if the groups of interest have a practical sample size. Most researchers define their groups of interest as the focal and reference group. The focal group is considered "the group of primary interest (i.e. one against whom there is concern for bias), and term the reference group as the standard to which the focal group is compared" (Myers, Wolfe, Feltz, & Penfield, 2006, p. 221).

Selecting a Matching Variable. In addition to selecting groups for comparison, the researcher interested in DIF must select a criterion measure in which to match participants on ability (Myers, Wolfe, Feltz, & Penfield, 2006). Matching variables are either internal or external. An external matching variable ideally is a, "parallel measure of the same construct from a different instrument" (Myers, Wolfe, Feltz, & Penfield, 2006, p. 221). Yet, it is uncommon for researchers to use an external matching variable (Clauser & Mazor, 1998). Typically, DIF analyses use an internal criterion measure, such as the scores provided by the instrument of study. For example, the matching variable can be student scores on the written communication assessment of interest within the DIF study. Yet if the researcher is interested in using an external matching variable for a DIF study involving a written communication assessment, they may use Verbal SAT.

Whether the matching variable is internal or external, it needs to have sufficient reliability and validity evidence. For example, if a matching variable is unreliable then the, "DIF analysis may not be meaningful because responses are being compared

between participants who are unevenly matched" (Myers, Wolfe, Feltz, & Penfield, 2006, p. 221). In other words, if the scores of the matching variable are inconsistent, then it is not possible to make appropriate conclusions from the DIF analysis.

Different DIF models. There are multiple techniques used to assess statistical DIF which can be categorized into a 2 x 2 matrix (see Table 3). According to the Potenza and Dorans (1995) framework, these approaches are differentiated by whether the matching variable is observed versus latent and whether the test is parametric or non-parametric. Note two things from Table 3. One, the DIF models within the table are polytomous DIF techniques. The particular interest of the current study is polytomous DIF, and therefore only polytomous DIF models are shown. Second, the DIF models are also either parametric or non-parametric and have either observed or latent matching variables.

In particular, the matching variable can either be an observed score or a latent variable. The observed score method uses the raw scores as an estimate of ability, where in contrast, the latent variable method estimates ability for an unobservable variable (Potenza & Dorans, 1995). For example, a Partial Credit Rasch model is a latent variable approach. In contrast, a polytomous logistic regression, such as an ordinal regression, may be an observed variable approach.

Furthermore, DIF procedures are also distinguished by the relationship between the item score and the matching variable (i.e. parametric procedures) and whether the relationship between item score and matching variable is not required to take a specific form. According to Potenza and Dorans (1995), the parametric approaches to assess DIF,

"require the assumption that the model for describing the relationship between item performance and the matching variable is correctly specified" (p. 24).

Therefore, DIF detected by a parametric approach may be due to model misspecification. Specifically, if the parametric model is not specified correctly, DIF detected using these methods might be due to model specification, not true DIF. In contrast, non-parametric approaches are not as prone to this model misspecification error. For further information on different DIF detection methods, the reader should see to Potenza and Dorans (1995), Penfield and Lam (2000), and Penfield and Camilli (2007).

An Item Response Theory Framework for Conceptualizing DIF

Despite the many approaches and multiple frameworks available to assess DIF, all of them in some way involve testing the null hypothesis of no DIF being present. The null hypothesis can be depicted as:

$$f(Y \mid \theta, G = R) = f(Y \mid \theta, G = F) \qquad (1)$$

where θ references ability, *G* corresponds to the grouping variable, *R* corresponds to the reference group, *F* corresponds to the focal group, and *Y* corresponds to the item score. Specifically, the above equation states that the distribution of the item score conditional on ability in the reference group is the same as the distribution of the item score conditional on ability in the focal group. The reference and focal groups are common terms describing the two comparison groups in a DIF procedure. If the conditional probability of *Y* is not the same across reference and focal groups, then individuals with the same level of ability have differing probability distributions of *Y*. When this occurs, there is statistical DIF.

The following section explains common models for detecting DIF. The reader should assume these models pertain to dichotomously scored items. In other words, the subsequent sections pertain to items on an assessment when there is a distinct right and wrong response. This is followed by an overview of polytomous DIF.

Introduction to IRT Models. The most common latent variable approach in detecting DIF is through the Item Response Theory (IRT) framework. Within an IRT framework, a variety of models are available to identify DIF (Penfield & Camilli, 2007). Within these IRT models, statistical DIF is assessed by comparing Item Characteristic Curves (ICCs) across groups. According to Zumbo (1999), "if the ICCs are identical for each group, or very close to identical, it can be said that the item does not display DIF. If, however, the ICCs are significantly different from one another across groups, then the item is said to show DIF" (p. 19). Specifically, an ICC depicts the probability curve of getting an item correct across a latent trait ability dimension (Walker, 2011).

The x-axis of the ICC represents ability or proficiency level and the y-axis represents the probability of answering an item correct (or endorsing) an item. Each Item Characteristic Curve is dependent on the different parameters of the corresponding IRT model. There are three parameters within the IRT framework: difficulty (*b*), discrimination (*a*) and pseudo-guessing (*c*). Specifically, the 1 PL model (see Figure 1) estimates the *b* parameter of each item, and the 2 PL model (see Figure 2) estimates both the *b* and *a* parameter for each item. The 3 PL model (see Figure 3) also estimates the *a*, *b*, and *c* parameters (de Ayala, 2009).

Each of these parameters gives specific information regarding a particular item. The *b* parameter indicates the amount of ability needed in order to be have a .5 or greater

probability of getting an item correct. The *a* parameter pertains to the slope of the probability curve. A steeper slope indicates a more discriminating item. In other words, the item can better distinguish between examinees in terms of their ability levels. The *c* parameter takes into account student guessing within an assessment and is evidenced by the lower asymptote of the ICC being greater than 0. Said differently, if the *c* parameter is estimated in an IRT model, then the lower asymptote of the ICC will be something other than 0. If the *c* parameter is not included in a particular IRT model, than the lower asymptote will be 0.

IRT models and DIF. Along with different IRT models used to detect DIF, there are also different types of statistical DIF to consider. In the following section, each type of DIF is described with an example and a corresponding figure for visualization. The following types of DIF include: uniform, crossing non-uniform, and unidirectional non-uniform DIF (Mellenbergh, 1982).

Uniform DIF occurs when, "the difference between the item characteristic curves (ICCs) for each group remains constant across all levels of ability" (Walker, 2011, p. 367). Recall that the ICC represents $P(X = 1 \mid \theta)$ where the probability of getting the item correct is on the y-axis as a function of ability on the x-axis, across any given set of parameters (Walker, 2011). Uniform DIF seen in Figure 4 occurs when an item is consistently more difficult for one group across all levels of ability in comparison to another group. Uniform DIF is the only DIF possible for detection in 1 PL IRT models (see Figure 4). Specifically, consider an average ability estimate of 0 in Figure 4. Males with an ability estimate of 0 have a .90 probability of getting the item correct whereas the

probability is about .20 for females of the same ability estimate of 0. This implies that for examinees of the same ability level, the item is more difficult for females than males.

Other models such as the 2 PL and 3 PL IRT models can detect non-uniform DIF, where there are differences in discrimination across groups. In addition, non-uniform DIF can be either non-crossing (Figure 5) or crossing (Figure 6). Crossing DIF is also known as non-uniform disordinal DIF. This type of DIF occurs when, "an item is more difficult for one group of examinees at some levels of ability but easier for the same group of examinees at other levels of ability" (Walker, 2011, p. 368). For example, in Figure 6 males have a higher probability of getting the item correct at lower ability levels, but females have a higher probability of getting the item right at higher ability levels. Therefore, the item is easier for males compared to females at lower ability levels, but the item is easier for females at higher ability levels. In contrast, non-crossing DIF in Figure 5 indicates that the item is always easier for males than females across all ability levels. In other words, the item is constantly easier across all levels of ability for one group (e.g. males) compared to another group (e.g. females). Yet recall that the item with crossing DIF is easier for females at higher ability levels, but harder at lower ability levels. Therefore, in non-crossing DIF the item always favors one group over the other. In crossing-DIF the item is favorable for one group only at specific ability levels.

The Rasch Model and DIF

In general, the Rasch Model is conceptually analogous to a 1 PL IRT model (Wu & Adams, 2007). In other words, the Rasch model estimates the difficulty (*b*) parameter, but not the pseudo-guessing (*c*) or discrimination (*a*) parameters. Due to the the Rasch Model estimating the single b parameter, only uniform DIF is evidenced using the Rasch

approach. Specifically, this study pertains to the traditional 1 PL polytomous Rasch models such as the Rating Scale Model (Andrich, 1978) and the Partial Credit Model (Masters, 1982). Though the current study pertains to polytomous items, the dichotomous model is a foundation to the other two Rasch models presented here. Building upon the dichotomous model, the Rating Scale Model (Andrich, 1978) and the Partial Credit Model (Masters, 1982) accommodate polytomous items.

The Dichotomous Case. For dichotomous items there are only two possible outcomes: a correct or incorrect response. The Rasch (1960) equation for the dichotomous case is:

$$\ln\left[\frac{\mathrm{P}_{ni}}{1-\mathrm{P}_{ni}}\right] = \beta_n - \delta_i \qquad (2)$$

where

P_{ni} is the probability of person *n* with ability β succeeding on item *i* (i.e. a correct response)

$1\text{-}P_{ni}$ is the probability of person *n* with ability β not succeeding on item *i* (i.e. an incorrect response)

δ_i is the difficulty of item *i*

β_n is the ability of person *n.*

This equation represents the log odds of getting an item correct as a function of the item's difficulty and the person's ability.

The Polytomous Case. Early developments in IRT and Rasch focused on dichotomously scored items (Lord, 1980), but in the past 30 years there has been a growing application of IRT and Rasch to polytmous items (Penfield, 2014). Polytomous items refer to item-types that do not have a simple correct/incorrect response, but instead

have a range of response values. These items have been used in a variety of settings, such as “the scoring of rating tasks, the scoring of testlets or groups of dependent dichotomous items, innovative item types, multiple-choice items for which the distinction between all distractors are retained for scoring purposes, and rating scales used to measure a host of psychological and behavioral traits” (Penfield, 2014, p. 36). Rasch models accommodating these polytomous item-types are increasingly complex compared to the aforementioned dichotomous Rasch model. In particular, a key difference between dichotomous and polytomous Rasch models is the presence of multiple thresholds within the polytomous case.

Researchers such as Masters (1982), Muraki (1992), and Tutz (1990) refer to the step function as a fundamental piece of polytomous IRT models. For example, consider a polytomous item with four ordered score categories ($Y_i = 1, 2, 3, 4$). As a score increases, so does the level-of-correctness, where a score of 1 is ‘portion correct’, a score of 2 represents a performance that is ‘partially correct’, a score of three is ‘mostly correct’, and finally a score of 4 represents a ‘completely correct’ performance. According to Penfield (2014), “one can conceptualize the score an examinee receives as being determined by the success that she has had in transitioning, or stepping, to successfully higher score categories” (p. 39). In the case of the example item, there are three possible steps where “Step 1 reflects the transition from ‘no portion correct’ to ‘partially correct’, step 2 reflects the transition from ‘partially correct’ to ‘mostly correct’, and step 3 reflects the transition from ‘mostly correct’ to ‘completely correct’” (p. 39).

In general, most IRT models define step functions using one of four approaches: the adjacent categories approach, the continuation ratio approach, the cumulative

approach, or the nominal approach (Penfield, 2014). All of these models interpret step functions differently. Of particular interest to this study however, is the adjacent category Rasch model approach since it is applied in the Rating Scale Model (Andrich, 1978) and the Partial Credit Model (Masters, 1982). The adjacent-category approach interprets the step function as if the only score categories of interest are the two adjacent (e.g. $Y_i = 0$ and $Y_i = 1$; $Y_i = 1$ and $Y_i = 2$). For example, the first-step function in alignment with the above example represents the transition point from where person *n* has an equal probability of scoring $Y_i = 0$ and $Y_i = 1$. The second step function represents the transition point from where person *n* has an equal probability of scoring $Y_i = 1$ and $Y_i = 2$.

Rating Scale Model (RSM). In addition to its use with rubrics, this model is used with Likert questionnaires, which include response options in the form of ordered ratings such as: strongly disagree, disagree, agree, strongly agree (Wright & Mok, 2004). In particular, the equation for the rating-scale model builds off of the dichotomous model:

$$\ln\left[\frac{P_{nij}}{P_{ni(j-1)}}\right] = \beta_n - \delta_i - \tau_j \quad (3)$$

where

P_{nij} is the probability of person *n* with ability β receiving a score of *j* on element *i*

$P_{ni(j-1)}$ is the probability of person *n* with ability β receiving a score of j-1 on element *i*

δ_i is the difficulty of item *i*

β_n is the ability of person *n*

τ is the threshold, or the transition point from where person *n* has an equal probability of scoring in two adjacent categories

Specifically, the τ are Rasch-Andrich thresholds (Andrich, 1998; Bond & Fox, 2015; Eckes, 2009; Linacre, 2006). These thresholds can be illustrated by considering their application to a written communication rubric.

In terms of a rubric, there are elements and score categories. For example, the elements of the AAC&U VALUE written communication rubric are: context of and purpose for writing, content development, genre and disciplinary conventions, sources and evidence, and control of syntax and mechanics (AAC&U, 2009). The rating scale categories range from 1-4. For the RSM in particular, the distance between thresholds (i.e. Rasch-Andrich thresholds) across rating scale categories is consistent across all rubric elements. In other words, the distance between category thresholds are fixed across rubric elements (Engelhard & Wind, 2013). This property of the RSM is an overarching assumption of the RSM model. Due to this assumption of the RSM model, the number and type of response options must be the same across all items of an assessment. For example, if there is a 20 item assessment with some items on a 1-4 scale and other items on a 1-3 scale, then RSM is not appropriate for this entire assessment.

Partial Credit Model (PCM). In contrast, the Rasch Partial-Credit Model (PSM) allows distances between thresholds to differ across items (Masters, 1982). Therefore, the PSM model would be appropriate for the aforementioned 20 item assessment where different items have different response scales. In other words, the distance between thresholds across rating scale categories are allowed to differ across individual rubric elements (Engelhard & Wind, 2013). This property of the PCM is evidenced by its corresponding equation:

$$\ln\left[\frac{P_{nij}}{P_{ni(j-1)}}\right] = \beta_n - \delta_i - \tau_{ij} \quad (4)$$

Where

P_{nij} is the probability of person *n* with ability β receiving a score of *j* on element *i*

$P_{ni(j-1)}$ is the probability of person *n* with ability β receiving a score of j-1 on element *i*

δ_i is the difficulty of item *i*

β_n is the ability of person *n*

τ_{ij} is the threshold, or the transition point from where person *n* has a an equal probability of scoring in two adjacent categories

Notice the threshold term τ_{ij} includes an *i*, indicating the thresholds can differ across items or for this particular study, across elements of rubric.

In particular, both the RSM and the PCM receive widespread use in part due to their application with a smaller sample that otherwise would not be sufficient for other, more complicated polytomous models (De Ayala, 2009). Ideally in terms of parsimony, especially due to the small sample size, the simpler model (RSM) would be the best choice for the current study's DIF analysis.

A Conceptual Overview of Polytomous DIF. Recall that in general, DIF analyses test the null hypothesis that no DIF is present. This generalizes to the case for polytomous DIF, but the presence of DIF is conceptualized a bit differently. For example, dichotomous DIF pertains to whether students of the same ability across differing groups have different probabilities of getting an item *correct*. Yet polytomous items do not have a dichotomy of correct and incorrect. In other words, there is usually a range of values indicating *level of correctness* for polytomous items (Penfield & Camilli, 2007).

According to Penfield, Gattamorta, and Childs (2009), “tests of polytomous items address whether individuals having the same level of proficiency, but belonging to different groups, have the same chance of obtaining each score level of the polytomous response variable” (p. 39). Yet these omnibus polytomous DIF measures compare overall item difficulty estimates between groups. Therefore, many tests of DIF are omnibus since they do not assess how the score level categories in particular contribute to the DIF effect. In other words, omnibus DIF analyses consider an item-level difference across groups, but do not consider how each threshold contributes to DIF.

A differential step functioning (DSF) method can identify how score categories contribute to the overall DIF effect. In general, the DSF framework compares the difficulty of the thresholds across the focal and reference group. Furthermore, the DSF analysis provides the researcher with some advantages over the omnibus DIF analysis (Penfield, 2007). First, the item-level DIF effect has lower power when the DSF effect varies across steps of a polytomous item (Penfield & Algina, 2003). Second, the differences in magnitude and sign of DSF effects within a polytomous item can contribute to a non-detected omnibus DIF effect (Penfield, 2007). Third, patterns of DSF within an item can also aide in identifying the cause of DIF (Penfield, Gattamorta, & Childs, 2009).

Due to the different information provided by DSF and its mentioned benefits, it is recommended that both DIF and DSF be computed when analyzing DIF for polytomous items (Penfield, 2014). Yet there are some disadvantages to DSF and Rasch methods for DIF. In particular, when there are sparse cells within score categories, it is difficult to estimate DSF. In addition, the Rasch model uses the sum score as an estimate of ability.

In other words, individuals with the same sum score receive the same ability estimate. Yet if there is a DIF item in a set of small items in an assessment, using the sum score for the ability estimate may influence non-DIF items to evidence statistical DIF (Magis & Facon, 2013). One way to combat this issue is to use a DIF method that employs an external matching variable such as a logistic or ordinal regression.

The Current Study

The present study proposes the use of the RSM and an ordinal regression model to investigate polytomous DIF across two assessment situations. These assessment situations include a low-stakes assessment situation and an embedded assessment condition. The instrument of interest is the AAC&U VALUE Written Communication rubric. Written products from both assessment situations are scored using this VALUE written communication rubric.

The rubric has five elements: 'Content Development', 'Genre and Disciplinary Conventions', 'Context of and Purpose for Writing', 'Use of Sources and Evidence', and 'Control of Syntax and Mechanics' (See Appendix A). These rubric elements are synonomous to polytomous items in the DIF analyses. Yet because there is a small number of items, there is a higher type I error rate in evidencing DIF items if there is already a true DIF item for those DIF analyses using an internal sum score matching variable (Lee & Geisinger, 2016). Therefore, the Rasch analysis is complemented by the ordinal regression analysis. In particular, the ordinal regression analysis uses Verbal SAT as a matching variable for ability.

In general, the research questions pertain to whether the rubric functions differentially across the assessment situations. This is of particular interest due to

previous research indicating differential student performance across stakes in testing (Burke, 1991; Liu, Bridgeman, & Adler, 2012; Taylor & White, 1981; Wolf & Smith, 1995; Wolf, Smith, & DiPaulo, 1996). In addition, construct-irrelevant variance due to differential effort across stakes in testing threatens the validity of scores (Barry & Finney, 2009). In other words, if construct-irrelevant variance is present, then the scores from low-stakes assessment conditions may have a different meaning than scores from the embedded-assessment condition (Messick, 1995).

For example, what if there is DIF detected for the 'Context and Purpose' rubric element? In other words, what if the 'Context and Purpose' element of the AAC&U VALUE Written Communication rubric is more difficult in the low-stakes assessment condition compared to the embedded assessment condition for examinees with the same overall writing ability? Then students in the low-stakes assessment condition would need to be of higher ability than students in the embedded assessment condition to be equally probable of receiving the same score on Context and Purpose rubric element.

This would indicate that scores across assessment conditions may have different meaning due to construct-irrelevant variance. Due to these differences in meaning, the writing scores across assessment conditions should not be compared to one another. In addition, if assessment scores do not differ in meaning across different assessment conditions, some of the benefits of course-embedded assessments may be called into question. Therefore, the current study examines possible validity evidence in support or against the presence of construct-irrelevant variance by conducting DIF analyses. These analyses will provide answers to the following research questions:

1. Are there overall differences in rubric element scores across the assessment conditions before controlling for ability?
2. After controlling for ability, do the rubric elements differ in difficulty across assessment conditions?
3. After controlling for ability, are there differences in the P(Y = j) between assessment conditions for each rubric element, where k represents each available score category?

CHAPTER 3

Method

The present chapter outlines the participants, data collection procedures, and the measures included in the current study. I then provide a description of the preliminary analyses, followed by Stage I and Stage II analyses. All of these pieces come together to answer the research questions in the Literature Review.

Participants

Study participants were undergraduate college students at a mid-sized public university in the Mid-Atlantic region of the United States. Student demographics include gender, race, credit hours, and SAT scores across both assessment conditions (low-stakes and high-stakes assessment). Specifically, out of 157 students, 84 are female (53.5%) and 117 are white (74.5%). Ten percent of students identified as being two or more races, and about five percent of students identified solely as either Asian or African American. Furthermore, the average GPA of the sample is 2.31 (SD = 1.36), the average Verbal SAT score is 575 (SD = 69), and the average credits earned is 58.01 (SD = 46.55).

Data Collection Procedures

Data were collected differently across two assessment conditions. Student performances within the course-embedded condition were collected through a data collection plan as a participant in the MSC initiative. Student performances within the low-stakes condition were collected during a university-wide ‘Assessment Day’. The following sections include information regarding the assignments contained in each assessment condition and the specific data collection procedures employed within each condition.

Non-Embedded Assessment Condition. Specifically, there are 80 student products in the non-embedded assessment condition. Student products were collected through a university-wide Assessment Day which takes place on two occasions. The pre-test occurs for first-year students a few days before the start of their first semester. The post-test occurs after students have completed 45 to 70 credit hours. The data include both pre-test (66%) and post-test data (34%). No student was included in both the pre-test and post-test data. The pre-test data are from Fall 2015 Assessment Day, and the post-test data are from the Spring 2017 Assessment Day. The non-embedded assessment data did not include student performances from any Assessment Day make-up sessions.

During Assessment Day students are randomly assigned to a two-hour testing session and room using their university identification number. Because different assessments are given during different sessions and in different rooms, this procedure allows each assessment to be completed by a random sample from the population of students at the institution. Efforts are made to ensure that the students take the same series of assessments during after completing 45-70 credit hours. All students are required to participate in Assessment Day. Students who are absent during Assessment Day must attend a make-up session. Failure to attend the make-up session results in a hold being placed on the student's account. Though all students are required to participate in Assessment Day, there are no negative consequences for poor performance. Therefore, this condition is a non-embedded, low-stakes assessment.

The writing assessment allows students to respond to a prompt within a 60 minute time limit (See Appendix B for the prompt). In particular, this assessment task evaluates student performance before and after students take their 'Skills for the 21st Century'

General Education courses such as critical thinking, human communication, and writing. A committee of faculty who teach these General Education courses and an assessment consultant developed this particular assessment. Specifically, this assessment asks students to write an opinion article that would hypothetically be published in the University student paper. Students are encouraged provided with sources and asked to consider writing elements such as audience, purpose, organization, and language conventions.

Embedded Assessment Condition. There are 173 student products collected across five different assignments in the embedded assessment condition. In the academic year 2016-2017, undergraduate faculty across the University were asked to contribute student work representing written communication skills. The following data collection steps were implemented across the University to collect course-embedded written communication assignments of student work:

1. An email was sent to program directors and assessment coordinators asking for faculty volunteers who are interested in participating in the MSC initiative.
2. Volunteers were asked to identify specific courses within their program from which artifacts may be sampled.
3. Student ID numbers and demographic data were obtained for each student enrolled the courses that are identified in step 2.
4. All courses identified by faculty volunteers were included in the final sample unless two or more courses are taught by the same faculty member.

Initially according to the MSC guidelines, only students with at least 90 credit hours were eligible. In addition, only 10 artifacts per course were to be included in the sample. Due to time constraints and limited number of faculty volunteers, all student performances that were collected were included in the study. I provide differences in credit hours between the embedded and non-embedded assessment condition in the results section.

The five assignments included in the embedded assessment condition are from varying disciplines: English, history, psychology, intelligence analysis, and philosophy. Only the assignment from the English capstone course was used for the assessment of writing ability. Other assignments asked for a written product, but also assessed other constructs of interest.

In addition to differing disciplines, these tasks also differ in the amount of structure given to the student. For example, the English assignment was given to capstone students. These students were not given much direction or detail about their task. In contrast, the history assignment was a book review for underclassmen (i.e. mostly freshman and sophomores), where there was more detail given to the student in terms of expectations of performance. Yet all of the assignments are constructed-response involving written communication. Furthermore, each assignment has a differing amount of student products, ranging from 11 to 43 student products per assignment.

Measures

AAC&U Written Communication VALUE Rubric. AAC&U created the Written Communication VALUE rubric (see Appendix A) in the same way as the other 15 AAC&U VALUE rubrics of essential learning outcomes (AAC&U, 2011). According to AAC&U (2007), the VALUE project began the rubric development process by collecting, “rubrics for all of the essential learning outcomes from campuses and other organizations that have developed them for their own local purposes” (p. 64). Then, faculty from universities across the United States synthesized common criteria and expectations for their particular essential learning outcome into a VALUE rubric.

In particular, the Written Communication VALUE rubric has five elements which include: context and purpose for writing, content development, genre and disciplinary conventions, sources and evidence, and control of syntax and mechanics (AAC&U, 2009). Each of the dimensions have scoring categories that range from 1-4. Student products can also receive a score of zero if there is no evidence of performance for that particular rubric element.

In terms of the scoring process, all raters must complete an online rater training. The rating process and training also includes adjudication and calibration to the rubrics. In addition, not all student products are rated by two raters. In terms of rater agreement, the most recent AAC&U (2017) report *On Solid Ground*, reported inter-rater agreement ranged from .60 to .84 across Written Communication rubric elements.

Student Opinion Scale. The Student Opinion Scale (SOS) is a 10-item self-report questionnaire administered to students at the end of their entire testing session on Assessment Day (See Appendix C). Each testing session includes more than one assessment. Students report both the effort they invested in their assessments, as well as their perceived importance of the tasks they completed (Sundre & Thelk, 2007). Specifically, the SOS has two subscales: Importance and Effort. Each subscale contains 5 items rated on a 1 to 5 scale, where 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree. Separate scores are calculated for each subscale, where the possible scores range from 5 to 25. However, this study will only use the effort subscale, which is commonly used for motivation filtering (Swerdzewski, Finney, & Harmes, 2011; Wise & Kong, 2005; Wise, Wise, & Bhola, 2006). Sample items include: “I engaged in good effort throughout these tests” and “While taking these tests, I was able to

persist to completion of the tasks." (Sundre & Thelk, 2007, p. 5). The SOS was only collected on students who participated in the non-embedded, low-stakes assessment (α = .81).

Demographic Variables. Recall that the student demographics include gender, credit hours, GPA, and Verbal SAT scores. In particular, GPA is on a range from a 0 to 4.0 scale. In addition, the SAT or the Scholastic Aptitude Test is intended to evaluate reading, writing and language, and math. Specifically, the area of interest is SAT Verbal, a standardized multiple choice test. Therefore the scores for the SAT for just the Verbal section range can range from 200 to 800.

Furthermore, the Verbal SAT scores indicate evidence-based reading and writing (College Board, 2018b). This is composed of a reading test and a writing and language test. According to the College Board (2018a), measures a range of reading skills which are grouped into three main categories: 1) command of evidence, 2)words in context, and 3) analysis in history/social studies and in science. Specifically, the reading test desires students to do such things as use evidence from a passage to come to a reasonable conclusion, to use context clues to figure out meanings of words, and use information to examine hypotheses.

The Writing and Language test asks students to write and edit mistakes and weaknesses in a text, and then fix them. In general, the Writing and Language test is composed of the three skills mentioned in the Reading test along with two more skills: expression of ideas and standard English conventions (College Board, 2018c). Specifically, students are asked to sharpen an argumentative claim, make a passage more precise, make editorial decisions to improve history, social studies, and history passages,

identify organizational problems within a text, and to evaluate building blocks of writing (e.g. sentence structure, word usage, verb tense, comma use).

Discussion of the alignment of this matching variable can be found in the Discussion chapter. Specifically, there may be some construct-underrepresentation and construct-irrelevant variance concerns with Verbal SAT pertinent to its use as a matching variable. In addition, further investigation of these demographic variables previously mentioned can be found within the Results section.

Data Analysis

The data analysis portion of this chapter is divided into three parts: preliminary analysis, Stage I data analysis, and Stage II data analysis. First, the preliminary analysis investigates data deletion due to the presence of zeros and motivation filtering. In addition, the preliminary analysis examines demographic comparisons and mean differences across non-embedded and course-embedded groups. Stage I analysis provides evidence supporting the following model assumptions: unidimensionality, and model fit. Stage II describes how the current study investigates DIF between the assessment conditions (low-stakes and course-embedded assessments). Specifically, Stage II uses a Rasch approach for investigating DIF. Finally, Stage III investigates DIF between assessment conditions by using Verbal SAT using an as an external criterion of ability.

The preliminary analyses and data management were performed using IBM SPSS version 24. The following analyses in Stage I and Stage II were performed using FACTOR (Lorenzo-Seva & Ferrando, 2006) and FACETS software (Linacre, 2017a). Stage III was performed using SAS 9.4 software.

Data Deletion. Before any analyses, including any preliminary analyses, we implemented particular data deletion procedures. For example, motivation filtering was used to delete student data indicating low student effort during the assessment. In addition, student performances rated with a zero were deleted from the data. Therefore, within the subsections below I first describe the reasoning and method behind data deletion due to the presence of zero scores. Then, I describe the data deletion due to the process motivation filtering.

Zeros. Recall the scoring categories for the Written Communication rubric elements range from 1 to 4. Yet the MSC allows for raters to report a zero. AAC&U's *On Solid Ground* (2017) states that, "A 'zero' score on any piece of student work is best described as reflective of an absence of evidence of student learning for that specific criterion" (p. 32). However, this absence of evidence represents two possibilities: (a) the student had low performance on the criterion, or (b) the student's assignment did not prompt the student to demonstrate their ability for that particular criterion. The first of these possibilities provides a score indicative of student ability (or lack thereof) on the particular rubric criterion (e.g. 'Sources and Evidence). The latter of these possibilities is not indicative of student ability. This is particularly problematic in a DIF analysis.

Recall that DIF analyses identify whether individuals in different groups of the same ability have differing probabilities of getting a particular score. Yet if a student's observed score is not indicative of their ability, then the estimated ability and the DIF analysis will be biased. This is because in a Rasch analysis, examinees with the same observed score get the same ability estimate (Frederiksen, Mislevy, & Bejar, 2012). Therefore, the current study is deleting the zeros from the analysis across all rubric

elements. One problem with this approach is that the initial sample sizes across assessment each condition would decrease.

In a previous study, the rubric element 'Sources and Evidence' had a large amount of zero scores (Hathcoat & Gregg, 2018). Therefore, in order to maintain the largest sample size possible for both assessment conditions, if one particular rubric element is causing a substantial amount of data deletion, then that particular rubric element was deleted from the analysis. In other words, instead of five rubric elements in the DIF analysis, fewer rubric elements were analyzed for DIF.

Motivation Filtering. After data deletion due to zero scores, motivation filtering was completed before anything was further analyzed from the data. One may argue that differences in motivation alone across assessment conditions may contribute to the evidence of DIF. In other words, differences in student motivation between assessment conditions may contribute to possible construct-irrelevant variance identified by DIF. Therefore, I performed motivation filtering using the SOS motivation data collected for the low-stakes condition.

Yet before motivation filtering can occur, two criteria must be met. First, there must be a correlation between test performance and test motivation. Second, there should not be a relationship between test motivation and ability (Wise, Wise, & Bhola, 2006). To check these criteria, a correlation between scores across particular rubric elements were correlated with the Effort SOS subscale score. In addition, a correlation was computed between both SOS subscale scores and Verbal SAT scores, where SAT scores were utilized as an estimate of ability. In particular, the correlational value considered to be "no relationship" between the SOS effort scale and Verbal SAT scores is any correlation

below positive or negative .3. Anything below this value indicates that there is no more than 9% of variance shared between SOS effort scores and Verbal SAT scores.

This coincides with motivation filtering practices conducted by Sundre and Wise (2003), Wise and DeMars (2005), and Wise, Wise, and Bhola (2006). For the data with SOS responses, the motivation filtering technique deleted the student performances associated with a SOS score less than 12 on the Effort subscale the Effort subscale. Recall the SOS scores for each subscale ranges from 5 to 25. A score of 12 symbolizes an estimate that on average, students tended to either respond neutrally or disagree with the SOS subscale items.

Preliminary Analyses. After the data deletion procedures, I ran preliminary analyses to investigate differences between the assessment conditions (i.e. non-embedded and course-embedded). First, I investigated demographic differences between the two assessment conditions, followed by an evaluation of raw score differences between the assessment conditions across all rubric elements.

Demographic Comparisons. Separate independent sample t-tests were conducted to evaluate statistical differences between SAT scores, GPA, and credit hours across the two assessment conditions. In addition, a chi-square analysis was used to determine if gender is independent of the students' assessment condition.

Mean Differences. There is an expected difference between student performance across assessment conditions (DeMars, 2000). Due to evidence of lower student performance in low-stakes situations, it is expected that the student performance for the low-stakes non-embedded assessment condition is lower, on average, than the average student performance for the embedded assessment condition. The non-parametric Mann-

Whitney U Test was used to investigate differences in student performance across assessment conditions for each of the five Written Communication elements.

Stage I: Assumptions and Model-Data Fit. The purpose of Stage I analyses is to investigate certain assumptions of the Rasch model. These assumptions include: specification of correct form (e.g. model fit) and unidimensionality (DeMars, 2010). In the following subsections I explain each of these assumptions and how I identify if the model adequately meets that particular assumption. The FACETS (Linacre, 2017a) program and SAS 9.4 software is used to evaluate these particular assumptions within Stage I.

Overall Model Fit. Both the overall fit and item-fit relate to the correct form assumption consideration of IRT-related models. According to Eckes (2009), empirical data will never fit the Rasch model perfectly, and therefore the real interest of overall fit pertains to the practical utility of a model. One way to assess overall fit is to examine unexpected responses (Fischer, 2007). Standardized residuals for individual persons can indicate the frequency of unexpected responses (Eckes, 2009; Linacre, 2008). A standardized residual can be computed first by computing the raw residual:

$$R_{nij} = Y_{nij} - E_{nij} \quad (7)$$

where

Y_{nij} = the observed score j for person *n* on item *i*

E_{nij} = the expected score j for person *n* on item *i*

And then computing a standardized residual by:

$$Z_{nij} = \frac{R_{nij}}{\sqrt{W_{nij}}} \quad (8)$$

where

R_{nij} = raw residual of person *n* on item *i* for score *j*

W_{nij} = model variance for person *n* on item *i* around its expected score *j* under the Rasch-model

The calculated residuals can then be used to evaluate overall model fit. Specifically, Linacre (2008) states satisfactory model fit is indicated when less than 5% or less of standardized residuals are ≥ 2 and less than 1% of standardized residuals are ≥ 3. Therefore, overall model fit was investigated by comparing the percentage of standardized residuals of individual respondents using the strategy proposed by Linacre (2008).

Item-Fit. After model comparisons and overall model fit was assessed, item fit was investigated. Two common fit statistics in Rasch analysis are Infit and Outfit statistics. For both infit and outfit statistics a mean-square (MS) is computed from standardized residuals (Linacre, 2012a). Then, the standardized residual is squared and averaged to compute an Outfit MS value where:

$$Outfit_{MS} = \sum_{n=1}^{N}\sum_{i=1}^{I} Z_{ni}^2 \Big/ N \cdot I \quad (9)$$

Note that the Outfit MS does not weight the variance of the residuals. In contrast, Infit MS are squared standardized residuals that are weighted by the variance of the residuals:

$$Infit_{MS} = \sum_{n=1}^{N}\sum_{i=1}^{I} W_{ni} Z_{ni}^2 \Big/ \sum_{n=1}^{N}\sum_{i=1}^{I} W_{ni} \quad (10)$$

Therefore, MS outfit statistics tend to be sensitive to outliers, whereas the MS Infit statistics are not as sensitive to outliers (DeMars, 2010). Infit and Outfit can give slightly different information to the researcher, and therefore it is recommended that both the Infit and Outfit statistics are considered in terms of item-fit.

MS Infit and Outfit are used as effect sizes for item misfit. For example, cutoffs for appropriate MS Infit and Outfit values differ depending on the proposed use of score and type of item (Linacre, 2012a). In general, items are flagged if the Infit/Outfit MS statistic goes outside the recommended range of .6 to 1.3 or 1.5 (Engelhard, 1992; Linacre, 2006; Wright & Linacre, 1994). In addition, the MS Infit and Outfit can be transformed into a t-distribution for statistical significance testing (Wright & Masters, 1982).

Unidimensionality. Recall that correct form is one of three assumptions of unidimensional IRT and unidimensional IRT-related models (DeMars, 2010). Unidimensionality is also an assumption underlying the Rasch model. In particular, unidimensionality is defined as, "a single latent trait being able to account for the performance on items forming a questionnaire" (Brentari & Golia, 2007, p. 253). In order to assess the dimensionality of the data, I performed a PCA of the Rasch standardized residuals. Specifically, this procedure was completed using the FACTOR program (Lorenzo-Seva & Ferrando, 2006)

In particular, a PCA of residuals investigates the hypothesis, "that the residuals are random noise by finding the component that explains the largest possible amount of variance in the residuals" (Linacre, 2017b, para 10). According to Linacre (2008), evaluation of unidimensionality can include multiple criteria. For example, if the

explained variance is more than 40% and less than 20% of the variance is unexplained by the first contrast of the residuals, then the unidimensionality assumption is seemingly met. In addition, Linacre (2008) states that if the eigenvalue is less than 1.4 for the first component, then unidimensionality is most likely met. Therefore, these criteria were used to determine the dimensionality of the current data from the Written Communication VALUE rubric.

Stage II: Differential Item Functioning Assessment with Rasch. Similar to Stage I, Stage II utilizes the FACETS software program (Linacre, 2017a). In particular, Stage II investigates the overall omnibus test of DIF across rubric elements (i.e. the rubric elements of this study). Recall that FACETS uses Joint Maximum Likelihood Estimation.

The specific Rating-Scale model of interest within this analysis is:

$$ln\frac{P_{nijk}}{P_{nijk-1}} = \theta_n - \delta_i - \alpha_j - \varphi_{ij} - \tau_k \quad (11)$$

where

P_{nij} is the probability of person *n* with ability β receiving a score of *j* on element *i*

$P_{ni(j-1)}$ is the probability of person *n* with ability β receiving a score of j-1 on element *i*

δ_i is the difficulty of rubric element *i*

θ_n is the ability of person *n*

α_j is the overall average ability of students in assessment condition *j*

φ_{ij} is the interaction term for the difficulty of item *i* and the average ability of students in assessment condition *j*

τ_k is the threshold, or the transition point from where person *n* has an equal probability of scoring in two adjacent categories

Within the FACETS (Linacre, 2017a) program, a bias/interaction analysis evaluates the statistical significance of the difference in the item difficulties between two assessment groups, after controlling for ability. The parameters in Equation 11 are estimated by a two-step calibration process (Linacre, 2012b; Myford & Wolfe, 2003).In the first calibration, all parameters except φ_{ij}were estimated. In the second calibration, the parameters from the first calibration were fixed and parameters for φ_{ij} were estimated (Eckes, 2009; Myford & Wolfe, 2003). In other words, "B" interaction is estimated after the initial calibration in the program (See Appendix D).

With this simultaneous estimation, the bias/interaction statistic in FACETS can then use the bias statistic below:

$$t_{ij} = \frac{\widehat{\varphi}_{ij}}{SE_{ij}} \tag{12}$$

In particular, the bias statistic approximates the distribution of a *t* statistic where the numerator is the estimated rubric element difficulties per each assessment condition estimated by the φ_{ij} facet in the model. In addition the denominator is the standard error of the φ_{ij}parameter estimate, and degrees of freedom for this statistic is number of observations – 1 (Eckes, 2009). Essentially, this statistic tests the differences of item difficulty estimates for each group over a standard error estimate.

Stage III: DIF Assessment with Ordinal Regression. Verbal SAT was used as the matching variable in an adjacent category logit ordinal regression. Specifically, one ordinal regression analysis is used for each rubric element individually. Furthermore, this model uses Verbal SAT scores and group membership (non-embedded and course-embedded) to predict the log odds of students scoring in specific score categories. In particular, this adjacent category model can be explained by:

$$\ln\left[\frac{P\,(Y=j_1)}{P\,(Y=j_2)}\right] = \tau_1 + \beta_1 Verb + \beta_2 Group \quad (13)$$

Where the log odds of obtaining a score of j_1 over a score of j_2 is a function of an intercept, verbal SAT, and group (embedded and non-embedded). Furthermore, τ is the intercept, where each regression equation for each rubric element models three different intercepts to represent the J-1 logit comparisons. For example, the τ_1 value is the intercept for the logit comparison between P(Y=1) and P(Y=2). Furthermore, τ_2 is the intercept for the logit comparison between P(Y=2) and P(Y=3.) Finally, τ_3 is the intercept for the logit comparison between P(Y=3) and P(Y=4). In order to evidence DIF, the slope for the group must be statistically significant.This would mean after controlling for ability (i.e. SAT Verbal), group is a significant predictor of the log odds for obtaining a particular score category. See Appendix E for syntax related to this analysis.

CHAPTER FOUR

Results

Recall that there are three pertinent research questions for the current study. These three research questions pertain to: (a) differences in raw rubric element scores between two assessment conditions, (b) differences between rubric element scores between assessment conditions, after controlling for ability, (c) differences in the $P(Y = k)$ between assessment conditions for each rubric element after controlling for ability, where k represents each available score category.

Data Management and Data Deletion

Zeros. Recall that though the rubric score categories range from 1 to 4, the reported scores range from 0 to 4. These zero scores can represent one of two meanings: (a) the student had low performance on the criterion, or (b) the student's assignment did not prompt the student to demonstrate their ability for that particular criterion. Therefore, all zeros are taken out of the data before any analyses.

Specifically, the rubric element 'Sources and Evidence' contained 75 zero scores (29.5%). Due to such a large amount of zero scores from one rubric element, the 'Sources and Evidence' rubric element was taken out of the data. This decision avoids deleting 75 scored products all together from an already small sample of 254 products. The 'Genre and Disciplinary Conventions' rubric element had 17 zeros and the "Content Development' element had two zeros. In sum, the 19 cases and one rubric element were deleted from the data set due to zero scores. After this step in data deletion, 127 cases remained in the embedded assessment condition, and 60 cases were in the non-embedded assessment condition.

Motivation Filtering. After data deletion due to zero scores, motivation filtering was completed before the preliminary analyses. Yet before motivation filtering occurred, I investigated two major criteria necessary for motivation filtering. First, there must be a correlation between test performance and test motivation. Therefore, we computed correlations between scores across each rubric element and SOS effort subscale sum scores. Note that only the 'Genre and Disciplinary Conventions' rubric element scores were correlated with the SOS Effort subscale scores to the degree of statistical significance (r = -.281, p <.05). In addition, the correlation is negative meaning that as test motivation increases, 'Genre and Disciplinary Convention' scores tend to decrease. This is the opposite of what is necessary to meet this particular criterion for motivation filtering.

The second assumption requires there not to be a relationship between test motivation and ability (Wise, Wise, & Bhola, 2006). To check this assumption, a correlation was computed between both SOS subscale scores and Verbal SAT scores, where SAT scores were utilized as an estimate of ability (See Table 4). In particular, the correlation between SAT Verbal scores and SOS Effort subscale scores was not significant (r = -.05, p > .05). Therefore, this second criteria for motivation filtering was met.

Despite not meeting one criteria for motivation filtering, only one SOS Effort subscale score was below the cut score of 12. There werealso nine missing scores for the SOS Effort subscale for the Assessment Day post-test from Spring 2017. These missing scores were treated as if the Effort subscale score was below the cut score. The 9 missing scores and the one SOS Effort subscale score below 12 were deleted from the data.

Therefore, after motivation filtering, 127 cases remain in the embedded assessment condition, and 50 cases remain in the non-embedded assessment condition. See Appendix F for score category frequencies between assessment conditions for all rubric elements.

Other Data Deletion Procedures. In addition to data deletion due to zeros and low motivation, more data deletion was necessary from missing information. In particular, 20 student cases did not have information regarding their student IDs. Therefore, I could not match Verbal SAT with these cases. Recall that Verbal SAT scores are the matching variable for the Ordinal Regression analysis in Stage III. Therefore these 20 cases, 19 from the English assignment and one from the History assignment were deleted from the data. In addition, more cases were deleted due to the lack of Verbal SAT scores, despite knowing their JMU student ID.

In summation, after the completion of all data deletion procedures, there are 107 student products for the course-embedded assessment condition and 50 student products for the non-course embedded assessment condition. Therefore, in total there are 157 scored student products included in the following procedures. See Table 5 for the number of scored products per each assessment condition.

Preliminary Analyses

In addition to data deletion, preliminary analyses investigate differences between the students in each assessment condition (i.e. course-embedded and non-course embedded). Specifically, comparisons investigate differences in particular demographic variables across assessment conditions. Due to the lack of random assignment of students to each assessment condition, of particular interest is whether students differ on specific demographic variables. The differences between assessment conditions in raw scores for

each rubric element is also of interest. In addition, the raw score differences for each rubric element across assessment conditions answers the first research question of interest: *Are there overall differences in rubric element scores across the assessment conditions before controlling for ability?*

Demographic Comparisons. The demographic variables pertinent here are gender, credit hours, and Verbal SAT. The information below contains the appropriate statistical test investigating group differences, along with effect sizes and confidence intervals when applicable.

First, there are 34 females and 16 males in the non-embedded assessment condition. The embedded assessment condition consists of 50 females and 57 males. Gender is not independent of assessment condition, $\chi^2(1)= 6.198$, $p = .013$. Specifically, there are more females than expected in non-embedded assessment condition, and less females than expected in the embedded assessment condition. Though the chi-square is significant, the phi coefficient ($\phi = .19$) indicates that about 3.6% of the variance is shared between assessment condition and gender. In addition, females are only 1.4 times more likely to be in the non-embedded assessment condition than in the embedded assessment condition. Therefore, though independence does not hold between these two variables, the effect size of the relationship is small.

In addition, on average the course-embedded assessment condition has more credits hours ($M = 72.80$, $SD = 45.18$) than the non-embedded course embedded condition ($M = 26.34$, $SD = 31.33$). This difference was statistically significant, where $t(132.61) = -7.468$, $p < .001$, 95% CI [-58.77, -34.16]. Furthermore, the average amount of credit hours for the course-embedded assessment condition is 1.12 standard deviations

above the average amount of credit hours for the non-course embedded assessment condition. Finally, there was no difference between assessment conditions and SAT verbal scores $t(155) = 0.376$, $p > .05$, 95% CI [-18.95, 27.85], where on average the non-course embedded assessment condition SAT scores ($M = 578.00$, $SD = 62.04$) are only .06 standard deviations higher than the average course-embedded assessment condition SAT scores ($M = 573.55$, $SD = 72.21$).

Mean Differences. It is expected that the student performance for the low-stakes non-embedded assessment condition is lower, on average, than the average student performance for the embedded assessment condition (DeMars, 2000). The non-parametric Mann-Whitney U Test indicated statistical differences in performance between assessment condition and performance across all four rubric elements (see Table 7). Specifically, student performance was higher in the course-embedded assessment condition than the non-course embedded condition for all rubric elements. For example, students in the non-embedded assessment condition ($M = 1.98$, $SD = .820$) scored significantly lower than the course-embedded assessment condition ($M = 2.42$, $SD = 1.00$) for the 'Context of and Purpose for Writing' rubric element, where $Z = -2.597$, $p = .009$. Therefore, to answer the first research question: yes, there are overall differences in rubric element scores across assessment conditions before controlling for ability.

Stage I: Assumptions and Model-Data Fit

Overall Model fit. Recall that percentage of extreme standardized residuals indicate the overall model fit. Specifically, satisfactory model fit is when less than 5% or less of standardized residuals are ≥ 2 and less than 1% of standardized residuals are ≥ 3 (Linacre, 2008). After running the Rating Scale Model in FACETS (Linacre, 2017a),

there are 628 observations, and therefore 628 residuals. The number of observations is computed by multiplying the number of student products (N = 157) and rating scale categories (n = 4).

Out of these 628 standardized residuals, 5 are ≥ 3 and 40 are ≥ 2. Therefore, 0.80% of the standardized residuals are ≥ 3 and 6.40% of the standardized residuals are ≥ 2. According to Linacre (2008), the current Rating Scale Model would be judged as satisfactory according to one criteria but unsatisfactory when using the 5% cut-off.

Finally, a second measure for overall model fit is the Rasch separation index. In general, this index describes the plausible amount of statistically distinguishable measurement strata among the Written Communication rubric. In this case, the Separation index is 5.76, indicating there may be about 5 or 6 distinguishable strata. In other words, this indicates about 6 groups of distinguishable sets of people based on their ability estimates.

The two additional global fit statistics are item reliability and a chi-square test of fit between rubric elements. First, rubric criteria reliability was .97, indicating adequate reproducibility of the relative measure location of the item parameter estimates. Furthermore, the rubric elements are not statistically the same χ^2 (3) = 1.343, p < .01.

Item-fit. Item fit can also be evaluated by individual item statistics such as Infit and Outfit statistics. Table 8 gives both Infit MS and Outfit MS. Recall that if the items' MS Infit or Outfit is below .6, then the item overfits the model. If the items' MS Infit or Outfit is above 1.3, then the item underfits the model (Wright & Linacre, 1994). Most of the rubric element infit and outfit MS indices are within the acceptable range specified by Wright and Linacre (1994). Specifically, the 'Context of and Purpose for Writing' rubric

element has an Infit MS value of .79 and an outfit value of .74. The 'Content Development' rubric element has an Infit MS value of .86 and an MS outfit value of .78. The 'Genre and Disciplinary Conventions' rubric element has an Infit MS value of 1.11 and a MS outfit value of 1.09.

Lastly, the 'Control of Syntax and Mechanics' rubric element has an MS outfit value out of acceptable range when using a 1.3 cut-off value. Specifically, this particular rubric element has a MS infit value of 1.14 and a MS outfit value of 1.50. Recall that the Outfit MS does not weight the variance of the residuals, but the Infit MS are squared standardized residuals that are weighted by the variance of the residuals. In other words, MS outfit indices are more sensitive to outliers than MS infit indices (DeMars, 2010). Therefore, the MS outfit indices for the 'Control of Syntax and Mechanics' rubric element indicates the scores may contain observations that are outliers.

Unidimensionality. In addition to model fit, unidimensionality is also an assumption underlying the Rasch model. Recall that unidimensionality is defined as, "a single latent trait being able to account for the performance on items forming a questionnaire" (Brentari & Golia, 2007, p. 253). In order to assess unidimensionality, a PCA was conducted on the standardized residuals using the FACTOR program (Lorenzo-Seva & Ferrando, 2006). In order to run the PCA in the FACTOR program, the residuals were first downloaded from the FACETS program (Linacre, 2017a).

Overall, the Rasch model explained 63% of the variance in the data. After taking into consideration the variability accounted for by the Rasch model, the first component had an eigenvalue of 1.43, which explained 35.6% of residual variance. The magnitude of this variance suggests that there may be another dimension (Linacre, 2008). Though it is

difficult to determine the extent to which there may be another dimension or whether this value is a reflection of the small number of criteria included within the analysis.

Stage II: Differential Item Functioning Analysis with Rasch

This section addresses the following research question: *After controlling for ability, do the rubric elements differ in difficulty across assessment conditions?* Yet before presenting results regarding this research question, I first describe Rasch estimates regarding rubric element difficulty, score category thresholds, and average ability estimates of students for both assessment conditions after collapsing across rubric elements. Then, I describe bias/interaction results. Recall these results indicate possible evidence of DIF in the rubric elements.

Rasch Model Estimates. The following information pertains to information from the Rasch Rating Scale estimates. Specifically, I provide descriptions regarding the difficulty estimates of the rubric elements, information on the score category thresholds, and estimates of ability level for each assessment condition. All of this information can be seen visually by a person-variable map (see Figure 7).

First, the measure column indicates the logit corresponding to particular ability and difficulty estimates. The student column represents student ability estimates, where each asterisk represents two people. The condition column represents the average ability of students in each assessment condition. Specifically, the higher the logit, the higher ability of the student and the more difficult the rubric element. Finally, the element column represents the rubric element difficulty and the scale represents the score category thresholds.

Rubric Element Difficulty Estimates. Overall, independent of assessment condition group, rubric elements differed in difficulty, χ^2 (3) = 134.3, p <.01 (see Table 8). For example, the 'Genre and Disciplinary Conventions' rubric element is the most difficult, with a logit value of 1.23 (*SE* = .19). The 'Control of Syntax and Mechanics' is the least difficult rubric element, with a logit value of -1.57 (*SE* = .19). These results relate to the raw score averages, where students scored highest on average on the 'Control of Syntax and Mechanics' rubric element, and the lowest on the 'Genre and Disciplinary Conventions' Rubric element (see Table 6). Furthermore, the 'Content Development' rubric element has an estimated difficulty of .78 (*SE* = .19), and the. 'Context of and Purpose for Writing' rubric element has an estimated difficulty of -.43 (*SE* = .18).

Score Category Threshold and Assessment Condition Information. Recall that the remaining scores in the dataset after data deletion range from 1 to 4. The frequencies of scores collapsing across rubric element can be found in Table 9. Specifically, score category 2 has the highest frequency (*n* = 242) and score category 4 has the lowest frequency (*n* = 56).

Furthermore, the Rasch-Andrich thresholds indicate, independent of assessment condition, the student ability level in logits where it is equally likely a student scores in the *k* or *k*-1 category. For example, the threshold -4.84 (*SE* = .14) indicates that a student with an ability level of -4.84 has an equal probability of getting a score of 1 and a score of 2. The threshold of -0.23 (*SE* = -.23) indicates that a student with an ability level of -0.23 has an equal probability of getting a score of a 2 and a score of 3. Finally, the

threshold of 5.07 (*SE* = .23) indicates that a student with an ability level of 5.07 has an equal probability of getting a score of a 3 and a score of a 4.

In addition to these score category information, I examined the average ability of students in both assessment conditions according to the Rasch model. Specifically, the non-embedded group is of lower ability on average than the embedded assessment condition $\chi^2(1) = 21.7$, $p < .01$.

Bias/Interaction Analysis. Recall that the bias/interaction analysis in FACETS (Linacre, 2017a) is essentially a t-test comparing the difficulty of each rubric element across the two assessment conditions (see Table 10). The contrast value in Table 10 represents the difference in the estimated rubric element difficulty across the two assessment conditions. This contrast value is then divided by the *SE* to get corresponding the *t*-statistic.

Out of the four rubric elements, two evidence possible DIF. The 'Content Development rubric element evidences DIF in favor of the embedded assessment condition, where $t(93) = 4.33$, $p < .001$. Specifically, the difficulty estimate for the 'Content Development' rubric element is 1.92 logits higher for the non-embedded assessment condition ($\delta_{non-embedded} = 2.16$) than the embedded assessment condition ($\delta_{embedded}$=.24). In other words, the 'Content Development' rubric element is more difficult for the non-embedded assessment condition than the embedded assessment condition.

In addition, the 'Control of Syntax and Mechanics' rubric element also evidences DIF $t(100) = 2.58$, $p < .05$. Yet the possible DIF evidenced in the 'Control of Syntax and Mechanics' rubric element favors the non-embedded assessment condition. Specifically,

the difficulty estimate for the 'Control of Syntax and Mechanics' rubric element is 1.06 logits lower for the non-embedded assessment condition ($\delta_{non-embedded} = -2.30$) than the embedded assessment condition ($\delta_{embedded}$ =-1.25). In other words, the 'Control of Syntax and Mechanics' rubric element is more difficult for the embedded assessment condition than the non-embedded assessment condition. Therefore, the two rubric elements with possible DIF favor different assessment conditions.

The other two rubric elements did not evidence DIF. Specifically, the difficulty estimates for each assessment condition were equal for the 'Genre and Disciplinary Conventions' rubric element, $t(99) = .01$, $p = .9894$. Lastly for the 'Context of and Purpose for Writing' rubric element, the difference between the non-embedded assessment condition ($\delta_{non-embedded}$ = -.94) and the embedded assessment condition ($\delta_{embedded}$ =-.20) was not statistically significant, where $t(100) = -1.85$, $p = .0667$.

Stage III: Differential Item Functioning Analysis with Ordinal Regression

Due to a small sample size and evidence of DIF in multiple items from the Rasch analysis, I also conducted an ordinal regression analysis. In contrast to the Rasch analysis, this ordinal regression analysis is an observed score DIF method using Verbal SAT scores as an external matching variable.

Within this section, I first describe the overall results of the ordinal regression. Each rubric element has its own regression analysis, with Verbal SAT and assessment condition as predictors and score category as the outcome. These results are all presented in the terms of log odds or logits (see Table 11). Yet log odds are not intuitive for many audiences, and therefore probabilities are primarily used to describe the results. All results are evidenced in tables and figures.

Overall Results: Logit Scale. Recall that the ordinal regression analysis is predicting the log odds of receiving a score in the j-1 score category over a score of j, where the predictors are Verbal SAT (i.e. ability) and assessment condition (i.e. group). Verbal SAT was centered, and assessment condition was dummy coded. Each of the four rubric elements had their own regression analysis. Furthermore, the interaction term between Verbal SAT and assessment condition was not significant in any of the four regression models for any of the rubric elements. Therefore, the interaction term, which would indicate non-uniform DIF was not included in the analyses. In addition, the Rasch model did not involve identifying non-uniform DIF, therefore the ordinal regression is in alignment with the same DIF investigation of the Rasch analysis.

Furthermore, Likelihood Ratio tests indicate the ordinal regression model with the assessment condition and Verbal SAT predictors significantly reduces the deviance compared to the intercept model (i.e. the null model). Specifically, the model with the Verbal SAT and assessment condition predictors fit significantly better than the intercept only model for the 'Context of and Purpose for Writing' rubric element $\chi^2(2) = 10.814$, $p = .004$, the 'Content Development' rubric element $\chi^2(2) = 35.520$, $p < .001$, the 'Genre and Disciplinary Conventions' rubric element $\chi^2(2) = 17.672$, $p < .001$, and the 'Control of Syntax and Mechanics' rubric element $\chi^2(2) = 15.069$, $p < .001$. In other words, the model predicting score category fits significantly better than the intercept only model for each ordinal regression analysis.

In addition to fit of the models, this adjacent category ordinal regression assumes equal slopes across all J-1 log odds being modeled. The following results assume this assumption is met. In order to assess whether the proportional odds assumption was met,

I compared the adjacent category model with the multinomial model using a Likelihood Ratio Test (LRT). A multinomial model allows the slopes of the predictors (Verbal SAT and assessment condition) differ for all *J*-1 category comparisons. Since there are four rubric elements, with four individual ordinal regressions, there was four model comparisons. All four LRT comparisons indicated that the multinomial model did not fit statistically significantly better than the adjacent category model. In other words, the proportional odds assumption for each of the four adjacent category models.

In general, the results indicate there is a statistically significant difference between assessment condition (i.e. embedded and non-embedded) and the log odds of scoring in the lower versus the higher of two adjacent categories, after controlling for Verbal SAT scores. In other words, after controlling for ability (i.e. Verbal SAT), students in each assessment condition differ in their log odds of receiving a particular score. Therefore, there is evidence of DIF for all rubric elements. Specifically, after controlling for Verbal SAT, students in the embedded assessment condition scored higher on average than students in the non-embedded assessment condition for the 'Control of and Purpose for Writing' rubric element, where $\chi^2(1) = 7.22$, $b = -.5252$, $p = .007$. This pattern is found in the remaining three rubric elements (See Table 11). Controlling for Verbal SAT, students in the embedded assessment condition scored higher on average than the students in the non-embedded condition for the 'Content Development' rubric element $\chi^2(1) = 22.73$, $b = -1.25$ $p < .001$, the 'Genre and Disciplinary Conventions' rubric element ($\chi^2(1) = 12.76$, $b = -.85$, $p < .001$), and the 'Control of Syntax and Mechanics' rubric element ($\chi^2(1) = 8.28$, $b = -0.67$ $p = .004$).

In addition to statistical significance, McFadden's R-squared and follow-up analysis of probabilities lend information regarding meaningful differences in the results. Specifically, the 'Content Development' rubric element had the largest McFadden's R-squared value, indicating a .0874 proportion of null deviance (i.e. the intercept only model) accounted for by the set of predictors (i.e. Verbal SAT and assessment condition). This is a medium to small effect size. The other rubric elements had small R-squared values, where 'Genre and Disciplinary Conventions' rubric element was the second largest with a .0456 proportion of null deviance accounted for by the set of predictors. The 'Context of and Purpose for Writing' rubric element had the smallest R-squared value of .0262, and the 'Control of Syntax and Mechanics' rubric element had the second smallest R-squared value of .0399.

Finally, follow-up analyses of probabilities suggest meaningful statistical DIF for the 'Content Development' and possibly the 'Genre and Disciplinary Conventions' rubric element. The pattern of probabilities for each assessment condition among the other criteria did not result in meaningful differences.

Continued Results: Probabilities. Within this particular section I describe meaningful trends in the probabilities of scoring in particular score categories per assessment condition across values of Verbal SAT. Table 12 organizes these score probabilities, and figures in Appendix G through J visualize this information. In particular, I first describe overall trends regarding P(Y=1) and P(Y=4) between assessment conditions, and then I discuss the meaningful DIF evidence regarding the P(Y=1) and P(Y=4) for the 'Content Development' and 'Genre and Disciplinary Conventions' rubric elements. The other two rubric elements did not evidence

meaningful differences in P(Y=1) and P(Y=4) between rubric elements, across values of Verbal SAT.

First, as Verbal SAT increases, probability of Y=1 decreases for both non-embedded and embedded assessment conditions. Yet the probability of scoring a 1 is consistently higher for the non-embedded assessment condition than the embedded assessment condition across values of Verbal SAT. In addition, the probability of Y=4 increases as Verbal SAT increases for both the non-embedded and embedded assessment conditions, though this is consistently higher for the embedded assessment condition. This trend occurs across all rubric elements, yet the magnitude of differences in P(Y=1) and P(Y=4) between assessment conditions was only judged to be meaningful for the Content Development and the Genre and Disciplinary Conventions rubric criteria.

Specifically, the 'Content Development' rubric element has the biggest difference in P(Y=1) between the assessment conditions across values of Verbal SAT. For example, the P(Y=1) for the non-embedded assessment condition is .411 higher than the P(Y=1) for the embedded assessment condition when Verbal SAT is 1 SD above the mean. Furthermore, the difference in P(Y=1) for the non-embedded assessment condition is .272 higher than P(Y=1) for the embedded assessment condition for the 'Genre and Disciplinary Conventions' rubric element, when Verbal SAT is 1 SD above the mean.

In contrast to the 'Content Development' and the 'Genre and Disciplinary Conventions' rubric element, the 'Context of and Purpose for Writing' and 'Control of Syntax and Mechanics' had small differences in P(Y=1) between the assessment conditions. Specifically, the difference in P(Y=1) between assessment conditions is .146 for the 'Context of and Purpose for Writing' rubric element and .072 for the 'Control of

Syntax and Mechanics' rubric element when Verbal SAT is 1 SD above the mean. Therefore, the differences in P(Y=1) between assessment conditions is not meaningful for two of the rubric elements, 'Context of and Purpose for Writing' and 'Control for Syntax and Mechanics'. Yet the differences in P(Y=1) between assessment conditions is meaningful for the 'Content Development' and 'Genre and Disciplinary Conventions' rubric elements.

Similar to P(Y=1), the 'Content Development' rubric element has the biggest difference in P(Y=4) between assessment conditions, yet the non-embedded assessment condition has a higher probability than the embedded assessment condition. For example, the difference in the P(Y=4) for the embedded assessment condition is .202 higher than the P(Y=4) for the non-embedded assessment condition for the 'Content Development' rubric element when Verbal SAT is 1 SD above the mean.

Furthermore, the difference in P(Y=4) between assessment conditions for no other rubric elements was judged as less meaningful. Specifically, the difference in P(Y=4) is .101 for the 'Genre and Disciplinary Conventions' rubric element, .121 for the 'Control of Syntax and Mechanics' rubric element, and .115 for the 'Context of and Purpose for Writing' rubric element when Verbal SAT is 1 SD above the mean. Therefore, the largest differences in the P(Y=4) is within the 'Content Development' rubric element, where students in the embedded assessment condition are more probable to score a 4 than their equal ability counterparts in the non-embedded assessment condition. In contrast, the 'Content Development' rubric and the 'Genre and Disciplinary Conventions' rubric element evidence the largest meaningful differences in the P(Y=1). Specifically, students in the non-embedded assessment condition are more likely to score a 1 than their equal

ability counterparts in the embedded assessment condition.**A Synopsis: Rasch and A A Synopsis: Rasch and Ordinal Regression Results**

In summation, there are raw mean differences in rubric element scores between the assessment conditions, where the embedded assessment condition scored consistently higher on average than the non-embedded assessment condition. In addition, the Rasch model evidenced plausible DIF in the 'Content Development' and 'Control of Syntax and Mechanics' rubric element. Specifically, the DIF evidenced for the 'Content Development' rubric element favored the embedded assessment condition. In contrast, the DIF evidenced for the 'Control of Syntax and Mechanics' rubric element favored the non-embedded assessment condition. These seemingly contradicting results will be discussed further in the 'Discussion' chapter.

Lastly, all rubric elements evidence plausible DIF from the Ordinal Regression model. However, further investigation suggests that there may only be meaningful differences between assessment conditions within the 'Content Development' and the 'Genre and Disciplinary Conventions' rubric elements. Specifically, these meaningful differences between assessment conditions are most prevalent for Y=1 and Y=4. Specifically, the non-embedded assessment condition has a higher P(Y=1) than the embedded assessment condition for all rubric elements across all values of Verbal SAT. The difference in probability for P(Y=1) between groups was found for the "Content Development' element and the 'Genre and Disciplinary Conventions' rubric element. In contrast, the non-embedded assessment condition has a lower P(Y=4) than the embedded assessment condition. The difference in the P(Y=4) between assessment conditions is evidenced in

the ‘Content Development’ element similar to the difference in the P(Y=1) between assessment conditions.

CHAPTER FIVE

Discussion

In the current study, I focused on the AAC&U Written Communication VALUE rubric, which is implemented by the Multi-State Collaborative in a nation-wide course-embedded assessment initiative. Specifically, I examined its functioning in non-embedded and course-embedded assessment conditions, and how these assessment conditions may influence possible DIF. Recall that DIF occurs when students in two different groups, but of the same ability, have differing probabilities of obtaining a particular score. In addition, recall that DIF may be an indication of construct-irrelevant variance. As applied in this study, construct-irrelevant variance reflects systematic variance not pertinent to written communication (Messick, 1995). For the current study, I investigated whether the score meanings held across assessment contexts. If they do not, there may be systematic variance involving some other construct or method effect not pertinent to written communication.

Within this study, DIF was examined using a Rasch model and an ordinal regression analysis. Across both approaches there is evidence of DIF for the 'Content Development' rubric element. Importantly, evidence of DIF was found after removing students from the sample who self-identified as unmotivated. The following sections focus on the possible causes of DIF which include: time constraints, task structure of the assignments, opportunity for feedback, and maturation differences between the two assessment conditions. Yet before I explain possible causes of the DIF results, I integrate the results from the two DIF methods.

Differences in the DIF Methods and Corresponding Results

The Rasch method indicated DIF for the 'Content Development' and the 'Control of Syntax and Mechanics' rubric elements. The ordinal regression analysis evidenced DIF in the 'Content Development' rubric element similar to the Rasch method, but did not find DIF in the 'Control of Syntax and Mechanics' rubric element. Instead, the ordinal regression analysis evidenced DIF in the 'Genre and Disciplinary Conventions' rubric element. These differences in results may be explained by the differences in the matching variable and the estimation of DIF between the two methods.

Specifically, the Rasch method estimates ability using what is analogous to a sum score of the Written Communication rubric assessment. In contrast, the Ordinal Regression method is an observed score method using an external matching variable of ability (i.e. Verbal SAT). In other words, these analyses differ in how they match students on ability. Therefore, different students may be compared at differing levels of ability across DIF methods, depending on the matching variable. For example, within the Ordinal Regression analysis, students with the same Verbal SAT score are compared in their probabilities of obtaining a particular score on each rubric element. In contrast, the Rasch method compares the probabilities of receiving a particular score for students with the same sum score across all rubric elements. Due to these differences, the results may not be exactly the same across both methods due to the difference in matching variables across both methods.

In addition to different matching variables, the two DIF methods differ in their estimation of DIF values. Specifically, the Ordinal Regression model estimated DIF in individual analyses for each rubric element. In contrast, the Rasch model estimated DIF

for all items simultaneously. Some possible issues arise for the Rasch model due to this simultaneous estimation. Specifically, there is a possible circularity problem with DIF methods which use the sum score as the matching variable (Navas-Ara & Gomez-Benito, 2002). The circularity problem arises when there is an item with DIF, then the matching variable is biased when investigating DIF. Yet there are multiple ways in which researchers attempt to avoid this circularity problem. For example, some researchers use a purification technique, where the greatest DIF presenting items are eliminated in the first stage of analyses and a second DIF analysis follows to identify the presence of any other DIF items (Clauser, Mazor, & Hambleton, 1993; Holland & Thayer, 1988).

Construct underrepresentation is a consideration for purification in the current study. Specifically, the purification procedure changes the very nature of what is assessed when there are such a small number of items (i.e. rubric elements). In this case, even with a large sample, it would be inappropriate to execute a purification technique because eliminating one rubric element alters the meaning of ability in written communication.

Furthermore, when one item (or rubric element) evidences DIF, then other items may evidence DIF that do not have true DIF (Lee & Geisinger, 2016). This may be occurring in the Rasch results for this current study. For example, the Rasch results indicate both the ‘Content Development’ and ‘Control of Syntax and Mechanics’ rubric element evidence DIF. Yet the DIF results in one of these rubric elements may have influenced DIF in the other rubric element. In other words, the possible true DIF in the ‘Content Development’ rubric element, may influence DIF evidence for the ‘Control of Syntax and Mechanics’ rubric element.

In addition, previous research indicated situations where items that do not have true DIF may evidence DIF favoring the opposite group than other DIF items (Magis & Facon, 2013). This may be plausible due to the DIF in the 'Control of Syntax and Mechanics' rubric element favoring the non-embedded assessment condition, and the DIF in the 'Content Development' rubric element favoring the embedded assessment condition. Intuitively, it is not clear as to why one rubric element would favor one assessment condition, and the other rubric element favor the other assessment condition.

Recall this circularity problem is most pressing when item parameters are estimated simultaneously. The ordinal regression estimates DIF using the external matching criterion, Verbal SAT, avoiding this circularity problem. Therefore, in addition to the issue of matching variables, the somewhat contradictory DIF results across both methods may also be due to the circularity problem in the Rasch analysis.

The evidence of DIF for the Control of Syntax and Mechanics' is weak due to this circularity problem in the Rasch method, and no evidence of DIF for the 'Control of Syntax and Mechanics' rubric element in the ordinal regression analysis. Both the Rasch analysis and the ordinal regression analysis indicate DIF in the 'Content Development' rubric element. Furthermore, the ordinal regression indicates meaningful differences in the P(Y=1) between assessment conditions across values of Verbal SAT for both the 'Content Development' and 'Genre and Disciplinary Conventions' rubric. In addition there are meaningful differences in P(Y=4) between assessment conditions across values of Verbal SAT for the 'Content Development' rubric element. Therefore, there is moderate evidence of possible for the 'Content Development' and the 'Genre and Disciplinary Conventions' rubric elements.

Possible DIF Explanations

There may be multiple ways to explain this possible evidence of DIF. In general, measurement researchers describe evidence of statistical DIF as an indication of possible construct-irrelevant variance. Recall that construct-irrelevant variance occurs if assessment scores contain systematic variance not pertinent to the construct of interest (Messick, 1995). For the current study, the assessment condition may be contributing construct-irrelevant variance in the AAC&U Written Communication scores. Within this section, I describe some possible contributors of construct-irrelevant variance in the rubric scores due to the assessment condition: 1) time, 2) task structure, 3) feedback, and 4) maturation.

Time. According to the AAC&U Written Communication rubric, in order for students to score higher (e.g. Y =4) in the 'Content Development' rubric element, students must use, "appropriate, relevant, and compelling content to illustrate mastery of the subject" (AAC&U, 2009). In contrast, in order to score a 1 a student must use, "appropriate and relevant content to develop simple ideas in some parts of the work"" (AAC&U, 2009). Therefore, to score higher on the 'Content Development' rubric element students must move from using simple ideas in some parts of their work regarding the content domain, to providing evidence of mastery of the relevant content domain. One may need more time to think, connect ideas, and organize subdomain knowledge in order to evidence mastery.

Possible construct-irrelevant variance for the scores from the AAC&U VALUE rubric for multiple rubric elements (not just for 'Content Development') may be due to a

time constraint. Specifically, there were two different types of time constraints between the assessment conditions. The non-embedded assessment condition had a 60-minute time limit but there was an extended amount of allotted time for the embedded assessment condition. Though the assignments within the embedded assessment condition had varying time limits, their due dates gave students more time than the 60 minute time constraint evidenced in the non-embedded assessment condition.

Some research indicates that time-allotted influences test performance. According to Powers and Fowles (1996), students performed significantly and practically better when given 60 minutes instead of 40 minutes to take the GRE Writing assessment. Though some researchers found allowing more time increases writing scores in essay exams (Biola, 1982; Hale, 1992) other investigators have failed evidencing this difference (e.g. Caudery, 1990; Livingston, 1987). Yet the current study did not only have assessment conditions differ in the amount of time allotted, but student in the embedded assessment condition also had more opportunities to revisit their assignment across multiple occasions. For example, students in the non-embedded assessment condition had one opportunity to work on their assignment, but the other group of students were allowed multiple opportunities to work on their performance within the time between given their assignment and the due date.

Many writing assessment theorists indicate timed-writing assessments underrepresent the written communication construct. Specifically, these timed assessments do not allow for the writing process, a key element to the written communication construct (CWPA, 2011). In addition, timed-writing assessments are first-drafts of a student's writing and therefore, "probably doesn't represent fairly or

accurately the full range of student's writing ability or event thinking" (Brown, 2010, p. 227). Furthermore, proficient writing requires flow between, "planning, generation, and reviewing, as the author attempts to solve the content problem of what to say and the rhetorical problem of how to say it" (Kellog & Whiteford, 2009, p. 255). Therefore, it seems it is necessary for writers to have a significant amount of time, and possibly more than one opportunity to produce proficient writing.

Maturation. The students in the non-embedded assessment condition had about 42 less credit hours on average than students in the course-embedded assessment condition. If students take about 30 credit hours a year, then the students in the course-embedded assessment condition are about 1.5 academic years ahead of the students in the non-course embedded assessment condition. Therefore, students in the embedded assessment condition may have a higher probability of getting higher scores compared to students in the non-course embedded assessment condition with the same Verbal SAT score simply due to maturation.

According to Kellog and Raulerson III (2007), deliberate practice in higher education is a means to improve writing skills of college students. In particular, students should have spaced practice, over time to improve their writing skills. In addition, Johnstone, Ashbaugh, and Warfield (2002) found that students who were accounting majors and took writing intensive courses had higher performance in writing than students who did not take the writing courses.

Given that students with more credit hours are more likely to have more practice writing, these students are more likely to perform better than students with less credit hours, who are not as likely to have as much writing practice. Therefore, the students in

the embedded assessment condition may have higher probabilities of getting a higher score across the rubric elements, simply due to these students having more opportunity for increased practice across more time than their non-embedded assessment counterparts.

Furthermore, Verbal SAT indicated ability for the ordinal regression procedure. Given that SAT is an indication of ability at the time of admittance to a university, maturation in writing ability may make the use of Verbal SAT problematic as an external criterion. In other words, students with more credit hours or writing experience from their coursework may not be matched to their current ability levels. In contrast, maturation is less problematic for the Rasch procedure, which uses what is analogous to a sum score across rubric elements as an indicator of ability.

Feedback. In addition to maturation and allotted time, students in different assessment conditions but of the same ability (e.g. Verbal SAT), may differ in their probabilities of getting particular scores due to faculty feedback. The non-course embedded assessment condition did not allow for feedback on student writing performance. In contrast, there was opportunity for students to get feedback on their performance for the embedded assessment condition.

In particular, getting feedback is part of the writing process where students have an opportunity to better their performance (Hull, 1987; Kellog & Raulerson III, 2007). According to Wingate (2010), "students who had utilized their feedback comments improved in the areas previously criticized, and did not receive the same criticisms again. They demonstrated awareness of many details of their feedback, and had taken action as a result" (p. 531). Assuming students in the embedded course assessment condition did

similar behaviors when given feedback, students in the course embedded assessment condition are more likely to perform better than their non-course embedded assessment counterparts of the same ability.

For example, in order for students to score a 4 in the 'Genre and Disciplinary Conventions' rubric element students must demonstrate, "detailed attention to and successful execution of a wide range of conventions particular to a specific discipline and/or writing task(s) including organization, content, presentation, formatting, and stylistic choices" (AAC&U, 2009). A student with feedback on a previous draft is more likely to master this demand for detailed attention and execution of conventions, not only because they have another opportunity to do so, but also because the instructor is likely to guide a student to execute higher performance in this domain.

In contrast, to score a 1 for the 'Genre and Disciplinary Conventions' rubric element, students must follow, "expectations appropriate to a specific discipline and/or writing task(s) for basic organization, content, and presentation" (AAC&U, 2009). This performance seems more likely of someone who does not get feedback in contrast to a score of a 3 or a 4 which would be more likely for a student who received previous feedback.

Task Structure. A possible fourth contributor to construct-irrelevant variance is the difference in task structure across the two assessment conditions. For the current study, the task structure references the prompt or assignment. Recall that task structure can differ depending on specific characteristics of the writing prompt such as the length of the prompt, specificity of the prompt, linguistic level of the prompt, and the content necessary to complete the task.

For the non-embedded assessment condition, the task structure was the same for all students (See Appendix B). The students in the embedded assessment condition had differing task structures across all five of the assignments. These assignments within the embedded condition differed on their specificity of task, word choice, and on the content necessary to complete the task. For example, the assignment for an introductory psychology course gave students examples and descriptions of key pieces to a paper: thesis, an argument, counterargument, response to a counterargument, and a conclusion. In contrast, when asked about the English Capstone assignment, the faculty member said there were no written instructions. In other words, the professor simply told them to write a paper on a topic of their choice.

In particular, the two assignments differ in the specificity of task where the introductory psychology class students were given detailed instructions and the English capstone course students were given little to no instruction for their assignment. These differences may contribute to why students of the same writing ability, have different probabilities of obtaining particular scores for particular rubric elements. In other words, these differences in the task structure may contribute to the possible construct-irrelevant variance in written communication assessment scores (Cooper, 1984; Huot, 1990; Schoonen, 2005).

Specifically, the position that students differ in their writing ability across content topics is not a profound idea for English teachers. According to Palmer (1966), “English teachers hardly need to be told that there exists a great deal of variability in student writing from one theme to another and from one essay to another. The most brilliant students may do well on one essay topic and badly on another” (p. 288). Furthermore,

generalizability studies provide numerous evidence of a task x person interaction. In other words, constructed responses assessments like the Written Communication assessment within the current study, often have scores with significant variability due to the task (Shavelson, 2013).

Generalizability theory would describe this effect as an interaction, where student performance depends on the task. Research within the generalizability framework indicates that it is necessary to have students complete a greater number of tasks in order to retain reliability, especially in performance tasks such as writing assessments (Lane & Stone, 2006). In particular, Lane, Liu, Ankenmann, and Stone (1996) investigated the number of tasks necessary to reach a generalizability coefficient of .80 for a math performance assessment. Thirty-six tasks were necessary to reach this coefficient. Though this was a math assessment and not a written communication assessment, research has also indicated that writing assessments have a significant person x task interaction as well (Lane & Stone, 2006).

Yet these differences in performance due to task may be from the specificity of the prompt (Brennan, Gao, Colton, 1995; Godshalk, Swineford, & Coffman, 1966; Shavelson, Baxter, & Gao, 1993), the type of writing (i.e. creative or analytic) demanded by the prompt (Bouwer, Beguin, Sanders, & Bergh, 2015, Crowhurst, 1980; Reed, Burton, & Kelly, 1985; Rosen, 1969), or the wording of the prompt itself (Abedi &Lord, 2001; Huot, 1990). There is research supporting how all these variations in task structure can influence writing performance. Yet there is no evidence of which difference in task structure is contributing to the possible construct-irrelevant variance in the scores of this particular study.

In addition, there is task variability within the embedded assessment condition, but not within the non-embedded assessment condition. Furthermore, there is task variability between the assessment conditions. Both the within variability of task structure of the embedded assessment condition and the variability of task structure between the two assessment conditions may contribute to the evidence of DIF and possible construct-irrelevant variance of the scores. In order to decrease this variability, or in other words decrease the noise within the variability between scores, the task should be the same across both assessment conditions. In other words, the specificity, wording, and content of the task should be consistent across both assessment conditions.

Limitations of the Current Study and Directions for Future Research

First, sample size for the current study is small. Recall that there were only 50 scored student performances for the non-embedded assessment condition, and 107 scored student performances for the embedded assessment condition. In addition, there is scarce performances for Y = 4 across all of the rubric elements. Due to such a small sample size, and scarce data for some score categories, the standard errors in the results are large. In addition, more data overall contributes to more stable estimates. In general, it is recommended to have at least 100 scores for each group when conducting a DIF analysis (Scott et al., 2009).

A second limitation of the current study is the control for motivation. Specifically recall that motivation researchers recommend data to meet two criteria before conducting motivation filtering (Wise, Wise, & Bhola, 2006). First, there must be a correlation between test performance and test motivation. If there is not a correlation between motivation and performance, filtering becomes obsolete. In other words, there is no

reason to filter. Three of the four rubric elements had no relationship with motivation (see Table 4). Yet there was a statistically significant negative correlation between the SOS Effort subscale scores and the 'Genre and Disciplinary Conventions' rubric element scores. Therefore, the first criterion was not completely met to conduct motivation filtering. In contrast, the second assumption pertaining to no relationship between test motivation and ability was met.

Though one of two criteria for motivation filtering was not met, recall only one student in the non-embedded assessment condition had a score less than 12 and was therefore deleted from the sample. The current study argues the difference in score meaning across the assessment conditions is not due to differences in motivation. Yet given that there was only 1 student below cutoff, it may be that motivation may not be as large of an issue in the non-embedded assessment condition. In other words, maybe motivation is not as problematic as suspected in non-embedded assessments, and therefore indicating motivation may not be as serious potential source of construct irrelevant variance. Future research should further investigate this claim of motivation and construct-irrelevant variance for the particular non-embedded assessment condition of this study.

A third limitation is the difference in credit hours between the assessment conditions. Recall that on average the students in the embedded assessment condition have more credit hours than students in the course embedded assessment condition. Due to this difference in credit hours, the evidence of DIF between the assessment conditions may be due to a maturation effect. In other words, students of the same writing ability may have differing probabilities of getting particular scores because students in the

embedded assessment condition had more opportunity to practice writing than their non-course embedded counterparts. Future research should either randomly assign students into each assessment condition to avoid this difference in credit hours, or in some way create groups with equal credit hour or writing experience. Therefore, researchers may be able to defend that these results of DIF may not be due to maturation.

A fourth limitation is the use of Verbal SAT as an external matching variable within the ordinal regression DIF analysis. In the current study, and likely for other educational researchers at a higher education institution, Verbal SAT scores are available and accessible. In addition, these scores are available for almost every student at the institution. Though these scores may be convenient, recall that the Verbal SAT score represents both reading and writing skills (College Board, 2018b). Therefore, Verbal SAT may not adequately represent writing ability, but instead contain construct-irrelevant variance within Verbal SAT scores due to the reading component of the measure (College Board, 2018a). Furthermore, the Writing and Language test portion of the Verbal SAT is a multiple-choice assessment. Therefore, there are going to be differences between what is being measured by the Verbal SAT multiple-choice assessment and what is being measured when a student creates a written product such as the scored written products in the current study. In other words, the scores of the Verbal SAT represent a ability than what is being measured by the AAC&U Written Communication VALUE rubric.

In addition, the Verbal SAT assessment contains strict time-constraints. Therefore, the issue of time-constraints contributing to construct-irrelevant variance in Verbal SAT scores is a concern. Readers should take into consideration the possibility of

construct-irrelevant variance in the external matching variable. In the future, researchers should consider using a better representation of writing ability for DIF analyses.

The final limitation is the difference in assignments across the assessment conditions. Specifically, there was only one prompt related to the student performances of the non-course embedded assessment condition. In contrast, the performances from the embedded assessment condition were a result of five different assignments across five differing content domains. In addition, all of these assignments differed in the length and specificity of instruction. Recall that research indicates that the quality of performances can differ due to different task structures (Huot, 1990). Therefore, in order to reduce the systematic variability between assessment scores simply due to differences in task structure, future researchers should use a common assignment. In other words, future researchers should investigate whether DIF occurs when holding task-structure constant between conditions. If DIF is present with a common assignment, then it is possible the DIF evidence in this current study is not due to differences in task across assessment conditions.

Implications and Conclusions

Written communication skills are necessary for success as a student in higher education, and as a post-graduate employee (Sparks et al., 2014). Furthermore, employers not only desire their employees to have high writing skills, in general, these employers are dissatisfied with the writing skills of post-graduate employees (ManpowerGroup, 2012; Markle et al., 2013; The Manufacturing Institute, 2011). Therefore, there is pressure on higher education to increase graduates' writing skills, and therefore there is no surprise that higher education prioritizes written communication skills across their

curriculum. In addition to curricular additions to writing, writing assessments are also in high demand within higher education.

Yet there is not a consensus across writing assessment researchers, and the practice of large scale writing assessments, about the best way to measure student proficiency (O'Neill & Murphy, 2012). One major concern is the use of timed writing assessments that do not reflect writing ability, and are not reflective of the type of writing expected within a curriculum (Calfee & Miller, 2007; O'Neill & Murphy, 2012). Another major concern is the use of low-stakes assessments where students have lower motivation to perform to the best of their ability (DeMars, 2000). In other words, students participating in low-stakes assessments, on average, perform lower than students who have consequences for their performance. Recall that this motivation issue increases for low-stakes constructed-response assessments compared to low-stakes selected-response assessments (Liu, Bridgeman, & Adler, 2012).

Within the low-stakes assessments where students are more likely to have low motivation, their scores contain variability due to something other than the ability of the intended construct being measured. In other words, the scores contain construct-irrelevant variance, or the variability in scores or not necessarily due to differences in ability but to differences in motivation (Barry et al., 2010). Due to such concerns, some have argued to adopt a course-embedded assessment approach (Coates & Seifert, 2011). Specifically, The Multi-State Collaborative is a national initiative which has adopted course-embedded assessments in an attempt to alleviate the contribution of construct-irrelevant variance in the scores due to low motivation in low-stakes assessments (AAC&U, 2017).

One of the assessments of the Multi-State Collaborative is the AAC&U Written Communication VALUE rubric. The current study investigated whether, after controlling for motivation, there was still possible construct-irrelevant variance between assessment conditions that were either a) low-stakes such as a non-course embedded assessment or b) high stakes such as a course-embedded assessment. The findings in this study indicate that there may be more than a motivation issue to consider between low and high-stakes assessments. Other possible contributors to construct-irrelevant variance include: time constraints, availability of feedback, and differences in task structures across the assessment conditions within this particular study. Future researchers should investigate these possible contributions to construct-irrelevant variance within the assessment scores.

Though further research is necessary, the current evidence poses there are issues with non-course embedded or high-stakes assessments other than motivation. These possible issues include the limited time and limited availability of feedback often found in non-embedded or high-stakes assessment situations. Yet approaches such as assessment day give researchers and assessment coordinators a random sample and the ability to randomly assign students to a particular battery of assessments. Yet evidence from the current study indicates that students of the same ability have lower probabilities of getting higher scores when in the non-embedded assessment condition compared to the embedded assessment condition.

Therefore, should assessment specialists avoid non-embedded writing assessments? Though course-embedded approaches overcome some of the challenges of non-course embedded assessments, they also have their own limitations. For example, if

we use a course-embedded approach we can potentially overcome issues with time, feedback, and so forth. However, we still face the challenge of determining which assignments to sample. Generalizability theory implies that task-specificity research (person by task interaction) is a complicated issue given that who we think is doing better tends to change across multiple tasks. This happens even when tasks are intentionally designed to measure the same thing (McBee & Barnes, 1998). So, what assignments and how many do we sample per student to get a reliable and accurate indication of their writing ability?

Given the strengths and limitations of both non-course embedded and course-embedded assessments, assessment practitioners must weigh the strengths and weakness of different strategies as they decide how to assess writing. Moreover, additional research needs to be conducted so that we minimize the weaknesses of each approach. This is going to take a concerted effort across measurement specialist, assessment researchers, and content experts, with the ultimate goal of obtaining assessment scores as representative as possible of writing ability.

Table 1. *Mapping Written Communication elements to key Frameworks*

Dimensions of writing construct	Frameworks		
	CWPA, NCTE, & NWP	DQP	LEAP
Genre, context, and purpose	X	X	X
Audience Awareness	X	X	X
Use of sources and textual evidence	X	X	X
Processes (planning, drafting, revision)	X		
Modes and forms (multimedia, digital)	X	X	X
Language conventions (grammar, syntax, and mechanics)	X	X	X

Note. X = directly mentioned in the framework. CWPA= Council of Writing Program Administrators; NCTE = National Council of Teachers of English; NWP = National Writing Project; DQP = Degree Qualifications Profile; LEAP = AAC&U's Liberal Education and America's Promise. CWPA and NCTE collaborated in the National Writing Project (NWP)'s *Framework for Success in Postsecondary Writing* (2011). Table adapted from Sparks et al. (2014).

Table 2. *Mapping written communication elements to assessments.*

Components of writing construct	Assessments					
	CAAP		**Proficiency Profile**		**CLA**	**AAC&U VALUE**
	SR	CR	SR	CR	CR	CR
Genre, context, and purpose	X					X
Audience Awareness	X	X				X
Use of sources and textual evidence	X	X		X	X	X
Processes (planning, drafting, revision)						
Modes and forms (multimedia, digital)	X					X
Language conventions (grammar, syntax, and mechanics)	X	X	X	X	X	X

Note. X = aligned with assessment. CAAP = Collegiate Assessment of Academic Proficiency; CLA = Collegiate Learning Assessment; AAC&U VALUE = Association of American Colleges & Universities Valid Assessment of Learning in Undergraduate Assessment Rubric. SR = Selected Response; CR = Constructed Response. Table adapted from Sparks et al. (2014).

Table 3. *DIF Approaches*

Type of Procedure and Matching Variable	Parametric	Non-Parametric
Observed Score	Polytomous Logistic Regression (PLR)	Mantel-Haenszel
	Polytomous Logistic Discriminant Analysis (PLDFA)	Standardized Mean Difference (SMD)
Latent	Partial Credit Rasch Model	Polytomous SIBTEST

Note. Parametric = the approach incorporates the relationship between the item score and matching variable. Non-parametric = approach does not incorporate the relationship between the item score and the matching variable. Observed = the approach estimates ability using an observed score. Latent = uses an estimate of latent ability. Table adapted from Penfield & Lam (2000) and Potenza & Dorans (1995).

Table 4. *Correlations for motivation filtering*

	1	2	3	4	5	6
1. SAT1/verb	575(67.85)					
2. Context	0.122	2.28(.956)				
3. Content	0.124	0.795**	2.04 (.993)			
4. Genre	0.116	0.717**	0.742**	1.96(.884)		
5. Syntax	0.184*	0.696**	0.695**	0.611**	2.47(.813)	
6. SOS - Effort	-0.050	-0.113	-0.075	-0.281*	-0.090	19.80(3.11)

Note. Means are on the diagonals with standard deviations in the parentheses. * $p < .05$, ** $p < .01$

Table 5. *Number of scored products per each assessment condition after data deletion procedures*

Assessment Condition	Assignment	*N*
Non-Embedded	Assessment Day Pre-test	33
	Assessment Day Post-test	17
Embedded	English	16
	History	27
	Intelligence Analysis	10
	Philosophy	22
	Psychology	32

Note. For the non-embedded assessment condition, *N* = 50. For the course-embedded assessment condition, *N* = 107.

Table 6. *Descriptives for rubric element raw scores after motivation filtering*

Rubric Element	N	*M*	*SD*	Min	Max
Context	157	2.28	0.966	1	4
Content	157	2.05	1.011	1	4
Genre	157	1.97	0.909	1	4
Syntax	157	2.49	0.821	1	4

Table 7. *Differences in rubric element performance between assessment conditions*

	Non-Embedded	Embedded			
Rubric Element	M(SD)		*U*	*Z*	*p*
Context	1.98(.820)	2.42(1.00)	2016.00	-2.597	.009
Content	1.44(.644)	2.34(1.03)	1336.00	-5.304	<.001
Genre	1.58(.673)	2.15(.950)	1779.00	-3.580	<.001
Syntax	2.22(.708)	2.62(.843)	2000.50	-2.723	.006

Note. For the non-embedded assessment condition, $N = 50$. For the course-embedded assessment condition, $N = 107$.

Table 8. *Fit indices and difficulty estiamtes for each rubric element and assessment condition.*

					Infit		Outfit	
		Fair Average	Difficulty	*SE*	*MNSQ*	*ZSTD*	*MNSQ*	*ZSTD*
Rubric Element								
	Context	2.27	0.43	0.18	0.79	-1.90	0.74	-2.20
	Content	2.03	-0.78	0.19	0.86	-1.30	.78	-1.70
	Genre	1.96	-1.23	0.19	1.11	0.90	1.09	0.60
	Syntax	2.49	1.57	0.19	1.14	1.10	1.50	2.70
Assessment Condition								
	Non-embedded	2.09	2.03	0.17	0.92		0.89	
	Embedded	2.28	1.09	0.11	0.99		1.09	

Note. For the rubric elements, $\chi^2(3) = 134.3$, $p < .01$. Embedded group is of higher ability though the ability values for the non-embedded group are higher than the non-embedded group. See the variable map in Figure 7 for further explanation.

Table 9. *Score Category descriptive information based on the Rasch model.*

Score Category	Absolute Frequency	Relative Frequency	Outfit MNSQ	Threshold	SE
1	133	0.23	0.90		
2	242	0.41	0.80	-4.84	.14
3	153	0.26	1.40	-0.23	.16
4	56	0.10	1.00	5.07	.23

Note. Thresholds are Rasch-Andrich thresholds. *SE* = Standard error.

Table 10. *Bias interaction results*

	Assessment Condition					
Criteria	Non-embedded	Embedded	Contrast	Joint SE	*t*	df
Context	-0.94 (.33)	-0.20 (.22)	-0.74	.40	-1.85	100
Content	2.16 (.38)	0.24 (.22)	1.92	.44	4.33**	93
Genre	1.24 (.34)	1.24 (.23)	0.01	.41	0.01	99
Syntax	-2.30 (.34)	-1.25 (.23)	-1.06	.41	-2.58*	100

Note. * $p < .05$ ** $p < .001$

Table 11. *Ordinal regression results.*

	Context		Content		Genre		Syntax	
Parameter	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
Logit 1 vs Logit 2	-0.169	0.235	0.777**	0.244	0.425	0.232	-1.205**	0.306
Logit 2 vs Logit 3	0.742**	0.252	1.729**	0.347	1.415**	0.309	0.657**	0.246
Logit 3 vs Logit 4	1.169**	0.329	1.558**	0.412	1.756**	0.433	1.744**	0.353
Verbal SAT	-0.002	0.001	-0.002	0.001	-0.002	0.001	-0.004*	0.001
Assessment Condition	-0.525**	0.195	-1.250**	0.262	-0.845**	0.236	-0.672**	0.233

Note. Verbal SAT is centered at 0. The non-embedded condition is coded as 0, and the embedded assessment condition is coded as 1. * $p < .05$ ** $p < .01$

Table 12. *Differences in probabilities across assessment conditions for each rubric element.*

Rubric Element	Verbal SAT	ΔP(Y=1)	ΔP(Y=2)	ΔP(Y=3)	ΔP(Y=4)
Context	- 1 SD (505.97)	0.169	0.012	-0.104	-0.076
	Mean (574.97)	0.159	0.038	-0.102	-0.095
	+1 SD (643.97)	0.146	0.064	-0.095	-0.115
Content	- 1 SD (505.97)	0.401	-0.119	-0.163	-0.119
	Mean (574.97)	0.411	-0.076	-0.177	-0.158
	+1 SD (643.97)	0.411	-0.024	-0.185	-0.202
Genre	- 1 SD (505.97)	0.273	-0.083	-0.131	-0.059
	Mean (574.97)	0.275	-0.053	-0.144	-0.078
	+1 SD (643.97)	0.272	-0.017	-0.154	-0.101
Syntax	- 1 SD (505.97)	0.121	0.084	-0.141	-0.064
	Mean (574.97)	0.096	0.126	-0.132	-0.091
	+1 SD (643.97)	0.072	0.156	-0.108	-0.121

Note. A negative indicates the embedded assessment condition has a greater probability of receiving a particular score category (Y=*j*), for a specific level of Verbal SAT.

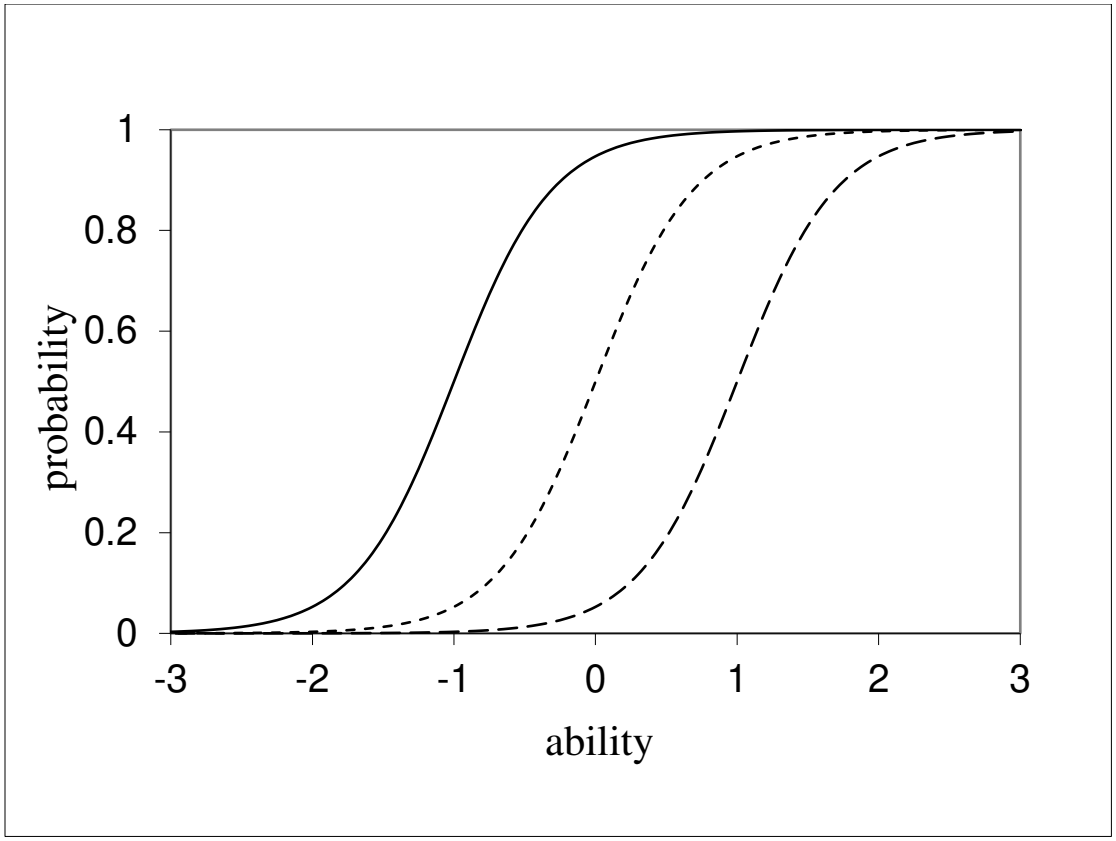


Figure 1. An example of a 1 PL model with three items. All items have a *c* parameter of 0, an *a* parameter of 1.7, and a range of *b* parameters at: -1, 0, and 1.

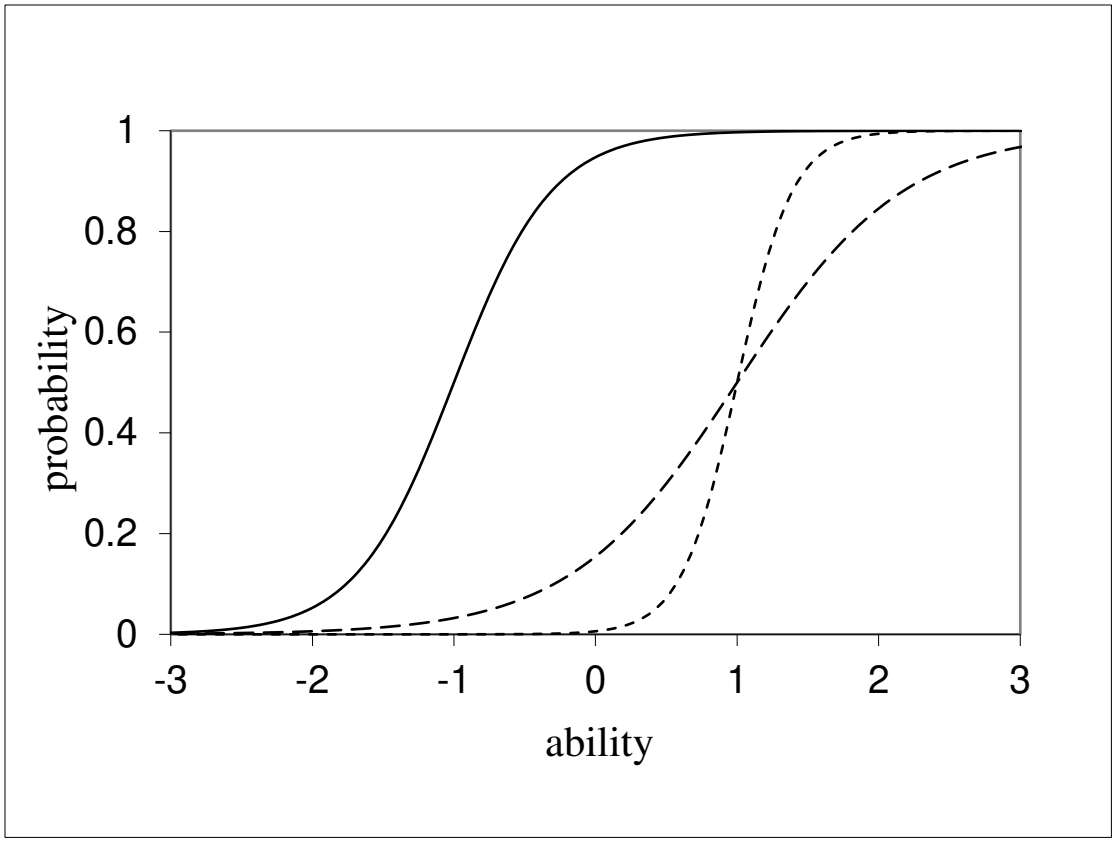


Figure 2. An example of a 2 PL model where the *a* and *b* parameters differ across three items, but the *c* parameter (i.e. lower asymptote) is held constant at 0.

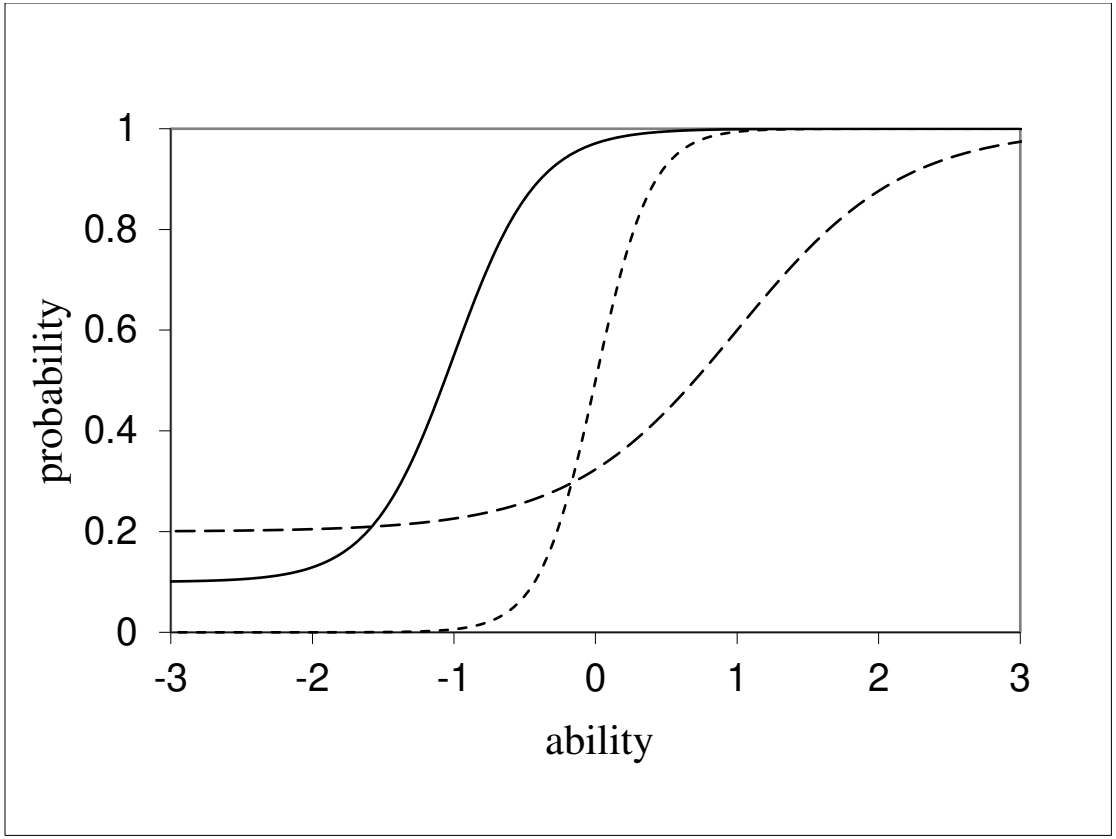


Figure 3. An example of a 3 PL model where the *a*, *b*, and *c* parameters differ across three items.

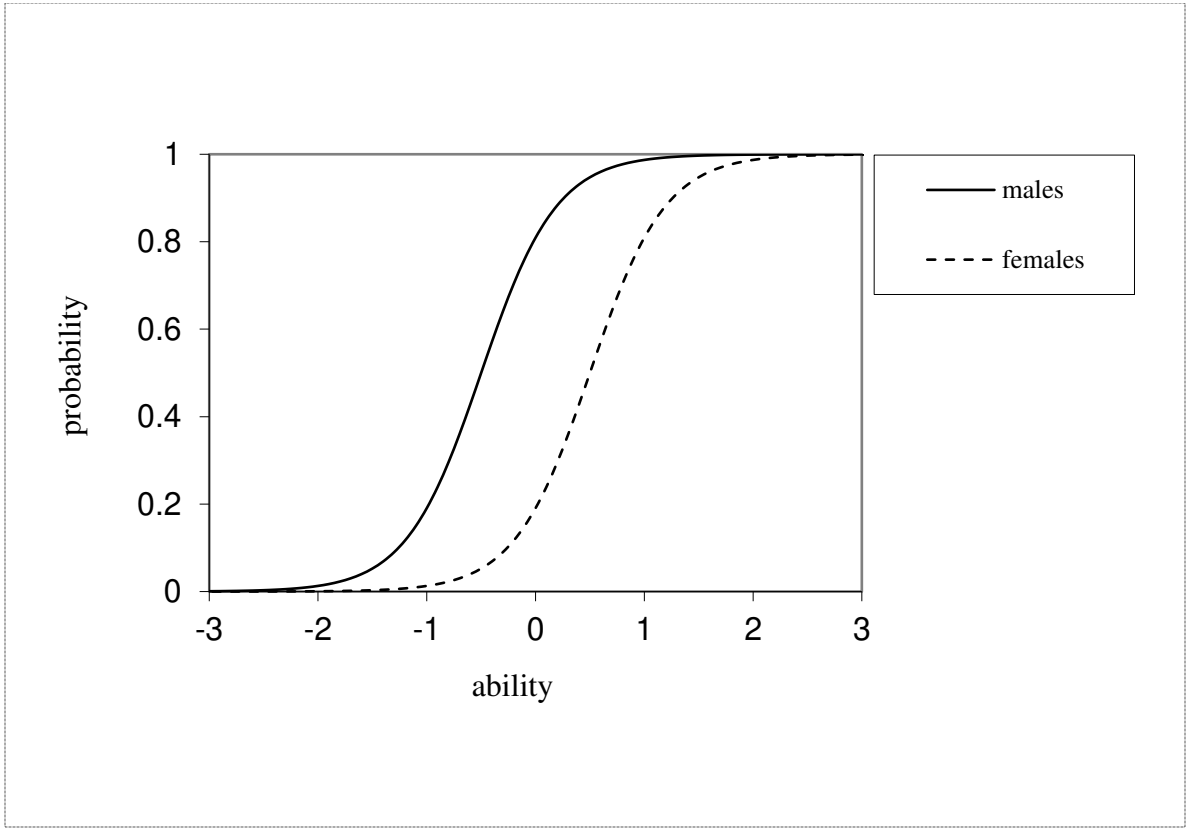


Figure 4. An example of uniform DIF.

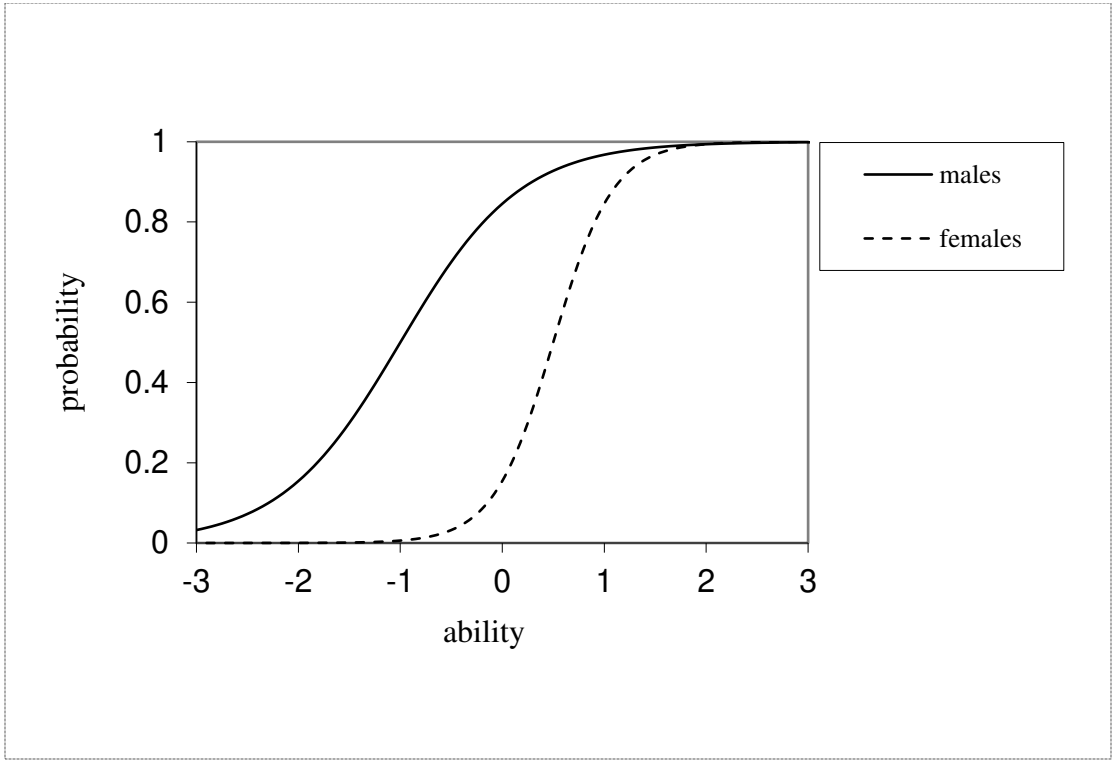


Figure 5. An example of Non-Crossing Non-Uniform DIF.

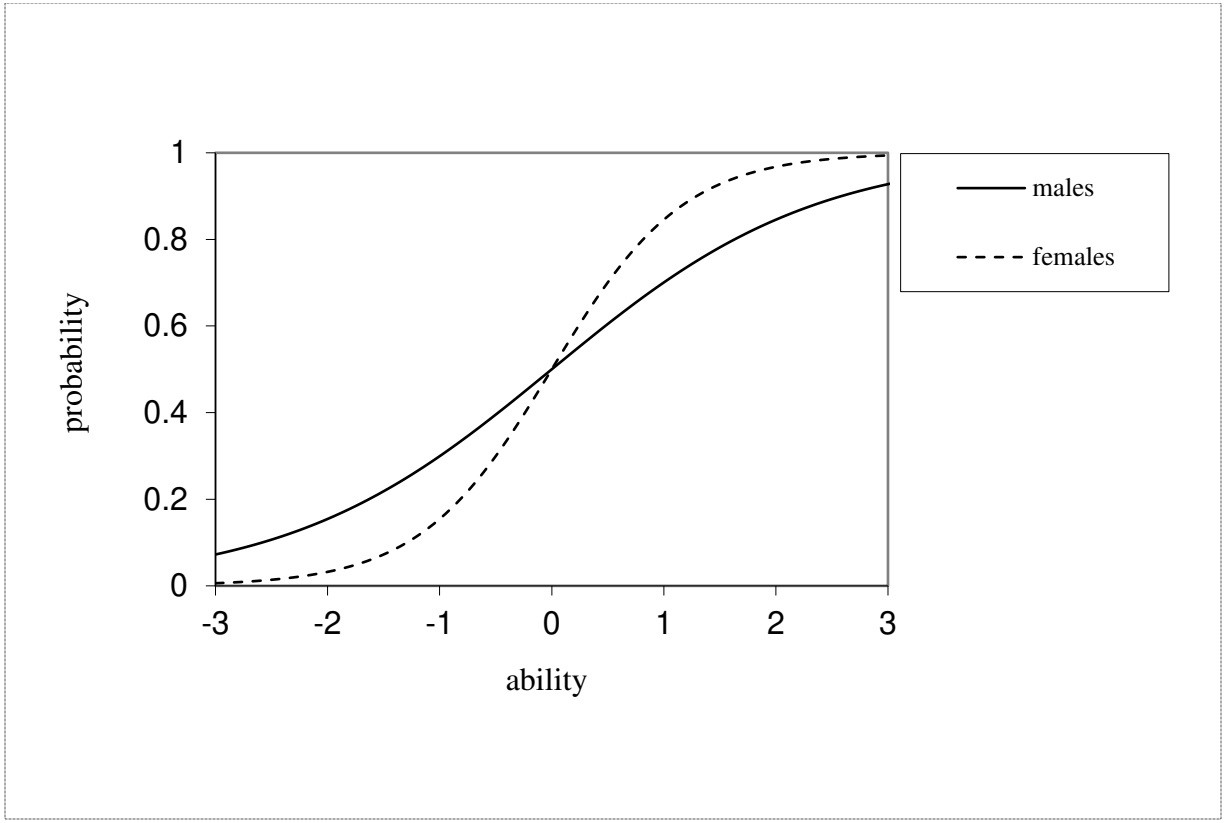


Figure 6. An example of Crossing Non-Uniform DIF.

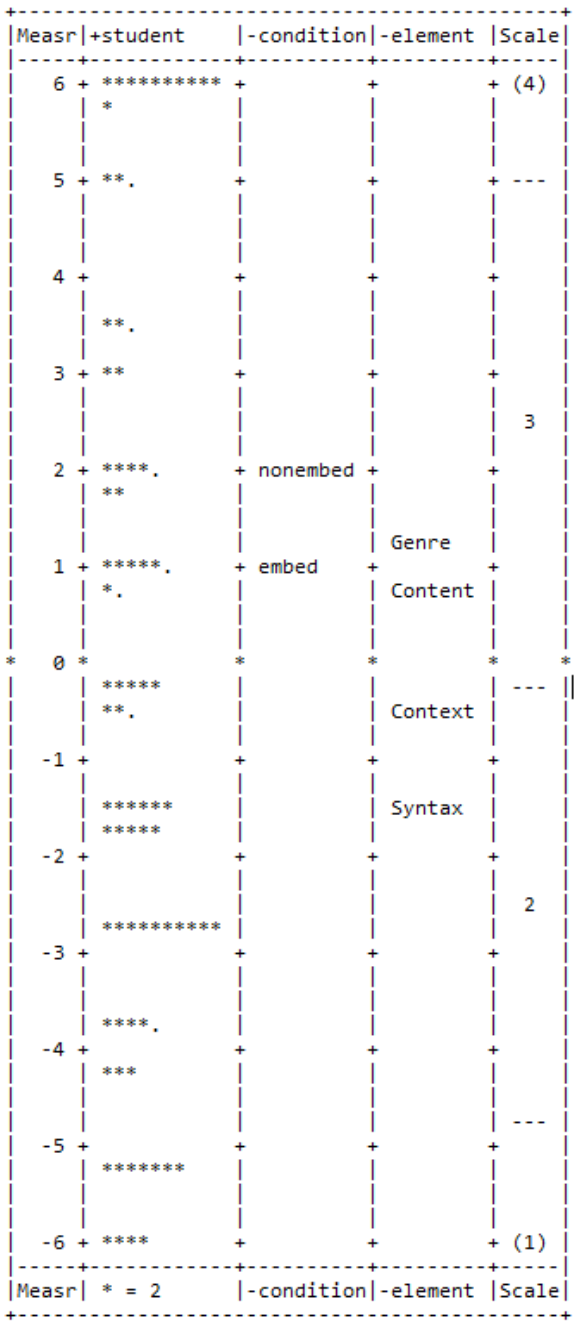


Figure 7. Measure is the corresponding logit to each of the ability and difficulty estimates. Student represents student ability, condition represents the average ability of students in each assessment condition, element represents the difficulty of each rubric element, and the scale represents the where the thresholds are between score categories.

Appendices

Appendix A

WRITTEN COMMUNICATION VALUE RUBRIC

Association of American Colleges and Universities

for more information, please contact value@aacu.org

The VALUE rubrics were developed by teams of faculty experts representing colleges and universities across the United States through a process that examined many existing campus rubrics and related documents for each learning outcome and incorporated additional feedback from faculty. The rubrics articulate fundamental criteria for each learning outcome, with performance descriptors demonstrating progressively more sophisticated levels of attainment. The rubrics are intended for institutional-level use in evaluating and discussing student learning, not for grading. The core expectations articulated in all 15 of the VALUE rubrics can and should be translated into the language of individual campuses, disciplines, and even courses. The utility of the VALUE rubrics is to position learning at all undergraduate levels within a basic framework of expectations such that evidence of learning can by shared nationally through a common dialog and understanding of student success.

Definition

Written communication is the development and expression of ideas in writing. Written communication involves learning to work in many genres and styles. It can involve working with many different writing technologies, and mixing texts, data, and images. Written communication abilities develop through iterative experiences across the curriculum.

Framing Language

This writing rubric is designed for use in a wide variety of educational institutions. The most clear finding to emerge from decades of research on writing assessment is that the best writing assessments are locally determined and sensitive to local context and mission. Users of this rubric should, in the end, consider making adaptations and additions that clearly link the language of the rubric to individual campus contexts.

This rubric focuses assessment on how specific written work samples or collectios of work respond to specific contexts. The central question guiding the rubric is "How well does writing respond to the needs of audience(s) for the work?" In focusing on this question the rubric does not attend to other aspects of writing that are equally important: issues of writing process, writing strategies, writers' fluency with different modes of textual production or publication, or writer's growing engagement with writing and disciplinarity through the process of writing.

Evaluators using this rubric must have information about the assignments or purposes for writing guiding writers' work. Also recommended is including reflective work samples of collections of work that address such questions as: What decisions did the writer make about audience, purpose, and genre as s/he compiled the work in the portfolio? How are those choices evident in the writing -- in the content, organization and structure, reasoning, evidence, mechanical and surface conventions, and citational systems used in the writing? This will enable evaluators to have a clear sense of how writers understand the assignments and take it into consideration as they evaluate

The first section of this rubric addresses the context and purpose for writing. A work sample or collections of work can convey the context and purpose for the writing tasks it showcases by including the writing assignments associated with work samples. But writers may also convey the context and purpose for their writing within the texts. It is important for faculty and institutions to include directions for students about how they should represent their writing contexts and purposes.

Faculty interested in the research on writing assessment that has guided our work here can consult the National Council of Teachers of English/Council of Writing Program Administrators' White Paper on Writing Assessment (2008; www.wpacouncil.org/whitepaper) and the Conference on College Composition and Communication's Writing Assessment: A Position Statement (2008; www.ncte.org/cccc/resources/positions/123784.htm)

Glossary

The definitions that follow were developed to clarify terms and concepts used in this rubric only.

- Content Development: The ways in which the text explores and represents its topic in relation to its audience and purpose.
- Context of and purpose for writing: The context of writing is the situation surrounding a text: who is reading it? who is writing it? Under what circumstances will the text be shared or circulated? What social or political factors might affect how the text is composed or interpreted? The purpose for writing is the writer's intended effect on an audience. Writers might want to persuade or inform; they might want to report or summarize information; they might want to work through complexity or confusion; they might want to argue with other writers, or connect with other writers; they might want to convey urgency or amuse; they might write for themselves or for an assignment or to remember.
- Disciplinary conventions: Formal and informal rules that constitute what is seen generally as appropriate within different academic fields, e.g. introductory strategies, use of passive voice or first person point of view, expectations for thesis or hypothesis, expectations for kinds of evidence and support that are appropriate to the task at hand, use of primary and secondary sources to provide evidence and support arguments and to document critical perspectives on the topic. Writers will incorporate sources according to disciplinary and genre conventions, according to the writer's purpose for the text. Through increasingly sophisticated use of sources, writers develop an ability to differentiate between their own ideas and the ideas of others, credit and build upon work already accomplished in the field or issue they are addressing, and provide meaningful examples to readers.
- Evidence: Source material that is used to extend, in purposeful ways, writers' ideas in a text.
- Genre conventions: Formal and informal rules for particular kinds of texts and/or media that guide formatting, organization, and stylistic choices, e.g. lab reports, academic papers, poetry, webpages, or personal essays.
- Sources: Texts (written, oral, behavioral, visual, or other) that writers draw on as they work for a variety of purposes -- to extend, argue with, develop, define, or shape their ideas, for example

Evaluators are encouraged to assign a zero to any work sample or collection of work that does not meet benchmark (cell one) level performance.

	Capstone 4	**Milestones** 3	 2	**Benchmark** 1
Context of and Purpose for Writing *Includes considerations of audience, purpose, and the circumstances surrounding the writing task(s).*	Demonstrates a thorough understanding of context, audience, and purpose that is responsive to the assigned task(s) and focuses all elements of the work.	Demonstrates adequate consideration of context, audience, and purpose and a clear focus on the assigned task(s) (e.g., the task aligns with audience, purpose, and context).	Demonstrates awareness of context, audience, purpose, and to the assigned tasks(s) (e.g., begins to show awareness of audience's perceptions and assumptions).	Demonstrates minimal attention to context, audience, purpose, and to the assigned tasks(s) (e.g., expectation of instructor or self as audience).
Content Development	Uses appropriate, relevant, and compelling content to illustrate mastery of the subject, conveying the writer's understanding, and shaping the whole work.	Uses appropriate, relevant, and compelling content to explore ideas within the context of the discipline and shape the whole work.	Uses appropriate and relevant content to develop and explore ideas through most of the work.	Uses appropriate and relevant content to develop simple ideas in some parts of the work.
Genre and Disciplinary Conventions *Formal and informal rules inherent in the expectations for writing in particular forms and/or academic fields (please see glossary).*	Demonstrates detailed attention to and successful execution of a wide range of conventions particular to a specific discipline and/or writing task (s) including organization, content, presentation, formatting, and stylistic choices	Demonstrates consistent use of important conventions particular to a specific discipline and/or writing task(s), including organization, content, presentation, and stylistic choices	Follows expectations appropriate to a specific discipline and/or writing task(s) for basic organization, content, and presentation	Attempts to use a consistent system for basic organization and presentation.
Sources and Evidence	Demonstrates skillful use of high-quality, credible, relevant sources to develop ideas that are appropriate for the discipline and genre of the writing	Demonstrates consistent use of credible, relevant sources to support ideas that are situated within the discipline and genre of the writing.	Demonstrates an attempt to use credible and/or relevant sources to support ideas that are appropriate for the discipline and genre of the writing.	Demonstrates an attempt to use sources to support ideas in the writing.

Control of Syntax and Mechanics	Uses graceful language that skillfully communicates meaning to readers with clarity and fluency, and is virtually error-free.	Uses straightforward language that generally conveys meaning to readers. The language in the portfolio has few errors.	Uses language that generally conveys meaning to readers with clarity, although writing may include some errors.	Uses language that sometimes impedes meaning because of errors in usage.

Appendix B

Non-Embedded Student Assignment

You have sixty minutes to write a letter to the editor that would be appropriate to appear in the James Madison University student newspaper, *The Breeze*, on the topic below. Letters to the editor express an opinion on a community matter. While your document should be no less than 250 words in length, it should be as long as necessary to get your idea across clearly and concisely. The supplemental materials on the following pages contain information that you may want to read prior to drafting your letter to the editor. You are encouraged to include in your letter any of the information contained in the provided readings that helps you to make your points. As you would do with any writing project, please be sure to review and modify your first draft throughout the session.

The topic for your letter to the editor will be: "Should chronological age (16, 18, 21) be the criterion by which adult responsibilities are granted?"

This assignment is designed to assess your ability to articulate and support complex ideas in writing. Keep in mind your intended audience (readers of *The Breeze*) and try your best to effectively convey your ideas. In evaluating your writing, we will consider your purpose, organization, complexity of ideas, style, and usage and mechanics.

- Features of purpose may include your thesis or central idea, topic selection, relevance, clarity, and focus.
- Features of organization may include the appropriateness of format, balance and ordering of ideas, flow, and transitions.
- Features of complexity may include your reasoning, evidence, detail, development, creativity, originality, and perspective.
- Features of style may include the tone, sentence length and structure, phrasing, and word choice of your letter.
- Finally, features of usage and mechanics may include your clarity, sentence structure, grammar, spelling, punctuation, and capitalization.

Please give this activity your best effort. We are interested in what you have to say as well as how you say it.

Appendix C

Student Opinion Scale

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below.

1. Doing well on these tests was important to me.
2. I engaged in good effort throughout these tests.
3. I am not curious about how I did on these tests relative to others.
4. I am not concerned about the scores I receive on these tests.
5. These were important tests to me.
6. I gave my best effort on these tests.
7. While taking these tests, I could have worked harder on them.
8. I would like to know how well I did on these tests.
9. I did not give these tests my full attention while completing them.
10. While taking these tests, I was able to persist to completion of the tasks.

Appendix D

FACETS Syntax

```
title = Thesis MSC Written Communication
facets = 3 ; three facets are id, condition, element
Delements = LN
noncenter = 2 ; id and element are centered at 0, condition is allowed to float
positive = 1 ; only for id does greater score mean greater measure
models = ; :
 ?,?B,?B,R4 ; where the B's represent the interaction between condition and element
*
labels=
1,student
1-157
*
2,condition
1 = nonembed
2 = embed
*
3,element
1= Context
2= Content
3= Genre
4= Syntax
*
```

Appendix E

SAS Syntax

```
proc logistic data = spss_data;
model context = xverbc d1
/link = alogit;
run;

proc logistic data = spss_data;
model context = xverbc d1
/link = alogit unequalslopes;
run;

proc logistic data = spss_data;
model content = xverbc d1
/link = alogit;
run;

proc logistic data = spss_data;
model content = xverbc d1
/link = alogit unequalslopes;
run;

proc logistic data = spss_data;
model genre = xverbc d1
/link = alogit;
run;

proc logistic data = spss_data;
model genre = xverbc d1
/link = alogit unequalslopes;
run;

proc logistic data = spss_data;
model syntax = xverbc d1
/link = alogit;
run;

proc logistic data = spss_data;
model syntax = xverbc d1
/link = alogit unequalslopes;
run;
```

Appendix F

Table 1. *Score frequencies between assessment conditions for the 'Context and Purpose of Writing' Rubric Element.*

Assessment Condition	Score Category				
	1	2	3	4	Total
Non-embedded	15	23	10	2	50
Embedded	22	36	31	18	107
Total	37	59	41	20	157

Table 2. *Score frequencies between assessment conditions for the 'Content Development' rubric element.*

Assessment Condition	Score Category				
	1	2	3	4	Total
Non-embedded	32	14	4	0	50
Embedded	25	40	23	19	107
Total	57	54	27	19	157

Table 3. *Score frequencies between assessment conditions for the 'Genre and Disciplinary Conventions' rubric element.*

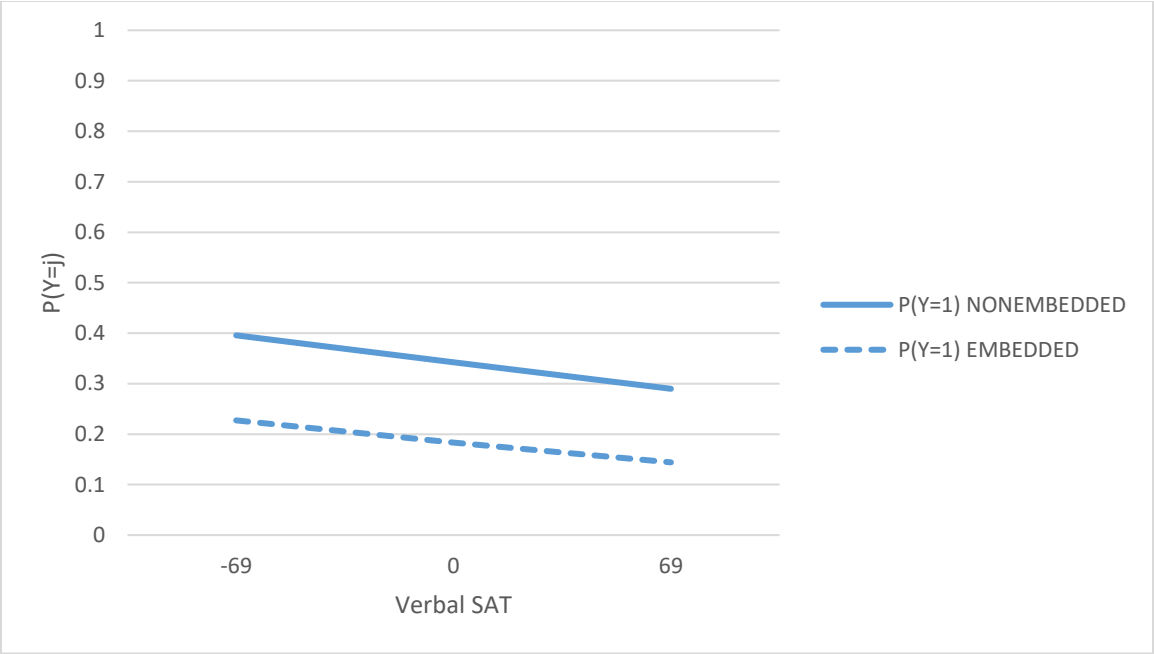
Assessment Condition	Score Category				
	1	2	3	4	Total
Non-embedded	26	19	5	0	50
Embedded	30	42	24	11	107
Total	56	61	29	11	157

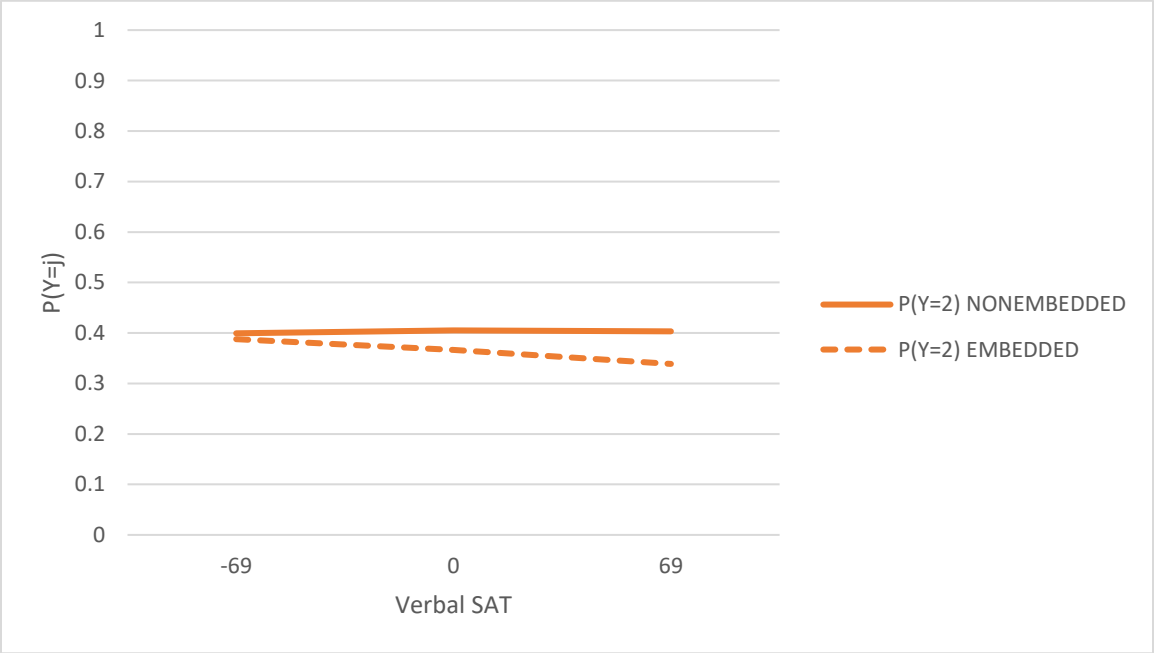
Table 4. *Score frequencies between assessment conditions for the 'Control of Syntax and Mechanics' rubric element.*

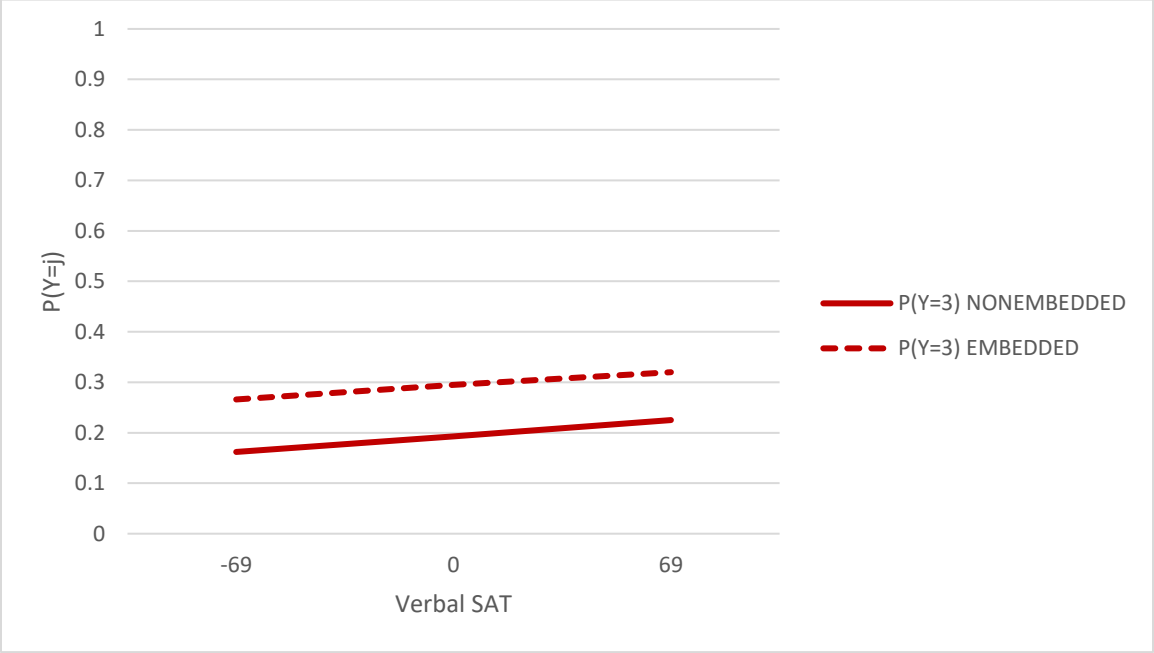
Assessment Condition	Score Category				
	1	2	3	4	Total
Non-embedded	7	26	16	1	50
Embedded	8	42	40	17	107
Total	15	68	56	18	157

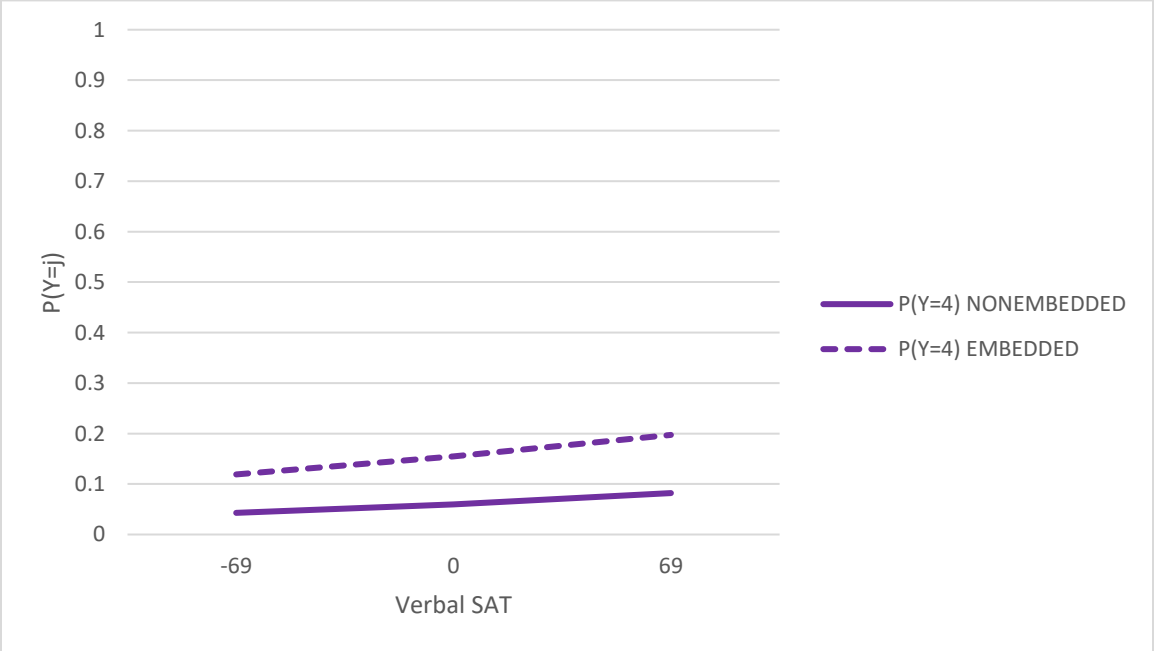
Appendix G

‘Context of and Purpose for Writing’ rubric element, P(Y=1) to P(Y=4)



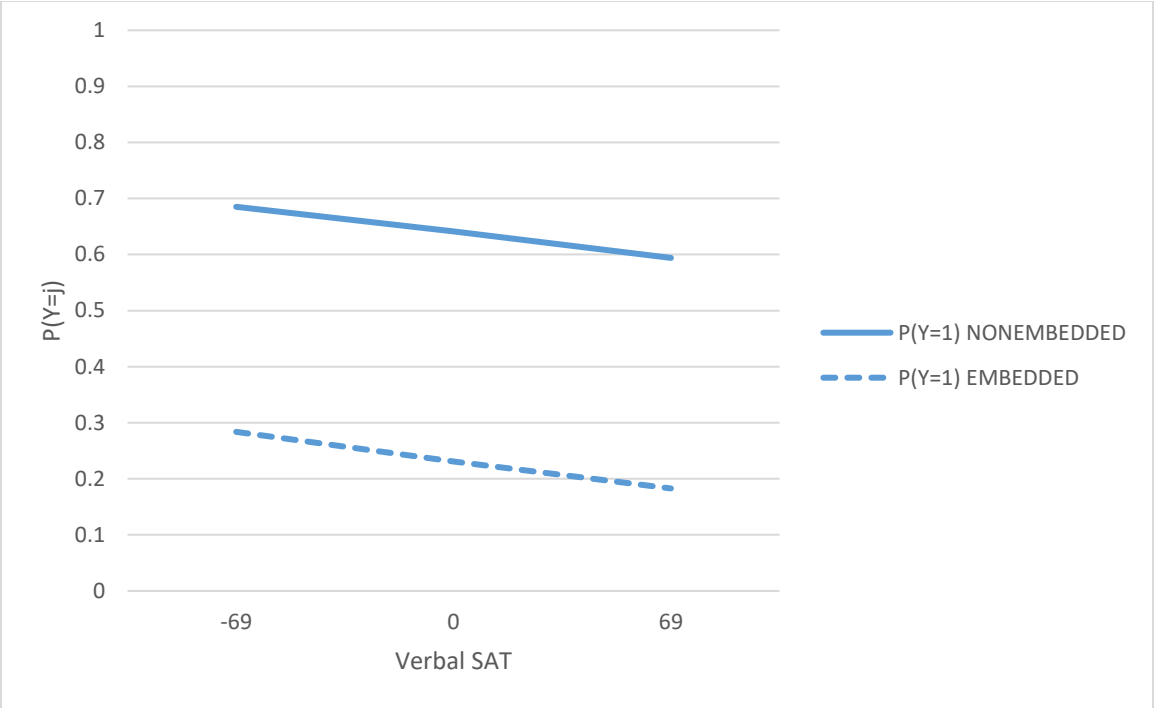


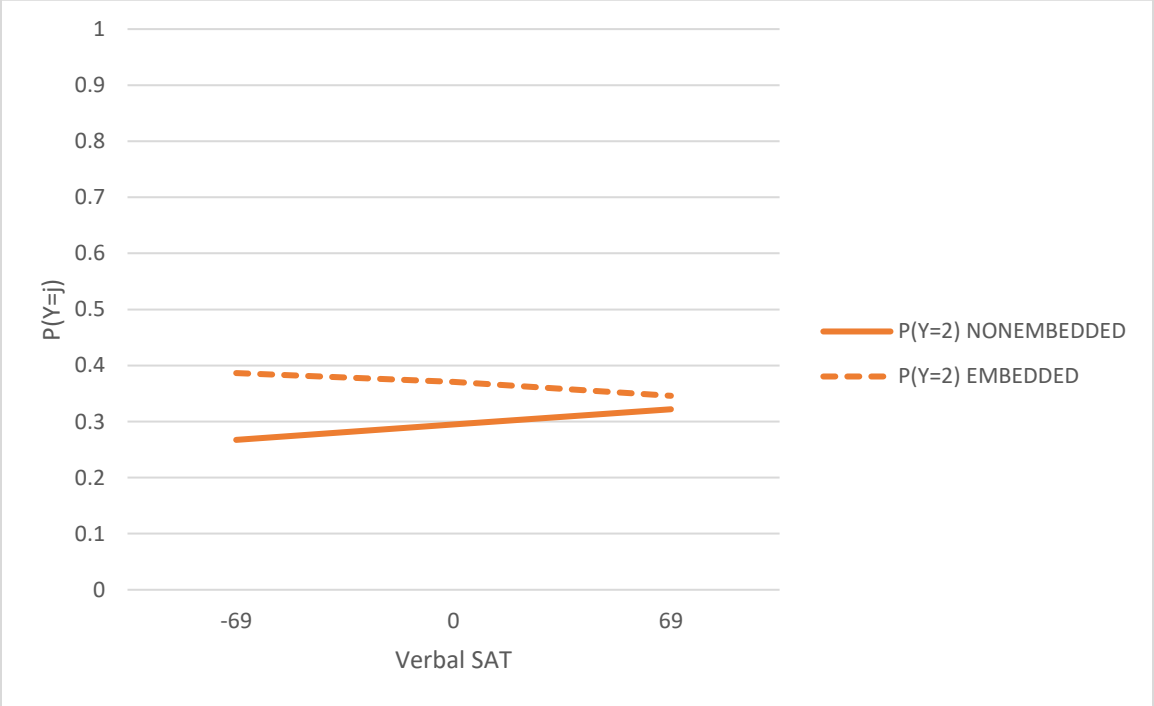

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=3) NONEMBEDDED
P(Y=3) EMBEDDED

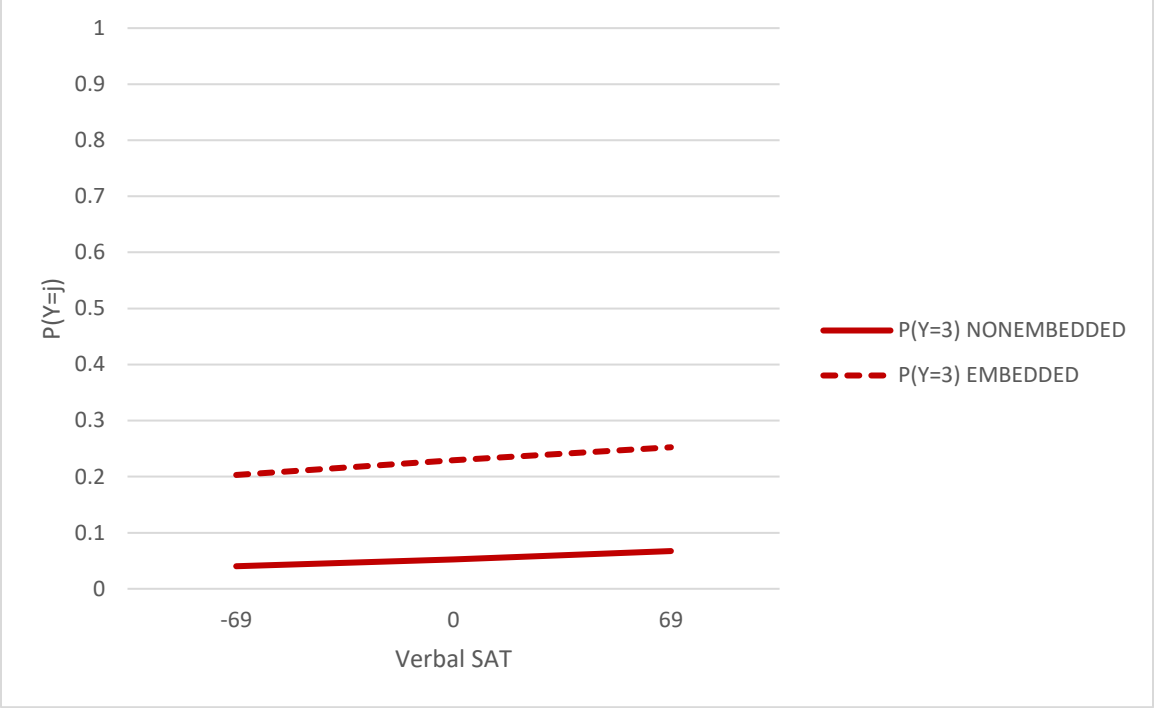

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=4) NONEMBEDDED
P(Y=4) EMBEDDED

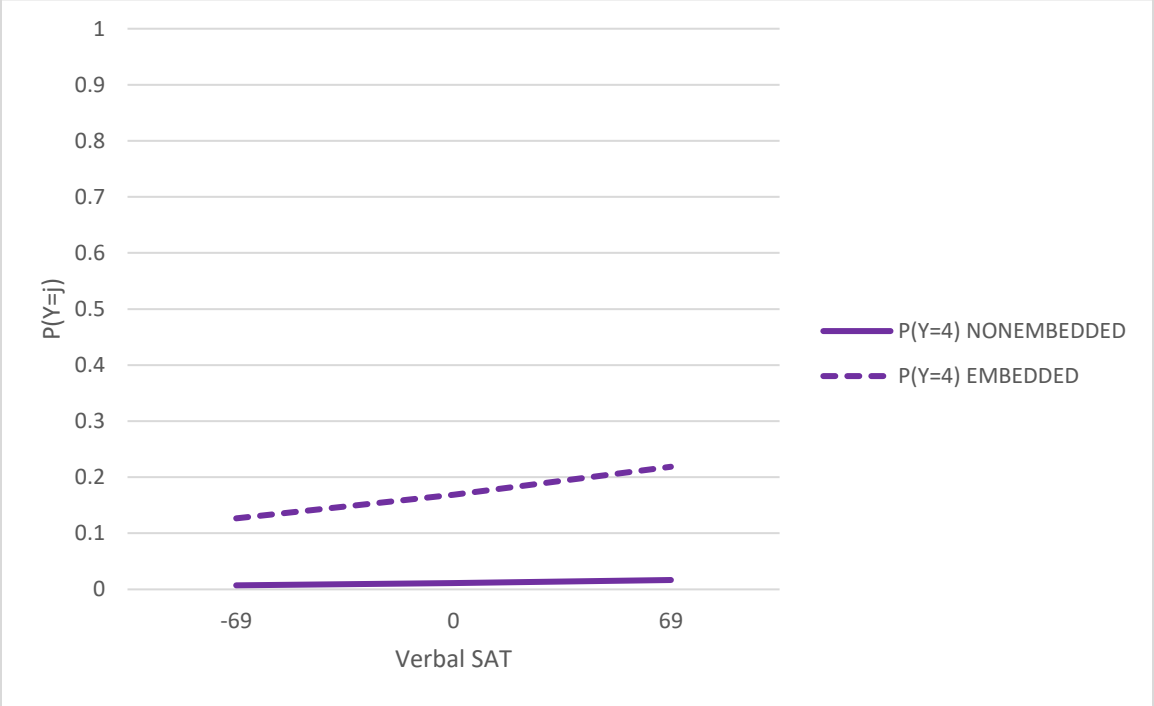
Appendix H

‘Content Development’ rubric element, P(Y=1) to P(Y=4)



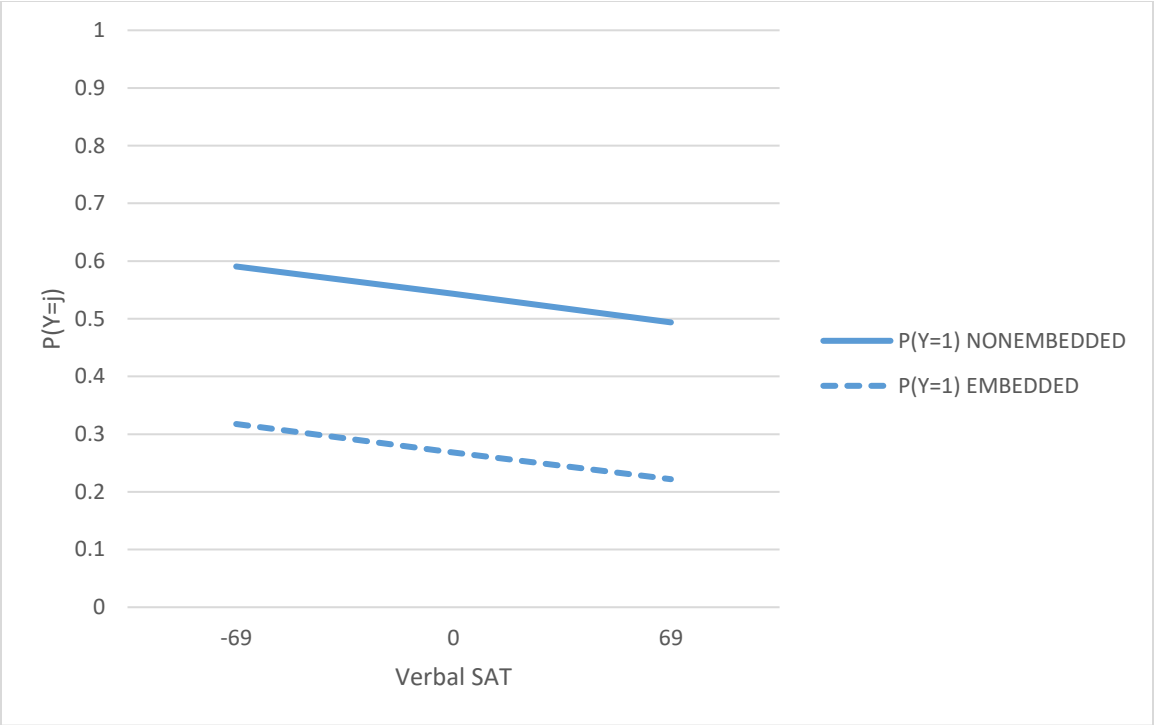


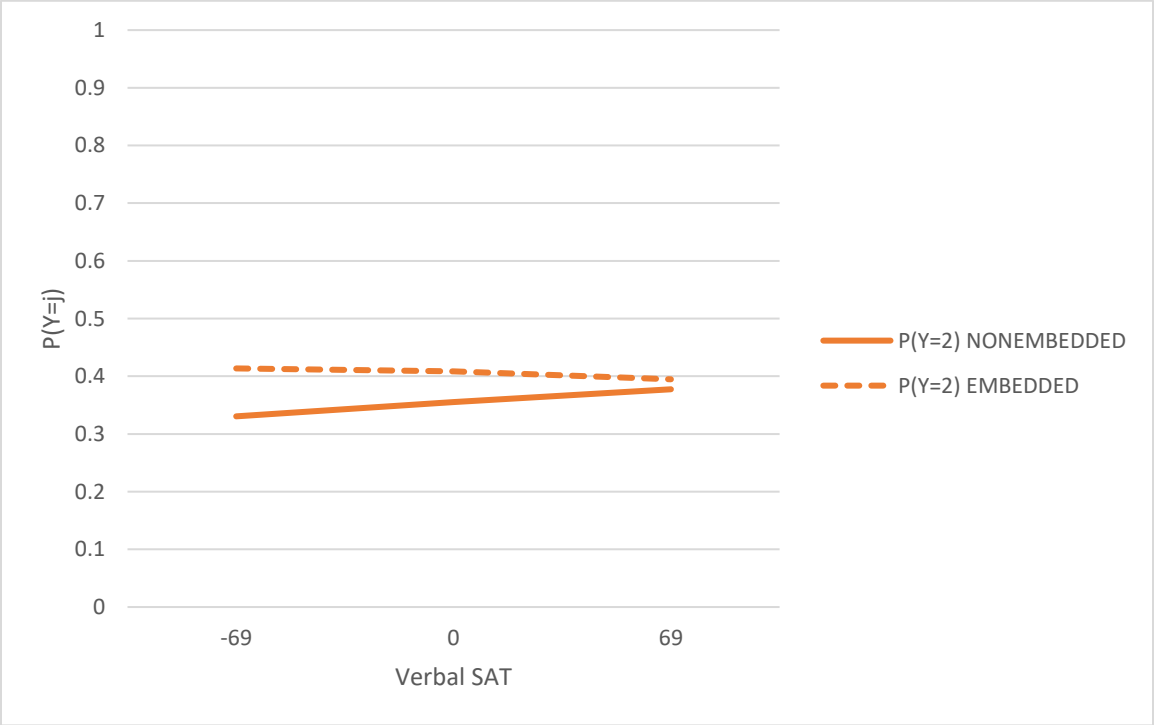

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=3) NONEMBEDDED
P(Y=3) EMBEDDED

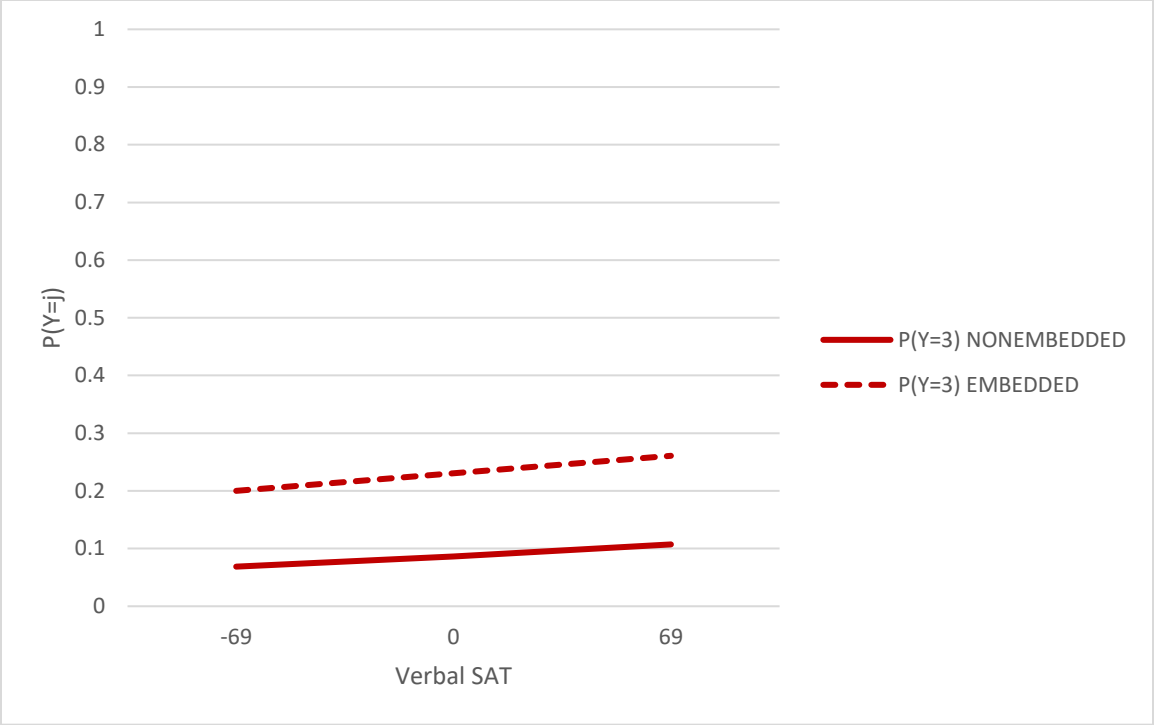

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=4) NONEMBEDDED
P(Y=4) EMBEDDED

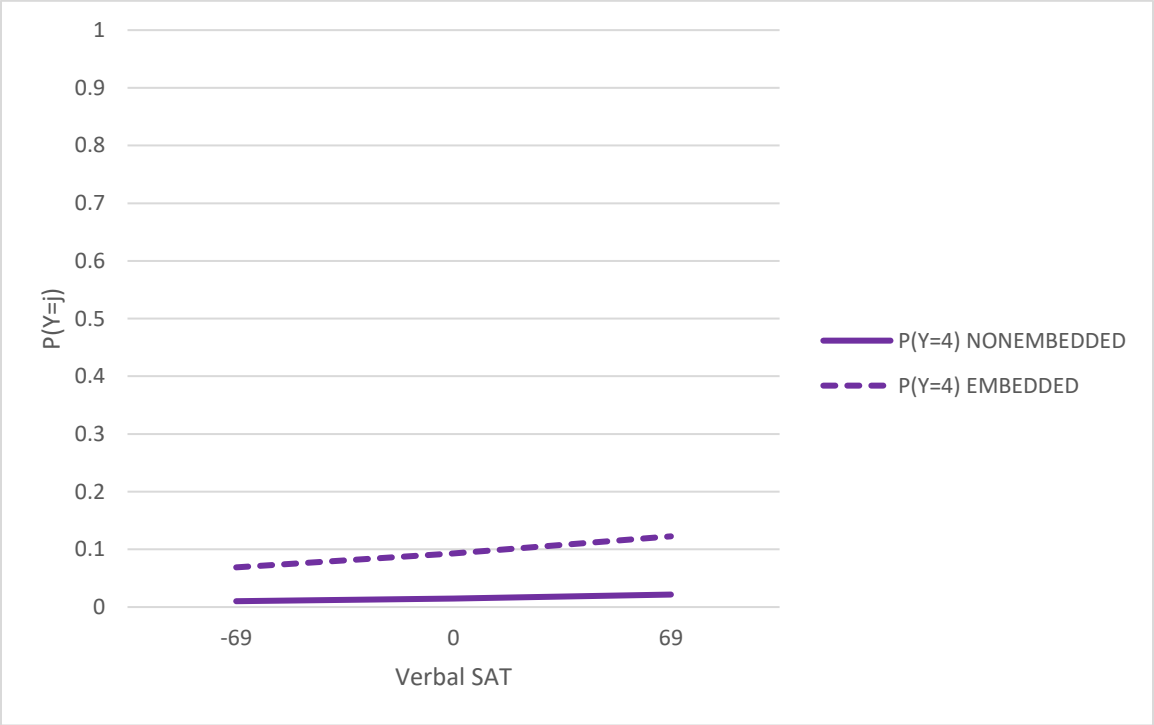
Appendix I

‘Genre and Disciplinary Conventions’ rubric element, P(Y=1) to P(Y=4)



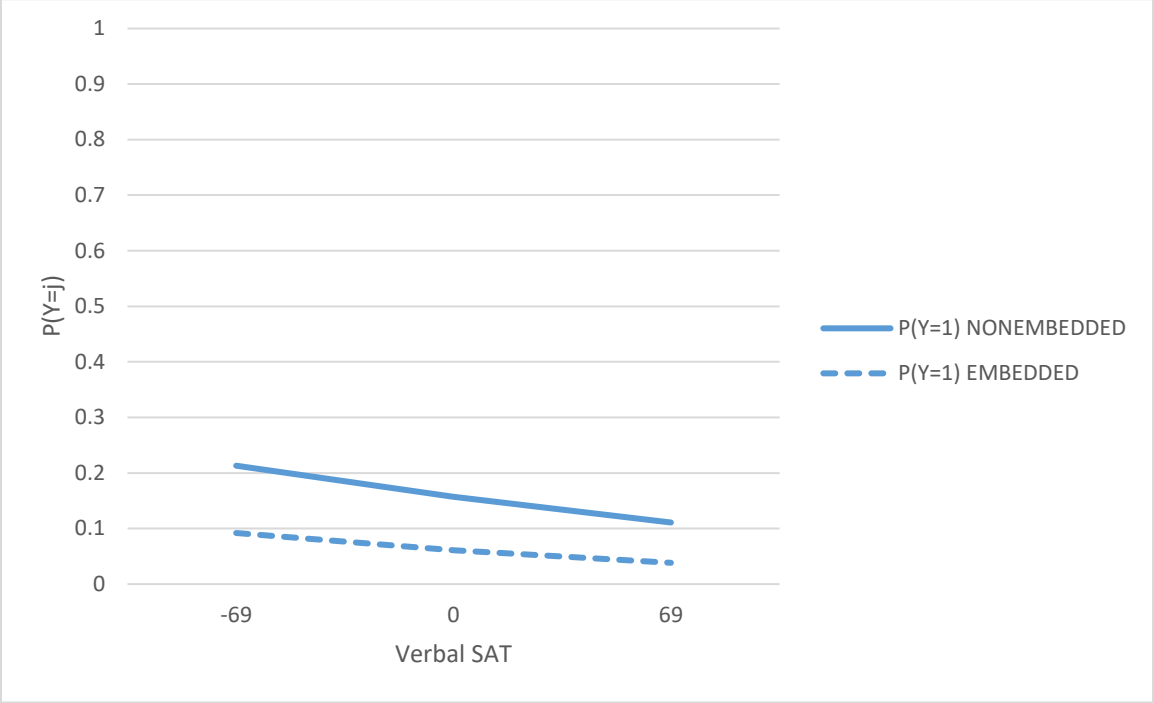


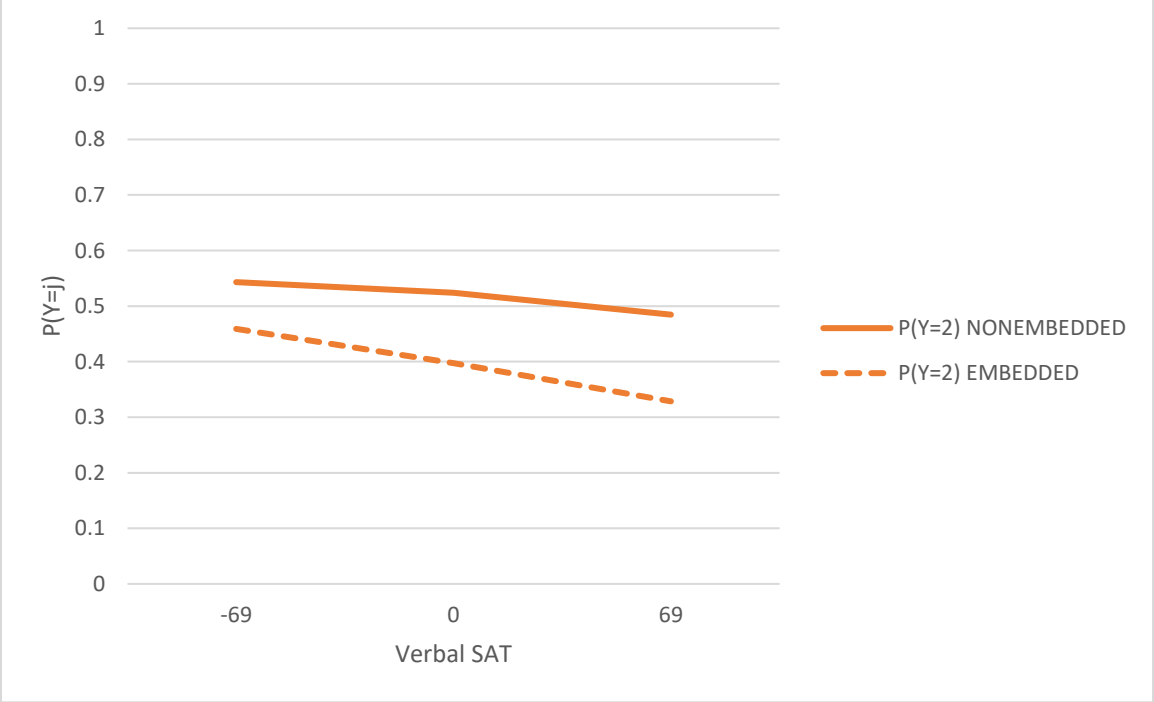

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=3) NONEMBEDDED
P(Y=3) EMBEDDED

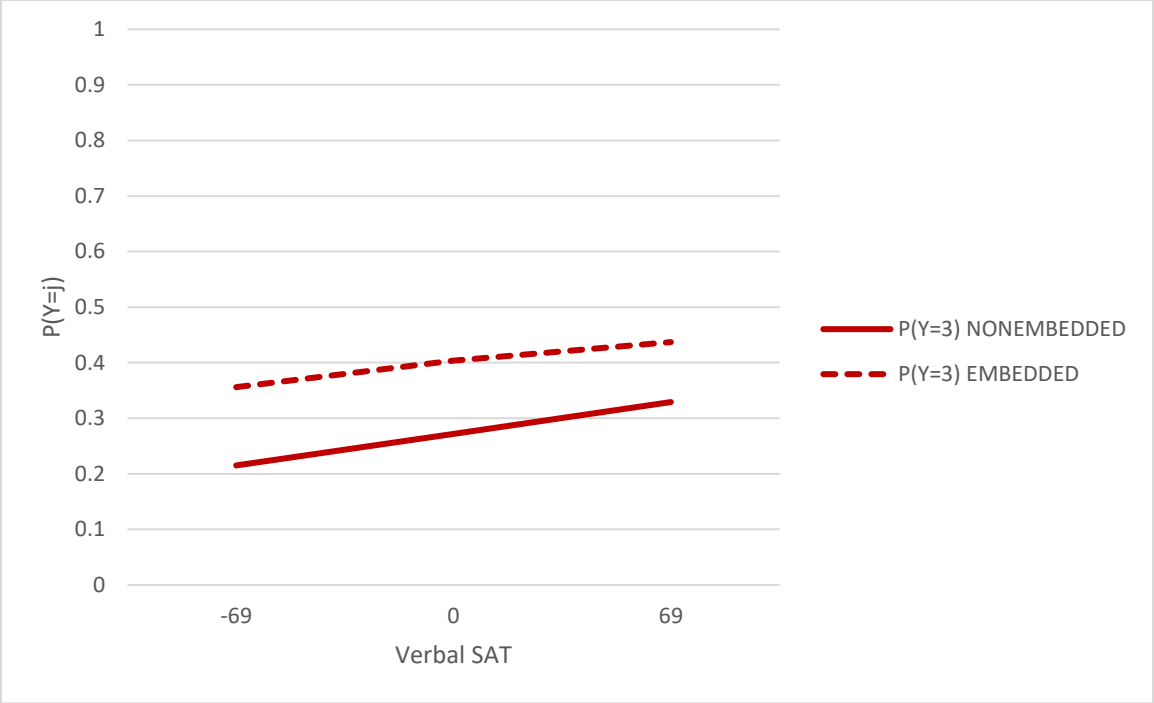

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=4) NONEMBEDDED
P(Y=4) EMBEDDED

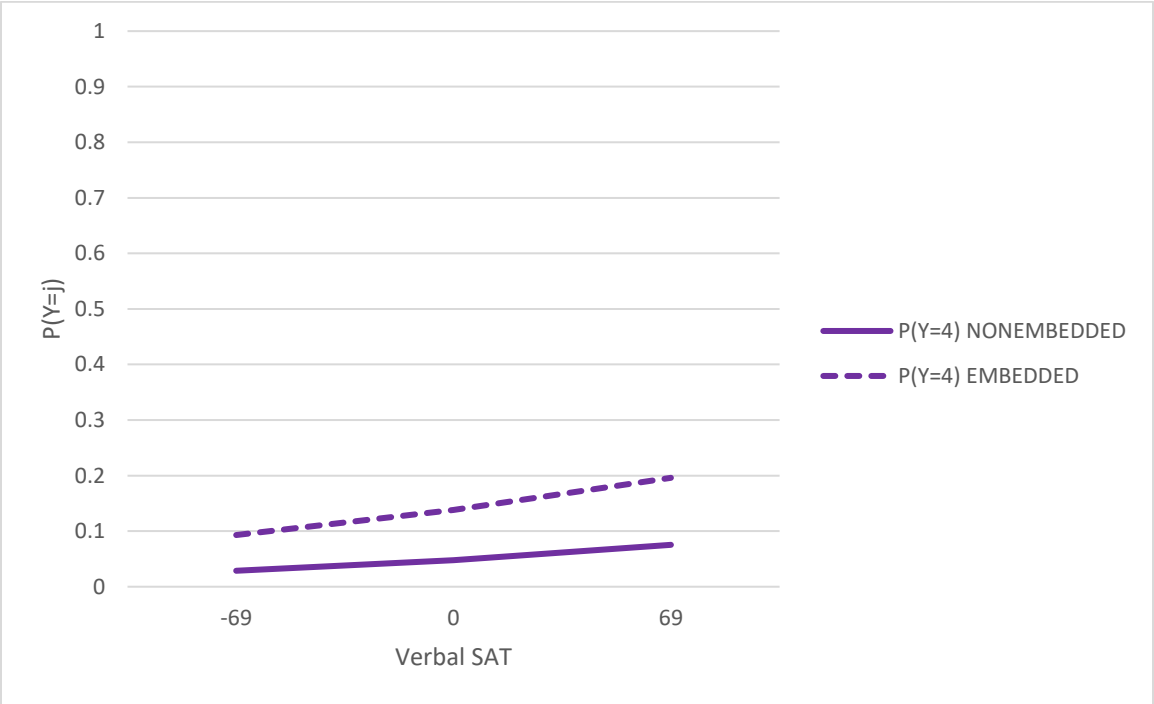
Appendix J

‘Control of Syntax and Mechanics’ rubric element, P(Y=1) to P(Y=4)






1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=3) NONEMBEDDED
P(Y=3) EMBEDDED


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
P(Y=j)
-69
0
69
Verbal SAT
P(Y=4) NONEMBEDDED
P(Y=4) EMBEDDED

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14,* 219-234.

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from

a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Adelman, C., Ewell, P., Gaston, P., & Schneider, C. G. (2011). *The degree qualifications profile*. Indianapolis, IN: Lumina Foundation.

American College Test. (2015). CAAP: Guide to successful general education outcomes assessment. Iowa City, IA: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Andrich, D. (1998). Thresholds, steps and rating scale conceptualization. *Rasch Measurement Transactions*, *12*(3), 648-649.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology, in *Differential Item Functioning*, P. W. Holland & H. Wainer, eds, Lawrence Erlbaum Associates, Hillsdale, pp. 3–23.

Association for American Colleges and Universities. (2017). On Solid Ground. Washington, DC: Author.

Association for American Colleges and Universities. (2011). The LEAP vision for learning: Outcomes, practices, impact, and employers' view. Washington, DC: Author.

Association of American Colleges and Universities. (2009). *Written Communication VALUE rubric. Retrieved from https://www.aacu.org/value/rubrics/written-communication*

Association for American Colleges and Universities. (2007). College learning for the new Global Century. Washington, DC: Author.

Banta, T. (2008). Trying to clothe the emperor. *Assessment Update, 20,* 3-4, 16-17

Banta, T. W., & Blaich, C. (2010). Closing the assessment loop. *Change: The Magazine of Higher Learning*, *43*(1), 22-27.

Barry, C. L., & Finney, S. J. (2009). Does It Matter How Data Are Collected? A Comparison of Testing Conditions and the Implications for Validity. *Research & Practice in Assessment*, *4*, 17-26.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342-363.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441.

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., & Rumble, M. (2010). Defining 21st century skills. In P. Griffin, B. McGaw, & E.

Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). New York, NY: Springer Science and Business Media B

Biola, H. R. (1982). Time limits and topic assignments for essay tests. *Research in the Teaching of English*, 16, 97-98.

Blaich, C., & Wise, K. (2011). From gathering to using assessment results. *National Institute for Learning Outcomes assessment*.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Bouwer, R., Béguin, A., Sanders, T., & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100.

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An Experimental Study of the Effects of Monetary Incentives on Performance on the 12th-Grade NAEP Reading Assessment. *Teachers College Record*, *113*(11), 2309-2344.

Brennan, R. L., Gao, X., & Colton, D. A. (1995). Generalizability analyses of Work Keys listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157-176.

Brentani, E., & Golia, S. (2007). Unidimensionality in the Rasch model: how to detect and interpret. *Statistica*, *67*(3), 253-261.

Brown, G. T. (2010). The validity of examination essays in higher education: Issues and responses. *Higher Education Quarterly*, *64*(3), 276-291.

Burke, P. (1991). *You can lead adolescents to a test but you can't make them try*. Final report (Contract No. OTA – H3-6110.0). Washington, DC: Office of Technology Assessment. (ERIC Document Reproduction Service No. ED 378 221).

Calfee, R. C., Miller, R. G., Graham, S., MacArthur, C., & Fitzgerald, J. (2007). Best practices in writing assessment. *Best practices in writing instruction*, 265-286.

Casner-Lotto, J., & Barrington, L. (2006). Are they really ready to work? Washington, DC: Partnership for 21st Century Skills.

Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. *ELT Journal, 44,* 122-131.

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT™ integrated writing tasks. *Language Testing*, *30*(4), 513-534.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing writing*, *8*(3), 165-191.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: issues and practice*, *17*(1), 31-44.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269-279.

Coates, H., & Seifert, T. (2011). Linking assessment for learning, improvement and accountability. *Quality in Higher Education*, *17*(2), 179-194.

College Board (2018a). Reading Test. *Retrieved April 2018 from https://collegereadiness.collegeboard.org/sat/inside-the-test/reading.*

College Board (2018b). SAT Suite of Assessments: Score Structure. *Retrieved April 2018 from https://sat-edit.collegeboard.org/educators/higher-ed/scoring/score-structure.*

College Board (2018c). Writing and Language Test. *Retrieved April 2018 from https://collegereadiness.collegeboard.org/sat/inside-the-test/writing-language.*

College Board. (2010). *Trends in college pricing*. New York: Author.

Cooper, P. L. 1984: The assessment of writing ability: a review of research. Princeton, NJ: Educational Testing Service. GRE Board Research Report GREB No. 82-15R/ETS Research Report 84-12.

Council for Aid to Education. (2017a). Assessment design and sample. *Retrieved August 2017 from http://cae.org/flagship-assessments-cla-cwra/cla/assessment-design-and-sample/.*

Council for Aid to Education. (2017b). CLA+ Scoring Rubric. *Retrieved August 2017 from http://cae.org/images/uploads/pdf/CLA_Plus_Scoring_Rubric.pdf*

Council of Writing Program Administrators, National Council of Teachers of English, & National Writing Project. (2011). *Framework for success in postsecondary writing*. Retrieved from http://wpacouncil.org/files/framework-for-success-postsecondary-writing.pdf

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Crowhurst, M. (1980). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English, 14*, 223-231.

de Ayala, R. J. (2009). The Theory and Practice of Item Response Theory. New York: Guilford Press.

DeMars, C. E. (2010). *Item response theory*. Oxford University Press.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.

Duckworth, A. L., Quinn, P. D., Lynam, D. R., Loeber, R., & Stouthamer-Loeber, M. (2011). Role of test motivation in intelligence testing. *Proceedings of the National Academy of Sciences*, *108*(19), 7716-7720.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*, 1-52.

Educational Testing Service. (2017). ETS Proficiency Profile Optional Essay. *Retrieved July 2017 from https://www.ets.org/proficiencyprofile/about/essay/*

Educational Testing Service. (2010). *ETS proficiency profile user's guide*. Princeton, NJ: Author.

Engelhard, Jr., G., & Wind, S.A. (2013). Rating quality studies using rasch measurement theory. *College Board Research Report,* 3.

Engelhard Jr, G. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*(3), 171-191.

Finney, S. J., Sundre, D. L., Swain, M. S., & Williams, L. M. (2016). The validity of value-added estimates from low-stakes testing contexts: The impact of change in test-taking motivation and test consequences. *Educational Assessment*, *21*(1), 60-87.

Fischer, G. H. (2007). Rasch models. In C. R. Rao & S. Sinharay (Eds.), Psychometrics (Handbook of statistics, Vol. 26, pp. 515–585). Amsterdam: Elsevier.

Fulcher, K. H., Good, M. R., Coleman, C. M., & Smith, K. L. (2014). A simple model for learning improvement: Weigh pig, feed pig, weigh pig. *NILOA Occasional Paper*, (23).

Frederiksen, N., Mislevy, R. J., & Bejar, I. I. (Eds.). (2012). *Test theory for a new generation of tests*. Routledge.

Godshalk, F. I., & Swineford, F. (1966). Coffman. *WEThe measurement of writing ability*.

Gulikers, J. T., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational technology research and development*, *52*(3), 67-86.

Hale, G. A. (1992). Effects of amount of time allowed on the Test of Written English. *ETS Research Report Series*, *1992*(1).

Hart Research Associates. (2010). Raising the bar: Employers' views on college learning in the wake of the economic downturn. Washington, DC: Author.

Hathcoat, J.D., & Gregg, N. (2018). Absence of evidence is not evidence of absence: The meaning of zero in the Multi-State Collaborative. Manuscript submitted for publication.

Hathcoat, J. D., Penn, J. D., Barnes, L. L., & Comer, J. C. (2016). A second dystopia in education: Validity issues in authentic assessment practices. *Research in Higher Education*, *57*(7), 892-912.

Hillocks, G. (2002). *The testing trap: How state writing assessments control learning*. Teachers College Press.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *Test validity*, 129-145.

Hull, G. (1987). The editing process in writing: A performance study of more skilled and less skilled college writers. *Research in the Teaching of English*, 8-29.

Huot, B. 1990: The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60*, 237-263.

Johnstone, K. M., Ashbaugh, H., & Warfield, T. D. (2002). Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of educational psychology*, *94*(2), 305.

Kellogg, R. T., & Raulerson III, B. A. (2007). Improving the writing skills of college students. *Psychonomic bulletin & review*, *14*(2), 237-242.

Kellogg, R. T., & Whiteford, A. P. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*(4), 250-266.Klein, S. C., Liu, O. L. E., Sconing, J. A., Bolus, R. C., Bridgeman, B. E., Kugelmass, H. C., Nemeth, A., Robbins, S., & Steedle, J. C. (2009). *Test Validity Study (TVS) Report*. Supported by the Fund for the Improvement of Postsecondary Education (FIPSE).

Kuh, G. D., Jankowski, N., Ikenberry, S. O., & Kinzie, J. L. (2014). *Knowing what students know and can do: The current state of student learning outcomes assessment in US colleges and universities*. Urbana, IL: National Institute for Learning Outcomes Assessment.

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*(1), 71-92.

Lane, S., & Stone, C. (2006). Performance Assessment. In R.L. Brennan. *Educational Measurement, Fourth Edition* (387-431). Portsmouth: Greenwood Publishing Group, Inc.

Lee, H., & Geisinger, K. F. (2016). The Matching Criterion Purification for Differential Item Functioning Analyses in a Large-Scale Assessment. *Educational and Psychological Measurement*, *76*(1), 141-163.

Lim, G. S. (2010). Investigating prompt effects in writing performance assessment. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 8, 95-115.

Linacre, J. M. (2017a). FACETS® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2017b). Dimensionality: contrasts & variances. *Retrieved October 2017 from http://winsteps.com/winman/principalcomponents.htm.*

Linacre, J. M. (2012a). Winsteps Rasch Tutorial 2. *Retrieved from the Winsteps website: http://www. winsteps com/tutorials.htm*.

Linacre, J. M. (2012b). Winsteps Rasch Tutorial 2. *Retrieved from the Winsteps website: http://www. winsteps com/tutorials.htm*.

Linacre, J. M. (2008). Facets Rasch model computer program [Software manual]. Chicago: Winsteps.com.

Linacre, J. M. (2006). Data variance explained by Rasch measures. *Rasch Measurement Transactions*, *20*(1), 1045.

Liu, O.L. (2011). Outcomes assessment in higher education: Challenges and future research in the context of voluntary system of accountability. *Educational Measurement: Issues and Practice, 30*, 2-9.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*(9), 352-362.

Livingston, S. A. (1987, April). The effects of time limits on the quality of student-written essays. In *Annual Meeting of the American Educational Research Association, New York, NY*.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lorenzo-Seva, U., & Ferrando, P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38*(1), 88-91.

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A counterexample with Angoff's delta plot. *Educational and Psychological Measurement*, *73*(2), 293-311.

ManpowerGroup. (2012). *Talent shortage survey: Research results*. Retrieved from http://www.manpowergroup.us/campaigns/talent-shortage-2012/pdf/2012_Talent_Shortage_Survey_Results_US_FINALFINAL.pdf

Markle, R., Brenneman, M., Jackson, T., Burrus, J., & Robbins, S. (2013). Synthesizing frameworks of higher education student learning outcomes. *ETS Research Report Series*, *2013*(2).

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174.

McBee, M. M., & Barnes, L. L. B. (1998). The generalizability of a performance assessment measuring achievement in eighth-grade mathematics. *Applied Measurement in Education*, *11*, 179–194.

McPherson, P. & Schulenberg, D. (2006). *Toward a voluntary system of accountability program (VSA) for public universities and colleges*. Washingtion, DC: National Association of State Universities and Land-Grant Colleges.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of educational statistics*, *7*(2), 105-118.

Messick, S. (1995). Standards of validity and the validity of standards in performance asessment. *Educational measurement: Issues and practice*, *14*(4), 5-8.

Miller, M.A. (2012). Demonstrating and improving student learning: The role of standardized tests. *National Institute for Learning Outcomes Assessment: The Seven Red Herrings about Standardized Assessments in Higher Education [Occasional Paper #15]*.

Multi-State Collaborative (2017). MSC: A Multi-State Collaborative to Advance Learning Outcomes Assessment. *Retrieved August 2017 from http://www.sheeo.org/projects/msc-multi-state-collaborative-advance-learning-outcomes-assessment*Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, *1992*(1).

Myers, N. D., Wolfe, E. W., Feltz, D. L., & Penfield, R. D. (2006). Identifying differential item functioning of rating scale items with the Rasch model: An

introduction and an application. *Measurement in Physical Education and Exercise Science*, *10*(4), 215-240.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of dif. *European Journal of Psychological Assessment*, *18*(1), 9.

O'Neill, P., & Murphy, S. (2012). Postsecondary writing assessment. In Secolsky,C., & Denison, B.D. (Ed.), *Handbook on measurement, assessment, and evaluation in higher education*. New York, NY: Routledge.

Palmer, O. (1966). Sense or nonsense? The objective testing of English composition. In C.I. Chase & H.G. Ludlow, (Eds.). *Readings in Educational and Psychological Measurement* (pp. 284–291). Palo Alto: Houghton.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, *33*(1), 36-48.

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38-49.

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*(3), 187-210.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In C.R. Rao & S. Sinharay (Eds.), Handbook of Statistics, Vol. 26: Psychometrics (p. 125-167). Amsterdam: Elsevier.

Penfield, R. D., & Algina, J. (2003). Applying the Liu-Agresti Estimator of the Cumulative Common Odds Ratio to DIF Detection in Polytomous Items. *Journal of Educational Measurement*, *40*(4), 353-370.

Penfield, R.D., & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations.

Polio, C, Fleck, C, & Ledger, N. (1998). If I only had more time: ESL learners' changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing, 7*, 43-68.

Potenza, M. T. & Dorans, N. J. (1995). DIF assessment for polytomously scored items: a framework for classification and evaluation, *Applied Psychological Measurement* 19, 23–37.

Powers, D.E., & Fowles, M.E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement, 33*(4), 433-452.

Rasch, G. (1960). Probabilistic models for some intelligence and achievement tests. *Copenhagen: Danish Institute for Educational Research*.

Reed, W. M., Burton, J. K., & Kelly, P. P. (1985). The effects of writing ability and mode of discourse on cognitive capacity engagement. *Research in the Teaching of English*, 283-297.

Rhodes, T.L. (2012). Getting serious about assessing authentic student learning. *National Institute for Learning Outcomes Assessment: The Seven Red Herrings about Standardized Assessments in Higher Education [Occasional Paper #15]*.

Rhodes, T. L. (2010). *Assessing outcomes and improving achievement: Tips and tools for using rubrics*. Washington, DC: Association of American Colleges and Universities.

Roohr, K. C., Liu, O. L., & Liu, H. (2017). Investigating Validity Evidence for the ETS® Proficiency Profile. *ETS Research Report Series*.

Rosen, H. (1969). *An investigation of the effects of differentiated writing assignments on the performance in English composition of a selected group of 15-16 years old pupils* (Doctoral dissertation, Institute of Education (University of London)).

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., Gundy, C., Koller, M., Perterson, M.A.,& Sprangers, M. A. (2009). A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales. *Journal of clinical epidemiology*, *62*(3), 288-295.

Shavelson, R. J. (2013). On an approach to testing and modeling competence. *Educational Psychologist*, *48*(2), 73-86.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*(3), 215-232.

Simmons, J. (1992). Don't settle for less in large-scale writing assessment. In K. Goodman, L. B. Vird, & Y.M. Goodman, *The whole language catalog:*

Supplement on authentic assessment (p. 160-161). Santa Rosa: CA: American School Publishers.

Sparks, J. R., Song, Y., Brantley, W., & Liu, O. L. (2014). Assessing Written Communication in Higher Education: Review and Recommendations for Next-Generation Assessment. *ETS Research Report Series*, *2014*(2), 1-52.

Steedle, J.T. (2014) Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education, 27*:1, 58-76

Sundre, D. L., & Thelk, A. D. (2007). The Student Opinion Scale (SOS): A measure of examinee motivation. Test Manual. *Harrisonburg: Center for Assessment and Research Studies, James Madison University*.

Sundre, D. L., & Wise, S. L. (2003, April). Motivation filtering: An exploration of the impact of low examinee motivation on the psychometric quality of tests. In *annual meeting of the National Council on Measurement in Education, Chicago, IL.*

Sundre, D. L. (1999). Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?.*Paper presented at the Annual Meeting of the American Educational Research Association in Montreal, Quebec, Canada.*

Taylor, C., & White, K.R. (1981, April). *Effects of reinforcement and training on Title I students' group standardized test performance*. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles. (ERIC Document Reproduction Service No. ED 206 655)

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). Assessment of higher education learning outcomes. *Feasibility study report*, *1*.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39-55.

U.S. Department of Education. (2006). *A test of leadership: Charting the future of U.S. higher education*. Washington, DC: Author.

Van Merriinboer, J. J. G. (1997). *Training complex cognitive skills: A four-component instructional design model for technical training.* Englewood Cliffs, NJ: Educa tional Technology Publications

Voluntary System of Accountability (2017). History of the VSA. *Retrieved August 2017 from http://www.voluntarysystem.org/about/vsa_history*

Walker, C.M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. Journal of Psychoeducational Assessment, 29(4), 364-376.

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education*, *35*(5), 519-533.

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185-205.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65-83.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational assessment*, *10*(1), 1-17.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183.

Wolf, L.F., & Smith, J.K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.

Wolf, L.F., Smith, J.K., & DiPaulo, T. (1996, April). *The effects of test specific motivation and anxiety on test performance*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Wright, B. D., & Mok, M. M. (2004). An overview of the family of Rasch measurement models. *Introduction to Rasch measurement*, 1-24.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370.

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis. Rasch Measurement*. MESA Press, 5835 S. Kimbark Avenue, Chicago, IL 60637.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Yancey, K. B., Fishman, T., Gresham, M., Neal, M., & Taylor, S. S. (2005). Portraits of composition: How writing gets taught in the early 21st century. In *Conference on College Composition and Communication Annual Convention. San Francisco, CA*.

Zahner, D., & James, J. (2015). Predictive validity of a critical thinking assessment for post-college outcomes. *New York, NY: Council for Aid to Education. Retrieved from*

http://cae.org/images/uploads/pdf/Predictive_Validity_of_a_Critical_Thinking_Assessment_for_Post-College_Outcomes.pdf.

Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.