THE EFFECTS OF PRIMACY ON RATER COGNITION: AN EYE-TRACKING STUDY

By

Laura Ballard

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies—Doctor of Philosophy

2017

ABSTRACT

THE EFFECTS OF PRIMACY ON RATER COGNITION: AN EYE-TRACKING STUDY

By

Laura Ballard

Rater scoring has an impact on writing test reliability and validity. Thus, there has been a continued call for researchers to investigate issues related to rating (Crusan, 2015). In the current study, I answer the call for continued research on rating processes by investigating rater cognition in the context of rubric use in writing assessment. This type of research is especially important for rater training and rubric development because, despite efforts to guide raters to a common understanding of the rubric criteria and to help raters converge on a common understanding of scoring bands, variance in rater scoring and rater behavior persists. Researchers have shown that trained raters do not always use rubric criteria in consistent ways, nor do they consistently use the same processes to score samples. This is relevant for the design of and use of scores from analytic rubrics, as raters are expected to allocate equal attention to each criterion within an analytic rubric, and non-equal attention has been shown to coincide with category reliability (Winke & Lim, 2015), and, therefore, overall test reliability. One factor which has not been investigated in assessment research is the role of information-primacy in rater cognition. Thus, in this study, I investigate the primacy effect in relation to rater-rubric interactions. Specifically, I investigate 1) whether the position of a category affects raters' assignment of importance to the category; 2) whether the position of a category affects raters' memory of a category; 3) whether raters pay more or less attention to a rubric category depending on its position in the rubric; 4) whether the position of the category affects the inter-rater reliability of a category; and 5) whether the position of a category affects the scores that raters assign to the

category. I employed a mixed-methods within-subjects design, which included eye-tracking methodology. Thirty-one novice raters were randomly assigned to two groups and were trained on two rubrics in two phases. The rubrics were a standard rubric (from Polio, 2013) and a reordered rubric (identical to the standard rubric, except with categories appearing in a mirrored order to the reordered rubric). In round 1, raters trained on one of the two rubrics and rated the same 20 essays using the rubric. The second round took place five weeks after the completion of the first. In round 2, raters trained on the alternate rubric and re-rated the same 20 essays. I utilized several data-collection tools to investigate rater's cognition and behavior related to their rubric of training. I examined raters' beliefs about category importance, raters' recall of the descriptors in each rubric category, raters' focus on the rubric criteria during essay rating, and raters' scoring consistency and severity for each rubric category. Results show that as novice raters train on a new rubric and assign scores using the individual categories on the rubric, the raters' behavior pertaining to the outer-most positions (e.g., left-most and right-most) was most susceptible to ordering effects. That is, the findings show that the category position affected the raters' beliefs about what criteria are the most and least important when scoring an essay, how many descriptors raters were able to recall from a category, how much attention raters paid to a category on the rubric while rating, and how severely raters scored a given category. Additionally, the findings provided evidence that there was an interplay between the category type and category positions, resulting in either more pronounced primacy effects or leveling effects for individual rubric categories. Based on these findings, I discuss rater training, rubric design, and test-construct considerations for rubric designers and test developers.

Keywords: primacy effects, ordering effects, rater cognition, rater behavior, rater training, rubric design

ACKNOWLEDGEMENTS

While Dr. Megan Smith likened the dissertation process to a hike out of the Grand Canyon (see Smith, 2016), I have thought of it more as a mountain-summiting experience. The key parallels are the emotions shared in both the summit and dissertation experience and the people that ensure I make it to the end. I'll start by describing the mountain metaphor and the prevailing emotions that I experience during the summiting journey.

At the base of the mountain, I always simultaneously experience three emotions: dread, appreciation, and excitement. Allow me to unpack these feelings.

The dread: at the outset of the hike, I literally have the entire mountain ahead of me. There is a knowingness of the inevitable difficulty of the climb. Depending on the distance and elevation I plan to conquer, the intensity of dread varies. For a beginner, even a small peak seems unconquerable. The only way that I've rationalized being able to do it myself is by reminding myself that many people before me have succeeded in the same task. After having a few peaks under my belt, the fear and dread at the base of the mountain is allayed by previous experiences, and I have felt more comfortable signing up for bigger hikes on taller mountains. The taller mountains again revive a sense of dread commensurate with their height. But slowly, over time, I have tackled longer trails and higher peaks, yet all the same, at the base of the mountain, the dread lingers.

The appreciation: I recognize that not everyone has the time, the skills, the gear, and the interest to go backpacking for a few days. These are precious resources that I have had available to me that have made it possible to gallivant on hiking trails. I've needed a flexible job, a physical fitness and willingness to sweat, a collection of packs and mess kits, and a deep love for

the great outdoors. Taking the time to appreciate the access to these resources makes the hike itself that much sweeter.

The excitement: the hike itself can be grueling. At several points on the ascent, I question why I would ever take on such a task; I curse my past self for subjecting my current self to such fatigue, pain, and harsh conditions. While trial markers help me to pace myself and offer tangible goals to reach, they also are cruel reminders of the remaining climb, yet I trek on. My mood changes once I first catch a glimpse of the view from a high elevation, confirming that the hard work is actually getting me somewhere. Eventually, the trail starts to level out at the top, and, at this point, relief sets in. Finally, I reach the vista, and the backpack comes off. When the weight of all the gear (that has left my hips bruised and my shoulders soar) is removed, I feel like I am floating. Finally, I can stand (or sit) at the peak and marvel at the view. Once there, soaking it all in, I reflect on how worthwhile the trek has been because of the great beauty of landscape that can only be fully seen and appreciated from the top. And I think about how few people have seen this view, solely because few are willing to put in the sweat and the commitment to the hardship of the hike. This is the excitement: the awe of the view, the worthwhile reward of the hours-long upward-bound sufferfest.

On the journey, there are several people integral to seeing me through the hiking endeavor. There are those that have done the hike before, see my enthusiasm and skill, and push me to challenge myself into a greater hike. There are those back at home who think I am crazy but admire the strength it takes to accomplish something of this nature and wish me well. There are those that I meet at various points in my journey (gear preping, packing, trail-head finding, etc.) that provide advice and prepare me with the skills, gear, and knowledge I need to actually get the job done. And perhaps most important are those who sign up to go with me on the trail,

who are a similar type of crazy and, in some ways, are kindred trail spirits who really understand the thrill of the summit. All of these people are essential, as I know I would never dare to tackle a hike on my own without the support, input, and companionship of others.

How does this summit metaphor relate to the dissertation process? The dread is real. At the beginning of such a mountain of a research project, the height, the length, and the difficulty of path feels insurmountable. It seems unreasonable. It is daunting. But the success (and encouragement) of other dissertation sojourners has coaxed me on and given me the courage to undertake such a massive project. This project, from conception to fruition, has taken more than a year and a half, more than 250 hours of data collection, more than $5000 of resources, and input from more than 45 people (committee members, expert raters, testing office personnel, and participants, to name a few). It has been a massive, daunting project.

The appreciation is fundamental. Like being a hiker on a mountain journey, being a PhD student is a privilege. Over the past six years during my Master's and doctoral program, I have the time, the mental space, the support, and the resources at my disposal to be able to pursue an advanced degree, and the wonder of this reality is not lost on me. It is a gift to be a student who studies such a narrow sliver of reality and to be part of a community who is committed to discovery.

The excitement is the driving force. At several points in the dissertation process, I found myself wondering why I would ever sign up for such a difficult and trying project. Discouragement, fatigue, and frustration have been present in the uphill climb. But the glimpses of the data narrative (i.e., the landscape) on the upward journey are compelling and have beckoned me to continue. When finally arriving at the summit, the excitement of being at the vista is tangible. This is the place where the data is in full view. On the upward climb, I've only

caught glimpses, but now at the top, I can see it all together and marvel at the beauty of the full story that the data landscape shows. While sitting on top and appreciating the view, I recognize that few will reach this point, not because few are able, but because few are willing to sign up for the sufferfest. But the reward is worthwhile. The view is remarkable, memorable, and captivating. I marvel at what I've been able to see and experience as part of the dissertation process and count myself lucky to be able to sit atop and appreciate a view that few will ever see.

On this dissertation journey, I have many people to thank for their various roles in helping me reach the summit. There are those that have done the hike before, who saw my enthusiasm and skill, and pushed me to a greater challenge. Paula Winke and Charlene Polio have been these encouragers. Were it not for them pulling me aside during my Master's program, I never would have considered lacing up my hiking boots and setting my eyes on the summit. I'm grateful for their gentle shove toward the trailhead.

There are those back at home who thought I was crazy for signing up for the sufferfest, but have been loving and supportive all the same. I have appreciated the encouraging words of my parents, sister, and brother, and the time to be goofy and wild with my niece and nephews. I have enjoyed the time I have had here in Michigan with my Granny and Aunt Van, and I am grateful for their tangible care through food and quality time. To Jessi and Kelly, I am grateful that they have understood the intensity and demand of this program and have been wildly generous in accepting the little I have been able to offer in our friendship over these past few years. I cherish our adventures together, and their friendship has meant so much to me.

There are those that I have met at various points in my graduate school journey that have advised me and prepared me with the equipment, the skills, and the knowledge I needed to actually get the dissertation done. To those who have helped fill my backpack with the gear I

need:

- Russ Werner, my technological savior. I feel indebted for all of his help over the past six years regarding my technological needs in the classroom, research lab, and life in general. I have asked many a stupid question, but he has been only generous with his time and expertise.
- Mike Kramizeh and Luca Giupponi. They have provided the recording gear and lab space I needed to make materials and collect data in this project.
- Dan Reed, Aaron Ohlrogge, and Andy McCullough. They have been instrumental in supporting this research and providing essays and rater-training materials around which to build this study.
- The expert raters and pilot-study participants. They have helped me to create and refine my data-collection materials, and thus improve the quality of this project.
- Cambridge Michigan Language Assessments and the Center for Applied Linguistics. The internship programs at these assessment organizations provided me an invaluable opportunity to do hands-on work in the assessment field, which captured my attention and fueled my desire to be on the trail. I am especially thankful to Meg Montee and Robbie McCord who invited me into the work they were doing and facilitated my internship experience.

To those who have sat down with me to look at the trail map and strategize the best route to take:

- Dan Isbell and Dr. Ryan Bowles. Dan sat down with me to do a lot of button-clicking and strategizing for my Rasch analyses in Winsteps, which prevented me from doing a lot of loops on a side trail or starting up a dead-end route. Dr. Bowles provided valuable input about the Rasch analyses, which helped me to visualize why I might consider

taking the switchbacks instead of just climbing up the side of a cliff.

- Tatyana Li. She spent many hours discussing the forks in the data-analysis trail and helped me to make the best decisions about which trail to take based on my skills and gear.
- Emily Weiss. She was sacrificial in picking up a A LOT of extra slack in the office, which allowed me to collect data and keep moving up the mountain.

To those who have cared for my needs along the trail, when I was hungry, discouraged, and physically ailing:

- My community at Capital City Vineyard. They have been a constant encouragement in my grad school journey. They have both literally and figuratively fed me in my weary moments.
- Dr. Guggenheim, the Olin PT staff, and Beth Malsheske. The difficulty of the journey was augmented by the physical challenges that I have faced. The medical professionals that met me on the trail and bound up my wounds have enabled me to finish the trek. I can't express my gratitude enough to these caring, compassionate, and competent individuals without whose care I would have given up.

To those who thought the journey worthwhile and were willing to financially invest in it:

- Second Language Studies, who provided research funds.
- The Graduate School, who provided a Research Enhancement Award.
- The College of Arts and Letters and The Graduate School, who provided a Dissertation Completion Fellowship that allowed me to collect data in the Fall 2016 semester.
- The College of Arts and Letters, who provided research funding through the Grant Initiative Award.

- The International Research Foundation for English Language Education (TIRF), who provided a Doctoral Dissertation Grant.
- Education Testing Services, who provided a TOEFL Small Grant for Doctoral Research in Second Language Assessment.

Finally, perhaps most important have been those who signed up to go with me on the trail, who are a similar type of crazy and, in some ways, are kindred spirits who really understand the thrill of the summit.

- The Graduate InterVarsity community, especially Mandy, Klodi, Jens, Jenn, Carolyn, Camille, Victor, Yeshoda, The Smiths, and The Ahlins. They have helped me to think through what it means to be a mindful, intentional, faithful graduate student, colleague, and researcher.
- The Fighting Narwhals and the Sad Song Band. They have provided an amusing outlet for athleticism and creativity. I know that I shant find another sports team or music group as fun to play with in all the land.
- The SLS student body, the SLS faculty, and the ELC faculty. They are brilliant, committed, and willing to wrestle though tough issues in our field. Rubbing shoulders with such people has inspired me to dig deep in my own academic journey.
- Dr. Megan Smith. She was the one who started the hike just before me, and when at the top, yelled down from the summit "If I can do it, so can you! You're almost there!"
- My housemates (both current and former), Erika, Skyin, Camille, Emily, Lauren, Erin, Jess, Spencer, and Zeke. They have kept me sane. They have commiserated in times of frustration, celebrated the milestones, and provided a place where I could let down my guard. They have been wily partners in crime and enthusiastic adventure buddies. You

have made my journey so rich and meaningful. I will sorely miss our heart-to-hearts and shenanigans.

- Dr. Jessica Fox. She is the lady from another planet who has helped me to see and understand life from a very different perspective. In terms of real-life skills and knowledge, she has taught me more than a graduate program ever could. She is a generous, kind-hearted, insightful, gracious teacher. I am deeply grateful for her friendship.
- The National Park Rangers (i.e., my dissertation committee), Dr. Paula Winke, Dr. Susan Gass, Dr. Charlene Polio, Dr. Aline Godfroid, and Dr. Dorry Kenyon. I am thankful for their investment in me and in this project. I recognize that it is a large commitment to accept a position on a dissertation committee; they committed to giving no small amount of their time and attention. They made sure I had all my gear in order at the mountain base. Once I started the trek, they monitored to make sure I didn't fall off the mountain. And at the summit, they are the experts that have helped me to appreciate the nuances of the panoramic landscape. I am especially grateful to Paula, my committee chair. She has been the first responder, the one on the other end of the walkie-talkie who has talked me through tough challenges on the trail and who has pointed me toward vital resources. She has been generous with her time, encouraging in every interaction, and has cared for me as a student and a person. I couldn't have asked for a more caring advisor.

Without the love, support, care, and encouragement of all of these kind souls, the dissertation wouldn't have been possible. I am deeply indebted and deeply grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Rater scoring has an impact on performance test reliability and validity. Thus, there has been a continued call for researchers to investigate issues related to rating (Crusan, 2015). Myford (2012) exhorted researchers and test designers to "do all that we can to help ensure that the ratings that raters assign are accurate, reliable, and fair" (p. 49). Second language testing researchers are committed to this goal and have been researching the various facets that affect test scoring processes for years (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). In second language writing assessment, such emphasis on investigating the scoring process and how raters arrive at particular scores have been seen as critical "because the score is ultimately what will be used in making decisions and inferences about writers" (Weigle, 2002, p. 108).

In the current study, I answer the call for continued research on the rating process by investigating rater cognition in the context of writing assessment. Research on raters' cognitive processes "is concerned with the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (Bejar, 2012, p. 2). A theme central to rater cognition is the way in which raters interact with rubrics. Only by understanding this interaction will test designers be able to improve rubrics, rater training, and test reliability and validity (Barkaoui, 2010). Performance test validity is tied to raters and rubrics, in particular, because there are certain propositions that must be counted as true in order for scores to be considered valid (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, see Standard 6.9). For

example:

- "Raters attend to the criteria included in the rubrics when making their judgments (i.e., they are using appropriate criteria when they are assigning their ratings).
- Raters use the categories on the rubrics in the intended manner, applying the rubrics consistently and accurately to judge each performance (or product)" (Myford, 2012, pp. 48-49).

In this study, I focus on rater-rubric interactions, which continue to be of interest because, despite rater-training efforts, variance in rater behavior and scores persist (Lumley & McNamara, 1995; McNamara, 1996; Weigle, 2002; Weir, 2005) which may lead to reliability problems. Though the goal of rater training is to give raters a common understanding of the rubric criteria and to help raters converge on a common understanding of scoring bands (Bejar, 2012; Roch, Woehr, Mishra, & Kieszczynska, 2012), many studies on rater behavior have shown that raters do not always use rubrics in a consistent way (i.e., they have low intra-rater reliability). Raters do not consistently score (i.e., they have low inter-rater reliability), and they do not use the same processes to arrive at a given score (Cumming, Kantor, & Powers, 2002; Eckes, 2008; Kondo-Brown, 2002; Lumley, 2002; Orr, 2002). As Winke and Lim (2015) suggested, one potential explanation for rater behavior and problems with inter-rater reliability may be the *primacy effect*. The primacy effect, discussed below in detail, is a psychological phenomenon that shows that the positionality of information in a list (e.g., a rubric) affects a listener's or reader's assignment of importance to that information (Forgas, 2011). This seems particularly relevant for helping to explain how raters pay attention to rubric criteria. Primacy may have a potential impact on inter-rater reliability on analytic rubrics. No researcher, however, has directly investigated the role of primacy in rater cognition and its potential effects on rater scoring (but see Winke & Lim, 2015,

who posited that primacy effects were observable in their study). Thus, in the current study, I propose to investigate primacy effects in relation to rater-rubric interactions, and I want to examine whether they affect behavior, such as mental-rubric formation, attention to criteria, and rater scoring, when raters use an analytic rubric.

The Primacy Effect and Decision Making

Definition and Hypotheses

In the field of psychology, the primacy effect (also known as serial position or ordering effects) is a long-researched phenomenon, and the theories and hypotheses that explain the phenomenon have shifted throughout the years (Asch, 1946; Crano, 1977; Forgas, 2011; Hendrick & Costantini, 1970; Tulving, 2008; Underwood, 1975). As I will explain below, two theories (*interference* and *inconsistency discounting*) have fallen out of favor in the psychology community, and the *attention decrement hypothesis* is upheld today. Researchers have largely agreed upon the existence and impact of the primacy effect; that information-presentation order matters in how people process, retain, and make decisions based on information presented to them. The *primacy effect* refers to the better recall of information that is presented first in a list of information. The *recency effect*, on the other hand, refers to the better recall of information that is presented toward the end of a list of information. Much of the research conducted on ordering effects has investigated the underlying cause of any ordering effect. The *interference theory*, which states that the rehearsal of information presented first leads a person to block out subsequent information, has not been supported (Crano, 1977; Underwood, 1975). Likewise, a theory of *inconsistency discounting* suggests that when information presented later in a list differs from (or is inconsistent with) information presented first, people discount the later information (Anderson, 1965). However, this theory has not been supported either (Hendrick &

Costantini, 1970). The most widely-accepted model that accounts for the primacy effect is the *attention decrement hypothesis* (Crano, 1977; Forgas, 2011; Hendrick & Costantini, 1970; Tulving, 2008). This hypothesis asserts that there is a "progressive decline in attention to trait descriptors over the course of a complete list… [and later information is] less heavily weighed in the process of impression formation. The relative influence of a descriptor varies as a function of its serial position" (Crano, 1977, p. 90). In other words, the reason that primacy effects occur is because people fail to process later items as carefully and attentively as earlier ones (Crano, 1977; Forgas, 2011). The attention decrement hypothesis is even thought to be supported by biological evidence. Tulving (2008) proposed that a biological process called *camatosis*, a slowing of neuron activity in the brain, is related to memory and information retention. Basically, when a list of information is presented, the brain becomes fatigued and has fewer resources available to attend to later-presented information. Tulving proposed that a measured decrease in neural activity could be observed throughout the course of information presentation, resulting in the most attention being paid at the beginning of information presentation, and a deceasing amount of attention being paid as information presentation continues. In other words, primacy is a psychological phenomenon undergirded by a biological phenomenon.

Primacy Research in Decision-making Tasks

Decision making has been one of the primary areas in which researchers have investigated the real-world implications of the primacy effect. This is because the order in which information is presented and the amount of attention paid to the information has a measurable impact on subsequent decisions (Forgas, 2011; Hendrick & Costantini, 1970; Luchins & Luchins, 1970; Rebitschek, Krems, & Jahn, 2015). What previous researchers have found is that primacy is the brain's default setting; as new information is presented, attention to that

information decreases. Stronger memory links are formed with first-order information than later information. This later leads to decisions that favor the first-order information. However, when a person makes an effort to pay equal attention to information presented in a list (for example, by repeating the information out loud rather than just listening to it), recency effects are the default, and later-presented information is better remembered by the person and has a larger influence on his or her subsequent decisions (Forgas, 2011; Hendrick & Costantini, 1970; Luchins & Luchins, 1970). This does not mean that people attempt to memorize the information; rather, people make an effort to equally attend to all information that is presented. To summarize, when equal attention is paid to all information, primacy effects disappear and recency effects manifest, likely because the most recently encountered information is remembered better (Hendrick & Costantini, 1970). The amount of attention paid to the information is key to understanding information retention patterns.

Another variable that plays a role in primacy is time. In a study on impression formation, Luchins and Luchins (1970) demonstrated that time lapse plays an important role in memory, information retention, and ordering effects. In their study, they gave two groups of participants a descriptive paragraph about a person, Jim, and these descriptions were somewhat conflicting. After reading the description, the researchers gave participants a paragraph to read that had contrary facts about Jim. Directly after reading each descriptive paragraph, and then ten minutes later, one week later, and one month later, the participants were asked to complete a questionnaire about Jim's personality. Results showed that directly after reading the paragraphs, recency effects influenced the participants' impressions of Jim. However, as more time lapsed, the primacy effects became stronger, showing that the participants' lasting impressions were more strongly influenced by whichever paragraph they read first. This study, along with others,

have shown that in short-term contexts, recency effects are more influential in immediate decision-making tasks, whereas in long-term situations, primacy effects play a larger role in decision-making tasks (Forgas, 2011; Hendrick & Costantini, 1970).

Rater Behavior and Decision Making

The act of scoring itself is a decision-making process that is meant to be informed by rubric criteria, which are presented and exemplified during rater training (Baker, 2012; Cumming et al., 2012; Roch et al., 2012). The goal of rater training is to bring raters into alignment to score by the rubric criteria, rather than by their own personal criteria (Weigle, 1994), thus achieving both intra-rater and inter-rater reliability in scoring. Rater training has been shown to increase rater alignment in scoring (Weigle, 1994; Weigle, 1998; Lim, 2011; Solano-Flores & Li, 2006; Roch et al., 2012), but other factors contribute to rater decision making and scoring (Lumley & McNamara, 1995; McNamara, 1996; Weigle, 2002; Weir, 2005), which often lead to problems with inter-rater reliability. In this study, I propose to investigate the role of primacy in rater-scoring behavior. I look particularly at how primacy could affect the way in which novice raters interact with and retain rubric criteria through rater training and during essay scoring. It may be particularly important to look at what criteria raters include in the construction of their "mental rubric" (Bejar, 2012) during rater training, and how this relates to rater behavior and rater reliability during operational scoring.

Different rubrics lead to different rater behavior (Knoch, 2009; Li & He, 2015). In language testing, there are two primary rubric types, holistic and analytic. With a holistic rubric, raters consider the text as a whole, and assign one score to the essay based on the raters' overall impression. In holistic scoring, the essay is thought of as "a whole entity" that should be judged as one unit because "the whole is not equal to the sum of the parts," but "the whole is equal to

the parts and their relationships" (Goulden, 1992, p. 265). With an analytic rubric, raters score an essay based on multiple individual facets of writing, and these scores are often summed to arrive at a total score. In analytic scoring, it is assumed that "the sum of the subscores for the parts is exactly equal to a valid score for the whole and, by evaluating the parts, the rater has evaluated the whole" (Goulden, 1992, p. 265). Key features of holistic rubrics are that the rubrics are less detailed. They allow raters to score based on impressions and bring in criteria that are not defined in the rubric criteria. Analytic rubrics, on the other hand, clearly define the assessment criteria, and the corresponding process assumes that raters will attend to all rubric criteria equally (Barkaoui, 2007; Goulden, 1994; Weigle, 2002).

I hypothesize that primacy effects would be most salient in the use of analytic rubrics. Due to the nature of holistic rubrics, I hypothesize that raters would perhaps be less vulnerable to primacy effects because holistic rubrics tend to be more impressionistic in nature, and raters can easily focus on any criteria to form their mental impressions of scores (Eckes, 2008; McDermott, 1986; Munro & Derwing, 1995). Analytic rubrics, on the other hand, may open the opportunity for ordering effects to play a larger role in raters' formation of mental rubrics[1] because of the vast amount of information within each category, which is divided across score bands. Analytic rubrics lay out detailed criteria descriptors listed by category and score band (Barkaoui, 2011; Weigle, 2002), but this also means that, because of the potentially long, detailed category-by-category layout, analytic raters would be the most susceptible to primacy effects, in which raters may pay less and less attention to category descriptors (or entire categories) as they read from left to right or top to bottom. As Hamp-Lyons and Henning (1991) opined, the large number of descriptors included in analytic rubrics may cause a heavy cognitive load, which may become

[1] The mental rubric is what raters are able to hold in their minds from the rubric (e.g., rubric descriptors) and how important they believe the criteria from the rubric to be.

unmanageable for raters. If raters do not have the attention or capacity to equally attend to all rubric descriptors, then primacy effects may be to blame. However, this is only speculation, as very little research has been done on the effects of rubric format (see Table 1 for a summary of research examining analytic rubrics and the rating process). Nevertheless, one study by Winke and Lim (2015) has investigated raters' cognitive processes while rating essays using a version of the Jacobs, Zingraf, Wormuth, Hartfiel and Hughey (1981) analytic rubric. They found that raters spent the majority of their rubric-use time attending to the left-most categories, and the least amount of time focusing on the right-most categories. The authors suggested that a probable explanation for the raters' lack of attention to the right-most categories in the rubric was due to the primacy effect. However, because their study was not set up to investigate primacy effects, they were only able to suggest primacy as a likely explanation for the observed rater behavior. To my knowledge, no study in language testing or educational measurement has explicitly investigated primacy effects in relation to rater cognition and rubric use.

Table 1

Summary of Research on Analytic Rubrics in Second Language Assessment

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Bacha (2001), System	English as a Second Language (ESL) Composition Profile (Jacobs et al. 1981)	Content Organization Vocabulary Language Mechanics	Raters (2): L1 or L2 English Essays (30): L1 Arabic, L2 English	To examine what two types of rubrics, analytic and holistic, can tell assessors in essay evaluation	1. What do two types of writing rating scales (analytic and holistic) tell teachers in evaluating their students' essays?	Holistic scoring may mesh many of the traits assessed separately in analytic scoring, making it relatively easier and reliable. However, it is not as informative for the learning situation (i.e., provides less detailed information) as analytic scoring. A combination of holistic and analytic evaluation is needed to better evaluate students' essay writing proficiency (pp. 380-381).

Table 1 (cont'd)

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Barkaoui (2007), Assessing Writing	Holistic rubric: EFL Placement Test rubric developed by (Tyndall & Kenyon, 1996) Multi-trait rubric: Composition Grading Scale (Brown & Bailey, 1984)	Content Organization Grammar Mechanics Style	Raters (4): L1 or L2 English Essays (24): L1 Arabic, L2 English	To improve the reliability and validity of the assessment by introducing a rating scale with explicit and standard guidelines for evaluating students' EFL writing and to determine which method should be used among holistic and multiple-trait scoring (p. 89).	1. What are the effects of holistic and multiple-trait rating scales on the dependability of the essay scores EFL teachers assign? 2. What are the effects of holistic and multiple-trait rating scales on EFL teachers' decision-making behaviors and the essay features they attend to? 3. How do EFL teachers perceive holistic and multiple-trait rating scales?	The holistic scale resulted in higher inter-rater agreement. Raters employed similar processes with both rating scales. Raters were the main source of variability in terms of scores and decision-making behavior (p. 86).

Table 1 (cont'd)

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Carr (2000), Issues in Applied Linguistics	UCLA English as a Second Language Placement Exam (UCLA ESLPE) Composition Rating Scale	Content Rhetorical control Language	Raters (?): not described Essays (83): L2 English	To examine how different composition rubrics (analytic and holistic) can differentially affect the aspects of academic English ability measured in an ESL proficiency test.	1. To what extent do holistic and analytic scales contribute differentially to total scores on a test of academic English ability? 2. To what extent does the test as a whole measure different aspects of language ability, depending on whether analytic or holistic composition scores are used? 3. To what extent does a particular rating scale type provide potentially useful information for placement or diagnosis, either alone or as part of a multi-component assessment? Multiple	Changing the composition rubric type not only changes the interpretation of that section of a test (within a larger multi-skill test) but may also result in total test scores which are not comparable. Each rubric is a different operationalization of the academic-writing-ability construct. Holistic scores provide an assessment of a single construct, whereas composite scores from an analytic rubric conflate the information from several constructs (p. 228).

Table 1 (cont'd)

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Knoch (2009), Language Testing	Diagnostic English Language Needs Assessment (DELNA) rating scale	Organization Coherence Style Data description Interpretation Development of ideas Sentence structure Grammatical accuracy Vocabulary Spelling	Raters (10): L1 or L2 English Essays (100): L2 English	To establish whether an empirically developed rating scale for writing assessment with band descriptors based on discourse analytic measures would result in more valid and reliable ratings for a diagnostic context than a rating scale typical of proficiency testing (p. 277).	1. Do the ratings produced using the two rating scales differ in terms of (a) the discrimination between candidates, (b) rater spread and agreement, (c) variability in the ratings and (e) what the different traits measure? 2. What are raters' perceptions of the two different rating scales for writing?	The trait scales on the new rubric resulted in a higher candidate discrimination, smaller differences between raters in terms of leniency and harshness, greater rater reliability, and fewer raters rating with too much or too little variation (p. 285). Rater feedback also showed a preference for the more detailed scale. (p. 275).

Table 1 (cont'd)

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Winke & Lim (2015), Assessing Writing	Jacobs et al. (1981) analytic rubric	Content Organization Vocabulary Language Mechanics	Raters (9): L1 English Essays (40): L2 English	To describe raters' cognitive processes when they use an analytic scale and to relate raters' scoring processes to inter-rater reliability.	1. To which part of the analytic rubric do raters pay the most attention? 2. Are inter-rater reliability statistics on the subcomponents of the rating rubric related to the amount of attention paid to those subcomponents?	Attention was associated with inter-rater reliability: Organization (the second category) received the most attention (slightly more than the first, Content). Organization also had the highest inter-rater reliability. Raters attended least to and agreed least on mechanics (the last category). Raters who agreed the most had common attentional foci across the subcomponents. A potential explanation for this behavior is primacy: raters paid the most attention to organization and content because they were on the left (and were read first by raters) (p. 37).

Table 1 (cont'd)

Author, Year, and Journal	Rubric	Analytic Rubric Categories	Raters and Essays	Purpose of Study	Research Questions	Primary Results
Wiseman (2012), Assessing Writing	In-house analytic and holistic rubrics (see Wiseman, 2005)	Task fulfillment Control of content development Organizational control Sociolinguistic competence Grammatical control	Raters (8): L1 or L2 English Essays (78): L1 Spanish, Chinese, Korean, Japanese, Polish or Russian; L2 English	To investigate decision-making behaviors of raters as they were scoring a performance-based assessment of second language writing ability using two rubrics, and to uncover the impact of rater effects on scores.	1. What are raters' decision-making behaviors as they score a performance-based assessment of second language writing ability using a holistic and analytic rubric?	Rater background may have contributed to rater expectations, which might explain raters' individual differences in the application of the performance criteria of the rubrics when rating essays. Additionally, rater ego engagement with the text and/or author may have helped mitigate rater severity, and self-monitoring behaviors by raters may have had a similar mitigating effect (p. 150).

Rater Cognition Research and Eye-movement Methodologies

The majority of research on rater behaviors and decision-making processes has been done using verbal protocols (Barkaoui, 2010, 2011; Li & He, 2015; Lumley, 2002; Weigle, 1994; Wiseman, 2012), where raters were asked to concurrently or retrospectively explain what they were thinking while reading and rating an essay. However, think-aloud protocols, or TAPs, have been shown to suffer from 1) veridicality, or raters thought processes not being fully or accurately expressed, and 2) reactivity, or interference with the process at hand (i.e., essay rating), which causes a change in the raters' actual rating behavior (Barakaoui, 2010; Lumley, 2005). A new methodology that has been recently used in rater-behavior research is eye tracking, in which raters' eye-movements are tracked on a computer while raters read and score an essay and read a rubric. Eye tracking offers the benefit of capturing what a reader is focusing on without interfering with the reading process (Godfroid & Spino, 2015), and eye movements can provide a window in readers' online cognitive processes (Pollatsek, Reichle, & Rayner, 2006). Borrowing these notions on reading and applying them to reading in a language testing context, Winke and Lim (2015) used eye tracking to uncover rater-rubric interactions during the rating process, showing how raters focused their attention on different categories in the analytic rubric. The authors showed that inter-rater reliability is related to the amount of attention raters paid to the individual subcomponents of an analytic rubric; raters paid more attention to and had higher inter-rater reliability on subcomponents that appeared first (to the left) on the rubric, which also suggests a possible relationship between primacy and reliability. Their study demonstrated how eye-tracking methods can be used to uncover rater behaviors during the rating process. The authors called for more research on rater-rubric interactions, but also cautioned that

researchers would need to control for how the rubric categories are presented and focused on during rater training, which they did not control for, and they suggested that researchers should also control for the number of words in each subsection and for the number of points possible in each subsection (i.e., how important raters perceive them to be based on how many points are possible in the category). Winke and Lim's rubric, unfortunately, was not balanced; rather, later subsections (those on the right) had fewer words and counted for fewer points, leaving the primacy effect in rubric use in need of further investigation.

Variables and Research Questions

Answering the call to investigate and understand rater cognition more fully, in the current study I examined ordering effects in raters' mental-rubric formation and the potential effects on raters' subsequent decision-making processes and essay scoring when using an analytic rubric. The variables that I investigated are operationalized in Appendix A.

In this research, I seek to answer the following research questions:

1. To what extent do raters show evidence of ordering effects in their mental-rubric formation after rater training?
2. To what extent do raters show evidence of ordering effects through their rubric usage during rating?
3. To what extent are raters' scores impacted by ordering effects?

CHAPTER 2

METHODS

Methods and Procedure

Participants

For this study, I recruited native-English-speaking undergraduate students who had no experience rating performance assessments. Seventy-three students expressed interest in partaking in the study, but only 31 students met the established criteria. The participants were undergraduate students from three colleges within the university (the College of Arts and Letters, the College of Education, and the College of Social Sciences), and I recruited the participants directly from Language Learning and Teaching courses (offered by the Linguistics department). I recruited others by posting flyers on campus that advertised my research study. Each participant received $150 for compensation for approximately 12 hours of data collection. The participants had a mean age of 21 (SD = 1.77). For background information about the participants, see Table 2.

Table 2

Participant Descriptives

Name	Sex	Age	College	L2 Language (Writing Proficiency)	Comfort in Applying Rubric	Perceived Capability as Rater
Bree	Female	21	Education	N/A	5	4
Bunita	Female	20	Arts and Letter	Spanish (4)	6	5
Chezbub	Female	20	Social Sciences	Chinese (2)	5	3
Chips	Female	21	Social Sciences	N/A	5	4
Diane	Female	19	Arts and Letters	Spanish (4)	4	4
Elsa	Female	21	Social Sciences	Korean (2)	5	4
Emily	Female	19	Arts and Letters	Japanese (2)	4	3
Henry	Male	21	Social Sciences	French (1)	6	4
Hermione	Female	19	Arts and Letters	Spanish(4)	6	5
I. Chesterton	Male	21	Education	Spanish (1)	4	4
Judy	Female	21	Arts and Letters	Spanish (3)	5	4
Kai	Male	25	Arts and Letters	Japanese (2)	5	3
Kaya	Male	21	Social Sciences	German (4)	3	3
Lo	Female	20	Social Sciences	Chinese (1)	5	4
Luna	Female	20	Arts and Letters	Spanish (1)	4	4
Mark	Male	20	Arts and Letters	Spanish (1)	5	4
Nat	Female	20	Education	Spanish (1)	6	5
Otter	Female	22	Communications	Spanish (4)	4	4
Patrice	Female	21	Social Sciences	N/A	4	4
Remus	Female	21	Arts and Letters	French (3)	4	4
Sam	Female	21	Education	French (1)	5	4
Selena	Female	21	Education	Spanish (3)	6	3
S. Jello	Female	18	Arts and Letters	Spanish (2)	4	4
Sylvia	Female	20	Arts and Letters	German (1)	5	3
T. Laurel	Female	27	Arts and Letters	N/A	5	3
Tish	Female	19	Social Sciences	Spanish (1)	5	4
Victoria	Female	21	Arts and Letters	Japanese (1)	4	4
Warlie	Male	23	Arts and Letters	Chinese (4)	6	5
Yulietta	Female	21	Social Sciences	Spanish (6)	4	4

Procedure

For this study, I employed a within-subjects mixed-methods design. Specifically, I used a concurrent nested model in which the qualitative methods were nested inside a larger

quantitative design. Though I collected qualitative data (i.e., interview data and decision-making-process outline data), they will not be discussed in this dissertation.

As shown in Table 3, I pseudo-randomly assigned participants to one of two groups (i.e., Group A and Group B) according to participants' availability for data collection. The groups differed only in the presentation order of the two rubrics. For rater training and rating, Group A used the standard rubric in Round 1, followed by the reordered rubric in Round 2, and Group B first used the reordered rubric in Round 1 followed by the standard rubric in Round 2.

The data-collection procedure is shown in Table 4. The study included two rounds of data collection, and each round consisted of two sessions: rater training and rating. The Round 1 and Round 2 procedures were identical, except that Round 1 included a consent form at the beginning of the rater training session, and Round 2 included a background questionnaire at the end of the rating session.

The purpose of the *rater training session* was to train the participants on a rubric and to collect their pre-training and post-training recall of and/or beliefs about the rubric criteria. In the rater training sessions, the following took place: pre-training criteria importance survey, rater training, post-training category recall task, and post-training criteria importance survey. This session took place on campus in groups of two to five participants in a conference room. In total, the rater training session lasted approximately three hours.

The purpose of the *rating session* was to have the participants rate essays, collect the participants' scores on these essays, collect their pre-rating and post-rating recall of and beliefs about the rubric criteria, and to better understand the participants' rating process. The data collection for the rating session took place individually in a Tobii eye-tracking lab on campus within 2 days (but not on the same day) of the completion of the rater training session. During

the rating session, all participants did the following: pre-rating category recall task, pre-rating criteria importance survey, rubric reorientation, essay rating, decision-making-process outline, interview, and post-rating criteria importance survey. Each component was completed via computer at the eye-tracking lab. All raters scored all essays (N = 20; 20 during Round 1 and the same 20 during Round 2) on a Tobii TX-300 eye-tracking computer. See *Figure 1* for a picture of the eye-tracking and essay-rating setup. Data collection with each participant for the rating portion took approximately three hours.

Round 2 of rater training and subsequent rating took place approximately five weeks after Round 1. The main difference between Round 1 and Round 2 was that raters were trained on and rated with a different rubric for each round (see Table 3). For Round 2, participants were told that the English Language Center Testing Office decided to make minimal changes to the rubric, and that for the study, I would train them on whatever rubric was currently being used by this office. The changes of the rubric were not discussed further.

Table 3

Rubric Counterbalancing

	Round 1	Round 2
Group A	Standard rubric	Reordered rubric
Group B	Reordered rubric	Standard rubric

Table 4

Procedure Summary

Phase	Round 1	Round 2
Rater Training	Consent form (5 min.)	CRT3 (20 min.)
	CIS1 (5 min.)	CIS5 (5 min.)
	Training/norming (2.5 hours)	Training/norming (2.5 hours)
	CRT1 (20 min.)	CRT4 (20 min.)
	CIS2 (5 min.)	CIS6 (5 min.)
Rating	CRT2 (20 min.)	CRT5 (20 min.)
	CIS3 (5 min.)	CIS7 (5 min.)
	Essay rating (2 hours)	Essay rating (2 hours)
	DMPO (15 min.)	DMPO (15 min.)
	Interview (20 min.)	Interview (20 min.)
	CIS4 (5 min.)	CIS8 (5 min.)
		Background Questionnaire (5 min.)

Note. CIS = Criteria Importance Survey. CRT = Criteria Recall Task. DMPO = Decision-making Process Outline.

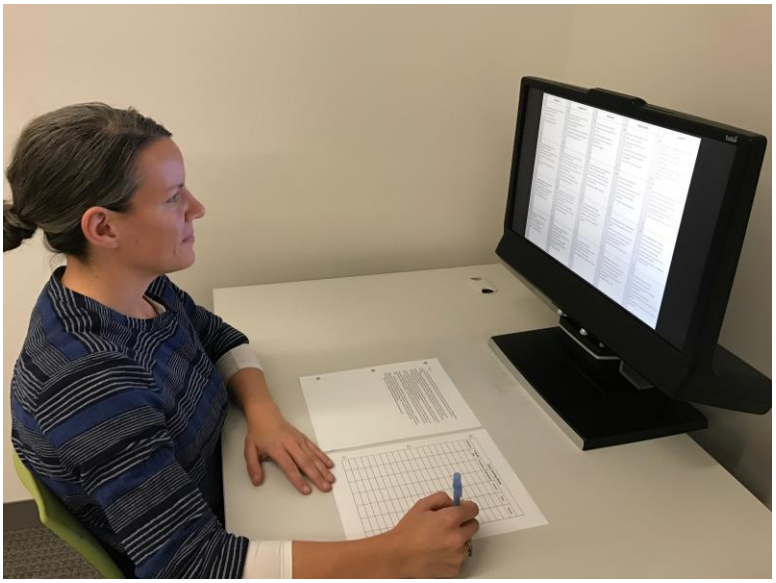


Figure 1. Photograph of the eye-tracking and essay-rating setup.

Materials

Essays. During the rating sessions, participants rated twenty essays. I selected these essays from a batch of forty-five handwritten essays, which were provided to me (without names) by the English Language Center. The essays were written by ESL students as a part of the university's English language placement exam, the Michigan State University English Language Test (MSUELT). The essays were written in response to independent prompts, which are shown in Appendix B. For the 45 essays, the MSUELT essay coordinator provided me with an equal distribution of essays from different score bands represented on the MSUELT scoring rubric by officials in the English Language Center, a different modified version of the Jacobs et al. (1981) rubric used in-house. For this current study, I typed each essay, maintaining all content and formatting from the original essay. Then, three expert raters rescored all 45 essays using the *Standard Rubric* developed for the current study, and I chose twenty of the rescored essays to be used for the participant rating sessions. I chose essays that represented a range of total scores, were less than one typed page, and received agreeing scores by the experienced raters. When used in the rating sessions, each participant rated the essays in a random order. During rating, the participants read a printed, hardcopy version of the essays, which was presented to each participant in a binder on the table in front of the participant. The participants recorded their scores for the essays by hand on a scoring sheet, which was taped on the table next to the essay binder.

Rubrics. I used two analytic rubrics that contained the same content but differed in layout. The rubric used was taken from Polio (2013), which was a revised version of the Jacobs et al. (1981) analytic rubric. The revised rubric is a five-category analytic rubric that contains the following categories, with their possible points in parentheses: content (20), organization

(20), vocabulary (20), language use (20), and mechanics (10). When Polio revised this rubric, she changed the number of words per category and point values to make each category more equally balanced than the original (1981) rubric.

In order to adapt the rubric for the current study, I modified the Polio (2013) rubric in the following ways: first, I increased the points in the Mechanics category from 10 to 20, thus making each category worth 20 points. I made this change in order to give equal emphasis and value to each category. Second, based on the expert raters' feedback, I included two new traits on the rubric (*capitalization* in the Mechanics category and *relationship to the prompt* in the Content category). I included these traits because the expert raters admitted to using them as criteria for rating even though they were not present in the rubric; the expert raters often discussed these traits in rating sessions and these traits clearly influenced score outcomes. Therefore, I included them on the rubric. Third, in order to reflect the typed nature of the essays, I modified the *layout* trait to include accepted formatting for typed essays (*appropriate layout with well-defined paragraph separation*) as opposed to handwritten conventions (*appropriate layout with indented paragraphs*). Fourth, for clarity, I modified two descriptors in the Language Use category; I revised *occasional errors in awkward order or complex structures* to *occasional errors in word order or complex structures,* and I revised *attempts, even if not completely successful, at a variety of complex structures* to *a variety of complex structures, even if not completely successful.* The group of three experienced raters agreed that these content modifications made the rubric clearer, easier to use, and more fitting for use with typed essays. Finally, in order to ensure optimal recording on the eye-tracking computer, I made slight modifications to the layout of the rubric; I increased the space between score bands (vertically) and increased the width of the boxes to make later fixations on the rubric categories easier to

differentiate.

During experimental rating, the participants viewed the rubric on a Tobii TX300 eye-tracking computer. The Tobii screen displayed only the rubric. When presented on the eye tracker, the rubric measured approximately 11 inches by 14 inches, with an approximate font size of 14 points.

Standard Rubric. The standard rubric (SR) has categories presented in the following order: content, organization, vocabulary, language use, and mechanics. The standard rubric can be found in Appendix C.

Reordered Rubric. The reordered rubric (RR) is identical in content and format to the standard rubric, except that the categories appear in the following order: mechanics, language use, vocabulary, organization, and content. Including this reordered rubric in the study allows for a comparison of rater behavior in rubric use based on category position on the rubric. The changed positions of the categories will allow an investigation of whether attentional focus on rubric categories is linked to the position on the rubric (i.e., primacy), or whether it is linked to the category itself. This order is a mirrored flip of the SR order, which will allow for statistical comparison of the effect of moving each category an equal distance relative to their position on the SR, using vocabulary as a control. The reordered rubric can be found in Appendix D.

Rater Training. During rater training, I trained the participants on the rubric and benchmark essays. The session included an orientation to the rubric, an introduction to and discussion of benchmark essays, and a round of practice scoring accompanied by group discussion. I modeled this protocol after the standard rater-training protocol at the English Language Center (ELC), but with special care to spend equal time focusing on each category of the rubric during rubric orientation and essay discussions, thus controlling for the time spent on

each category during the training.

Rubric orientation. In order to carefully control attention to the rubric categories during rubric orientation, I delivered the rubric category introductions and explanations via a video-based orientation created using Camtasia. In the video, I focused on each category an equal amount of time, relative to the total number of words in the category. I created separate training videos for the Standard Rubric and the Reordered Rubric; the video content was identical, but it presented each category in an order that reflected the category appearance (from left to right) on the rubric.

Rater training benchmark essays. The English Language Center Testing Office provided me with the current set of benchmark essays used for operational testing for the Michigan State University English Language Test (MSUELT). From this set, I created annotation for them that highlighted the category descriptors within the essays and provided the official scores for each category. The annotations were specific comments on essay features, which were in comment boxes in the margin. I ensured that there were an equal number of comments on each category across the set of benchmark essays. I created these annotations based on a discussion session about the benchmark essays with the group of experienced raters. See Appendix E for an example.

I used these annotated benchmark essays to give participants exemplar essays at different score points and to provide concrete examples of essay descriptors within a student essay. Based on these benchmark essays, I led a discussion about the essay characteristics, the score points, and their relationship to rubric descriptors. I also allowed participants to ask any questions they had about the essays, the assigned scores, and the essay's relationship to the rubric.

Rater training norming essays. Similar to the benchmark essays, the Testing Office also

provided me with the current set of norming essays for operational testing for the MSUELT. With this set, I annotated the essays in the same manner as the annotated benchmark essays. The norming set likewise included specific instances of descriptor exemplifications and official scores for each category. I developed these annotations from the experienced-rater discussion about the norming essays.

I also created a clean set of the norming essays (without any commentary or scores) for practice rating, and the annotated set was provided to participants after they rated each essay on their own. This gave participants the opportunity to compare their scores with the experienced rater scores.

Criteria Importance Survey. The criteria importance survey (CIS) is a Likert-scale questionnaire that I modeled after the criteria importance questionnaire used by Eckes (2008). For this questionnaire, I asked participants to mark on a ten-point Likert scale how important they thought each criterion was when scoring an essay. Each criterion was taken from the analytic rubric used in this study (with only minor modifications to clarify descriptors in context). The CIS is included in Appendix F. I administered the CIS eight times throughout the study, and each time, the items were presented in a different, semi-randomized order. Table 5 summarizes the time points for each CIS administration.

Criteria Recall Task. The criteria recall task (CRT) is a free-recall task that presents raters with a completely blank rubric; that is, I gave the participants a rubric in which the rubric content was deleted, leaving only a rubric shell. Participants filled in as many descriptors and titles as they could remember. Through this task, I measured what participants recalled from the rubric without any help from other sources, and it provided additional insight into participants' memory of rubric categories and category descriptors. The CRT can be found in Appendix G, and

Table 5

Summary of CIS Administrations

Round	Administration	Description	Rubric
1	1	Pre-rater training	1
	2	Post-rater training	1
	3	Pre-rating	1
	4	Post-rating	1
2	5	Pre-rater training (not yet exposed to new rubric)	1
	6	Post-rater training	2
	7	Pre-rating	2
	8	Post-rating	2

Note. Rubric 1 is SR for Group A and RR for Group B. Rubric 2 is RR for Group A and SR for Group B.

Table 6 summarizes the time points for each CRT administration.

To prepare the data for analysis, I formed a coding scheme that identified whether or not the participant provided an acceptable representation of each descriptor and category title from the rubric. Using this coding scheme (see Appendix H), one rater coded each CRT on whether the participant satisfactorily reproduced each individual descriptor or category title. After coding, I counted the number of rubric components (category titles and descriptors) each participant appropriately produced for each rubric category. I summed these numbers and converted them into a percentage of descriptors provided per category (%= number of elements provided by the participant/number of elements in each category), hereafter referred to as the *recall accuracy score*. A second rater coded 42% of the data, and the Cronbach's alpha for interrater reliability was .99 (CI 95% .989, .990).

Table 6

Summary of CRT Administrations

Round	Administration	Description	Rubric
1	1	Post-rater training	1
	2	Pre-rating	1
2	3	Pre-rater training (not yet exposed to new rubric)	1
	4	Post-rater training	2
	5	Pre-rating	2

Note. Rubric 1 is SR for Group A and RR for Group B. Rubric 2 is RR for Group A and SR for Group B.

Rubric Reorientation Task. Before rating, participants refamiliarized themselves with the rubric. For all raters, their eye movements were recorded to measure how much time they spent reviewing the descriptors within each category. All raters completed this task on the rubric they used for the actual essay rating (i.e., standard rubric or reordered rubric). This also included a practice rating to orient participants to the procedure and have an opportunity to ask questions before beginning the rating of the set of 20 essays.

Decision-making Process Outline. In the decision-making process outline (DMPO), raters made a step-by-step written outline of how they approached essay rating during the rating session. Each rater typed out their outline in a Word document.

Rater Interviews. At the end of the rating session, I conducted semi-structured interviews with each participant. I asked the participants to explain their decision-making process. I probed into what each rater focused on as they read an essay, used the rubric, and arrived at a score. The interview questions are in Appendix I.

Background Questionnaire. Each participant completed a background questionnaire about their education, language-learning experience, and rating experience. I adapted this questionnaire from Winke and Lim (2015). The questionnaire is presented in Appendix J.

Analysis

Raters' beliefs about criteria importance

In analyzing the CIS data, I sought to investigate whether the rubric that a rater trained on, that is, the order that they encountered each category, affected how important the rater considered the (descriptors in each) category to be. I employed a repeated-measures analysis of variance (RM ANOVA) to uncover any differences in category importance over time and between the groups.

I collected CIS data at eight time points during the study (see Table 5); the first five are based on the rubric of initial exposure (i.e., SR for Group A and RR for Group B), and the last three are based on rubric of second exposure (i.e., RR for Group A and SR for Group B). Because the repeated-measures, cross-over design would make it difficult to meaningfully analyze all the data in one model, I selected specific administrations of the CIS (Round 1 Pre-rating, Round 2 Pre-training, and Round 2 Pre-rating) to analyze; I selected the administrations that (1) optimally demonstrated ordering effects, if ordering effects were present, and (2) overlapped with administrations of the CRT (i.e., instances where the participants completed a CRT and CIS consecutively). Round 1 Pre-rating provided a snapshot of the participants' beliefs about criteria importance after a short-term delay in exposure to the initial rubric (i.e., one- to two-day lapse since exposure); Round 2 Pre-training provided a snapshot of a long-term delay in exposure to the initial rubric (i.e., five-week lapse since exposure), and Round 2 Pre-rating provides a snapshot of a short-term delay in exposure to the second rubric (i.e., one- to two-day lapse since exposure). When paired with the corresponding CRT analyses results, these specific CIS data provided a clearer, more comprehensive picture of the raters' mental-rubric formation and how the raters' mental-rubric formation was affected by category-ordering effects.

For the CIS data, I first inspected the descriptive statistics for each of the five categories at the eight time points for the two groups to understand general trends in the data. Next, I examined the CIS data through RM ANOVA to uncover differences in category-importance beliefs over time and between the two groups. I computed separate RM ANOVAs for the two rubric periods (Rubric 1: rubric of first exposure/training; Rubric 2: rubric of second exposure/training). For the first rubric period, I computed a 2 (Group: A, B) x 2 (Time: Round 1 Pre-rating, Round 2 Pre-training) x 5 (Category: Content, Organization, Vocabulary, Language Use, Mechanics) RM ANOVA. For the second period, I computed a 2 (Group: A, B) x 5 (categories) RM ANOVA at one time point (Round 2 Pre-rating). I used Bonferroni adjustments in all analyses to account for multiple analyses on the same data set (with a significant *p* value equal to or less than .025). I examined the data for a normal distribution, equal variances, sphericity, distribution of the residuals, and equal variances of the residuals. The data were neither normally distributed nor had equal variances. Mauchly's test of sphericity was statistically significant ($p < .05$) which indicated that compound symmetry in the variance-covariance matrix was not present, and an examination of the residual SSCP matrix revealed a violation of the sphericity assumption. Given this violation, I report a Greenhouse-Geisser correction, which is used when sphericity is not present (see Field, 2009). I did not find any important deviations from normality or homogeneity of variances of the residuals of the data entered into these models.

Raters' criteria recall

In analyzing the CRT data, I sought to investigate whether the rubric that raters trained on, that is, the order in which they encountered each category, affected their memory of (the descriptors in) each category. As with the CIS data, I employed RM ANOVA to look at

differences in category memory over time and between the groups, who trained on different rubrics.

I collected the CRT data at five time points during the study (see Table 6); the first three are based on the rubric of initial exposure (i.e., SR for Group A and RR for Group B), and the last two are based on rubric of second exposure (i.e., RR for Group A and SR for Group B). Again, because the repeated-measures cross-over design makes it difficult to meaningfully analyze all the data in one model, I selected specific administrations of the CRT (Round 1 Pre-rating, Round 2 Pre-training, and Round 2 Pre-rating) to analyze; I selected the administrations based on the same selection criteria I used with the CIS data. Namely, the CRT (1) should optimally demonstrate ordering effects, if the ordering effects were present, and (2) should overlap with administrations of the CIS (i.e., instances where the participants completed a CRT and CIS consecutively). Round 1 Pre-rating provided as a snapshot of the participants' memory after a short-term delay in exposure to the initial rubric (i.e., one- to two-day lapse since exposure). Round 2 Pre-training provided a snapshot of a long-term delay in exposure to the initial rubric (i.e., five-week lapse since exposure). And Round 2 Pre-rating provided a snapshot of a short-term delay in exposure to the second rubric (i.e., one- to two-day lapse since exposure). As above, when paired with the corresponding CIS analyses results, the specific CRT data revealed that the raters' mental-rubric formation and shed light on how category-ordering effects influenced the mental rubric formation.

For the CRT data, I first inspected the descriptive statistics for each of the five categories at the five time points for the two groups to understand general trends in the data. Next, I examined the CRT data through RM ANOVA to uncover differences in category memory over time and between the two groups. I computed separate RM ANOVAs for the two rubric periods

(Rubric 1: rubric of first exposure/training; Rubric 2: rubric of second exposure/training). For the first rubric period, I computed a 2 (Group: A, B) x 2 (Time: Round 1 Pre-rating, Round 2 Pre-training) x 5 (Category: Content, Organization, Vocabulary, Language Use, Mechanics) RM ANOVA, and for the second period, I computed a 2 (Group: A, B) x 5 (categories) RM ANOVA at one time point (Round 2 Pre-rating). I used Bonferroni adjustments in all analyses to account for multiple analyses on the same data set (with a significant *p* value equal to or less than .025). I examined the data for a normal distribution, equal variances, sphericity, a good distribution of the residuals, and equal variances of the residuals. The data were neither normally distributed nor had equal variances. Mauchly's test of sphericity was not statistically significant ($p > .05$), and examination of the residual SSCP matrix did not reveal a notable violation of the sphericity assumption. However, I did not find any important deviations from normality or homogeneity of variances of the residuals of the data entered into these models, so I proceeded with the analysis.

Primacy and raters' order of attention

To uncover the order in which raters attended to the rubric categories and to see how the primacy effect may have played a role in rubric-use behavior, I analyzed the time to first fixation (TFF) data. The eye-movement data were recorded by the eye-tracking computer during the two rating sessions. To analyze the TFF data, I first examined the TFF data for each participant on each rubric. If a TFF value was below 45 seconds, I visually examined the corresponding recording to determine whether the fixation was accidental (a random, quickly-passing fixation) that suggested that the rater was not reading the text in the category. If the fixation was of this nature, I manually adjusted the recording segment time to count later fixations as the true first fixations for each category (this is the way Tobii support recommended cleaning the data; personal communication, February 2017). Next, I exported the TFF value for each participant on

each of the 20 rubrics at the two time points (Round 1 and Round 2). I visually inspected the data for trends, and then collapsed the data into one mean TFF value for each participant for each of the five categories in each of the two rounds. I then examined the descriptive statistics for each group, replacing any extreme outliers with the mean plus two standard deviations. I present the descriptive statistics below.

Rubric category importance and the primacy effect

I sought to uncover how the primacy effect may have played a role in raters' attention to the rubric categories. To do this, I investigated how much total time raters spent fixating on (i.e., reading) each rubric category and how many times each rater visited (read within) a category. To do this, I analyzed two types of eye-movement data: the total fixation duration (TFD) data and the visit count (VC) data. Next I answer how I specifically used both sets of data to investigate primacy and raters' rubric category attention and attentional behaviors.

Raters' concentrated attention (measured via TFD) to the rubric categories

To prepare the TFD data for analysis, I followed the same collapsing procedure outlined for TFF. Next, I computed a mean difference TFD value (TFD[SR] – TFD[RR]) for each participant and for each category. Then, because each of the categories had a different number of words, which may have resulted in different fixation times based on word count alone, I computed a controlled TFD value (as done by Winke & Lim, 2015, and by McCray & Brunfaut, 2016) that takes the number of words in each category into account. I divided the mean difference TFD value by the number of words in each category, thus creating a TFD value that is comparable across categories.

To analyze the data, I first conducted a 2 (Group) x 5 (Category) RM ANOVA in order to uncover whether the two groups had different TFD times for each category. Next, I conducted a

one-sample *t* test for each category and for each group in order to determine whether the groups' TFD for each category was statistically different between rounds (e.g., Content Round 1 and Content Round 2), which I calculated by comparing the mean difference TFD value to zero. I used Bonferroni adjustments in all analyses (with the *p* value set to .005) to account for multiple analyses on the same data set.

Before running any inferential statistical tests, I examined the descriptive data and the assumptions for each test. As discussed above, I replaced the value of any extreme outliers with the mean plus two standard deviations. After this adjustment, I examined whether the data met the statistical test assumptions (homogeneity of variance, normal distribution, linearity, etc.). For the RM ANOVA, Mauchly's test of sphericity was statistically significant ($p < .05$), and an examination of the residual SSCP matrix revealed a violation of the sphericity assumption. Given this violation, I report the Greenhouse-Geisser statistic, which is used when sphericity is not assumed. I also examined the residuals, and there were no strong violations of the assumptions.

Raters' frequency of attention (measured via VC) to the rubric categories

In analyzing the VC data, I sought to uncover how many separate visits the raters made to each of the rubric categories and how the primacy effect may play a role in this behavior.

I followed the same data-collapsing procedure outlined for TFF. I computed a mean difference VC value (VC[SR] – VC[RR]) for each participant for each category. To analyze the data, I first conducted a 2 (Group) x 5 (Category) RM ANOVA to uncover whether the two groups had different visit counts for each category. Next, I conducted a one-sample *t* test for each category and for each group in order to determine whether the VC for each category was statistically different between rounds (e.g., Content Round 1 and Content Round 2), which is

calculated by comparing the mean difference VC value to zero. Again, I used Bonferroni adjustments (*p* value = .005) in all analyses to account for multiple analyses on the same data set.

Before running any inferential statistical tests, I examined the descriptive data and the assumptions for each test. I replaced the value of any outliers with the mean plus two standard deviations. After this adjustment, I examined whether the data met the assumptions for the statistical tests (homogeneity of variance, normal distribution, linearity, etc.). For the RM ANOVA, Mauchly's test of sphericity was statistically significant ($p < .05$), and an examination of the residual SSCP matrix revealed a violation of the sphericity assumption. Given this violation, I report the Greenhouse-Geisser statistic, used when sphericity is not assumed. I also examined the residuals, and there were no strong violations of the assumptions.

In addition to visit count, I provide descriptive statistics for the number of times raters skipped (i.e., did not read) a category (as done in Winke & Lim, 2015), henceforth referred to as *category skipping*. From these data, I computed a mean category-skipping value and a mean difference category-skipping value, which I calculated by subtracting Round 1 mean category values from Round 2 mean category values. I present the descriptive statistics for category skipping below.

Interrater reliability

In the Winke and Lim's (2015) study, the authors found primacy effects by investigating the raters' scoring behavior as measured through eye-tracking metrics and intra-rater reliability measures. To compare rater behavior (through reliability estimates) to that of Winke and Lim, I computed the same reliability statistic, Intraclass Correlations (ICC). Intra-rater reliability is an estimation of rater scoring consistency, which is calculated by correlating two sets of scores produced by the same rater for the same examinees (Brown, 2005, p. 288). I calculated the ICC

statistic for each individual rubric category using IBM's Statistical Package for Social Sciences (SPSS) 24.

Rater severity

To investigate raters' scoring behavior while using each rubric, I conducted Multi-Faceted Rasch Measurement (MFRM). This measurement model offers the possibility to examine the impact of multiple factors, called facets, on raw test-score outcomes. MFRM estimates how a given facet (e.g., rubric, rater, task difficulty, etc.), entered by the researcher, contributes to score variation. These estimations are provided in the same model on a common scale, known as a logit scale, so that the impact of different facets can be compared (Bond & Fox, 2013). MFRM is also useful for estimating interactions between two or more facets to investigate how different facets (e.g., rubric and rubric category) together may impact score variation (Linacre, 1989). In short, MFRM allows test administrators to investigate how various factors influence score outcomes, and crucially, it is particularly useful for investigating rater behavior through scoring data.

In the present study, I used a fully-crossed design, meaning that every rater scored all 20 essays. To analyze the data, I used Facets 3.71.4 (Linacre, 2014). I specified six facets for the analyses: essays (20), raters (31), round (Round 1 and Round 2), rubric (SR and RR), group (Group A and Group B), and rubric category (Content, Organization, Vocabulary, Language Use, and Mechanics).

I originally modeled the rater scoring data on the full 20-point scale for each rubric category. However, Facets yielded many disordered average measures for score points within score bands (1-5, 6-10, 11-15, 16-20), and my initial analysis also revealed that there were fewer than 10 observations at many of the high (16-20) and low (1-5) score points. Thus, I recoded the

data in a 5-point scoring scale by collapsing the score points (as outlined by Eckes, 2011; Linacre, 2010; see also Janssen, Meier, and Trace [2015], who encountered similar score-point distinction problems with such a wide point range on the original Jacobs et al. [1981] rubric) as follows: 1 - 5 = 1; 6 – 10 = 2; 11 – 12 = 3; 13 – 15 = 4; and 16 – 20 = 5. I recoded each score band into one score point, except for the 11-15 band. I coded the score points in this band into two score points (3 and 4) because most of the observations fell within this band, and the Andrich thresholds (i.e., the boundaries between two score points) for score points 13, 14, and 15 suggested that these scores were more similar to one another (i.e., the thresholds were closer to one another) than different. In addition, conceptually, score point 13 was the transition point into the higher end of the score band, and thus it made theoretical sense to group score points 13 to 15 together.

Following data recoding, I conducted five Partial Credit Model MFRM analyses. The Partial Credit Model (PCM), unlike the Rating Scale Model, does not assume that the steps of a scale (i.e., score points) are equivalent across all elements of a given facet. The PCM allows the score points of different components of the rating scale to be estimated separately, thus showing differences in scale use between rubric categories. This is particularly important in the current study, as the PCM makes it possible to investigate differences in rater severity for each individual rubric category.

For each analysis, I designated the essay facet as non-centered and all other facets as centered. I also designated the essay facet as positively oriented, thus indicating that higher logit values signify higher ability of the essay writer.

For each data set, I inspected rater fit statistics, which provide information about whether raters are scoring as expected by the model. Misfitting raters are raters who do no behave (i.e.,

score) like other raters, and therefore are flagged for inspection. Because I am interested in looking at the effects of rubric category order on rater scoring behavior (and may expect to see some degree of misfitting raters), I inspected the rater fit statistics but did not remove any rater from the analyses because removing misfitting raters might have resulted in a reduction of variance related to rubric order.

In the first two analyses, I implemented a PCM to investigate rater severity within each category between the two rubrics. In the first analysis, I examined the differences between the SR and RR in Round 1, and in the second I examined the data from Round 2. In these models, I anchored the Group facet at zero logits to ensure connectivity in the data set. Due to the between-subjects design of the study, not anchoring a facet would otherwise lead to two disjointed subsets. Differences in category severity between the two rubrics would suggest that category order may have affected the raters' scoring behavior. I entered the PCM models as follows:

Model 1=
?,?,1,?B,,#B,R5
; Essay, Rater, Round, Group, Rubric, Category (Group/Category interaction)

Model 2=
?,?,2,?B,,#B,R5
; Essay, Rater, Round, Group, Rubric, Category (Group/Category interaction)

where *?* indicates that the facet was free to vary, *1* indicates that only Round 1 data was included in the model, *2* indicates that only Round 2 data was included in the model, *#* indicates that the Andrich thresholds were free to vary for each rubric category, *B* indicates that a bias term (i.e., interaction) is to be calculated between facets that include this code, and *R5* indicates that the highest score-point value allowed in the data was 5.

In the third and fourth analyses, I examined rater severity for each category within the same rubric between the two Rounds. In analysis three, I compared SR Round 1 with SR Round

2, and in analysis four, I compared RR Round 1 with RR Round 2. In these models, I anchored the Round facet at zero logits to ensure connectivity in the data set. Differences in rater severity for a given category between the two rubrics would suggest that the order of rubric exposure affected the raters' scoring behavior. I entered these PCM models as follows:

Model 3=
?,?,?,?B,1,#B,R5
; Essay, Rater, Round, Group, Rubric, Category (Group/Category interaction)

Model 4=
?,?,?,?B,2,#B,R5
; Essay, Rater, Round, Group, Rubric, Category (Group/Category interaction)

where *?* indicates that the respective facet was free to vary, *1* indicates that only the SR data was included in the model, *2* indicates that only RR data was included in the model, *#* indicates that the Andrich thresholds were free to vary for each rubric category, *B* indicates that a bias term (i.e., interaction) is to be calculated between facets that include this code, and *R5* indicates that the highest score-point value allowed in the data was 5.

My final MFRM analysis included the entire data set because my goal was to investigate rater severity for each category between the two rubrics and across the two rounds. In this model, I anchored the Rubric facet at zero logits to ensure connectivity in the data set. Here, any differences in rater severity for a given category would suggest an overall difference in raters' scoring behavior between the two rubrics. I entered the PCM models as follows:

Model 5=
?,?,?,?,?B,#B,R5
; Essay, Rater, Round, Group, Rubric, Category (Rubric/Category interaction)

where *?* indicates that the respective facet was free to vary, *#* indicates that the Andrich thresholds were free to vary for each rubric category, *B* indicates that a bias term (i.e., interaction) is to be calculated between facets that include this code, and *R5* indicates that the

highest score-point value allowed in the data was 5.

CHAPTER 3

RESULTS: MENTAL-RUBRIC FORMATION

To answer the first research question (To what extent do raters show evidence of ordering effects in their mental rubric formation?), I examined the Criteria Importance Survey (CIS) data and the Criteria Recall Task (CRT) data. I administered the CIS to measure how important each rater considered each rubric descriptor to be and to understand each rater's beliefs about the importance of each category as a whole. I administered the CRT to measure how many descriptors each rater remembered from each rubric category, thus indicating how present and important each category was in each rater's mental rubric. Taken together, these data provide a snapshot of the raters' mental-rubric formation (i.e., what is included in their mental rubric and how important they consider each piece of that rubric to be).

Results

Criteria Importance

Descriptive statistics for the CIS data are in Table 7, Figure 2, and Figure 3. On this Likert scale, the options were labeled as follows: 1 = unimportant, 4 = somewhat important, 7 = important, and 10 = very important (see Appendix F for a sample survey). Looking at general trends in the descriptive data, Group A starts with a lower, wider spread in importance ratings than does Group B. At Time 5 (after the five-week lapse), which is the most likely time for ordering effects to manifest, Group A's importance ratings replicate the general order of the Standard rubric, with participants rating the left-most categories (Content and Organization) as being most important, the right-most categories (Language Use and Mechanics) as least important, and vocabulary in the middle. Group B's ratings, however, have a less clear spread

between categories, with participants rating Content and Organization as the most important, and Vocabulary, Language Use and Mechanics clustered together slightly lower.

Another general trend shows that, for both groups, the ratings for all categories appear relatively similar within groups, with the majority of mean category ratings (36 of 40 for Group A and 40 of 40 for Group B) falling between 7 and 9 on the importance scale. Participants rated Organization and Content as most important after the groups trained on the initial rubric (Time 2-8). Between Time 5 and 6, in which participants trained on the new rubric, there is a slight increase in importance scores for Language Use and Mechanics for both groups. Additionally, comparing Time 3 and Time 7 (both pre-rating administrations), each group maintained similar beliefs about category importance between the two time points, even though the CIS administrations were based on different rubrics. Another trend is that for Group A, Mechanics is rated lowest for importance, but for Group B, Mechanics falls more toward the middle of the importance ratings.

For the first RM ANOVA model (including Round 1 Pre-rating and Round 2 Pre-training), the three-way interaction between Group, Time, and Category was not statistical per the Greenhouse-Geisser statistic ($F_{2, 56} = 1.060$, $p = .359$, $\eta^2_P = .041$), nor was the two-way interaction between Group and Category ($F_{3, 67} = 1.539$, $p = .216$, $\eta^2_P = .058$). For main effects, Category ($F_{3, 67} = 1.539$, $p < .001$, $\eta^2_P = .369$) was statistically significant, but Time ($F_{1, 25} = 0.160$, $p = .692$, $\eta^2_P = .006$) and Group were not ($F_{1, 25} = 0.123$, $p = .729$, $\eta^2_P = .005$).

Pairwise comparisons of Category for Group A at Time 3 (Round 1 Pre-rating) revealed that only Organization and Mechanics were statistically different from one another (mean difference= 1.013, $p = .006$, 97.5% CI [0.148, 1.878]). At Time 5 (Round 2 Pre-training), Organization was statistically different from Vocabulary (mean difference= 0.883, $p = .014$,

97.5% CI [0.054, 1.712]) and Mechanics (mean difference= 1.648, p = .002, 97.5% CI [0.387, 2.910]).

For Group B at Time 3, only Content and Organization were statistically different (mean difference= -0.656, p = .020, 97.5% CI [-1.295, -0.017]), and at Time 5, no categories were statistically different from one another.

Table 7

Criteria Importance Survey (CIS) Means

Round	Time	Group	Content	Organization	Vocabulary	Language Use	Mechanics
1	1 (pre-training)	A	7.7 (1.1)	7.3 (1.5)	5.7 (0.8)	6.4 (1.3)	6.5 (1.7)
		B	7.3 (1.2)	8.1 (1.1)	7.0 (0.9)	7.3 (1.4)	7.5 (1.3)
	2 (post training)	A	8.4 (1.2)	8.6 (1.2)	7.6 (1.4)	8.0 (1.2)	7.4 (1.3)
		B	8.3 (1.4)	8.7 (1.3)	7.8 (1.7)	8.2 (1.6)	7.9 (1.8)
	3 (pre-rating)	A	8.2 (1.3)	8.4 (1.1)	7.7 (1.4)	7.9 (1.4)	7.4 (1.2)
		B	7.9 (1.3)	8.5 (1.0)	7.8 (1.7)	7.9 (1.4)	7.8 (1.3)
	4 (post-rating)	A	8.2 (1.2)	8.8 (1.1)	7.5 (1.4)	7.9 (1.5)	7.3 (2.0)
		B	8.4 (1.3)	8.8 (1.3)	8.1 (1.5)	8.6 (1.1)	8.6 (1.2)
2	5 (pre-training)	A	8.1 (1.0)	8.5 (1.2)	7.6 (1.4)	7.6 (1.5)	6.9 (2.0)
		B	8.4 (1.2)	8.5 (1.4)	7.8 (2.0)	7.9 (1.9)	7.7 (2.2)
	6 (post-training)	A	8.3 (1.4)	8.6 (1.2)	7.8 (1.5)	8.3 (1.2)	7.4 (1.8)
		B	8.3 (1.2)	8.8 (1.2)	7.9 (1.8)	8.3 (1.4)	8.0 (1.5)
	7 (pre-rating)	A	8.4 (1.2)	8.6 (1.3)	7.8 (1.4)	7.7 (1.5)	7.5 (1.6)
		B	8.1 (1.4)	8.8 (1.3)	8.0 (1.8)	8.0 (1.6)	8.0 (1.7)
	8 (post-rating)	A	8.5 (1.3)	8.7 (1.3)	7.8 (1.3)	7.9 (1.5)	7.5 (1.5)
		B	8.5 (1.3)	8.9 (1.3)	8.1 (1.7)	8.2 (1.7)	8.3 (1.5)

Note. Standard deviations are in parentheses. Group A: n = 13. Group B: n = 14.

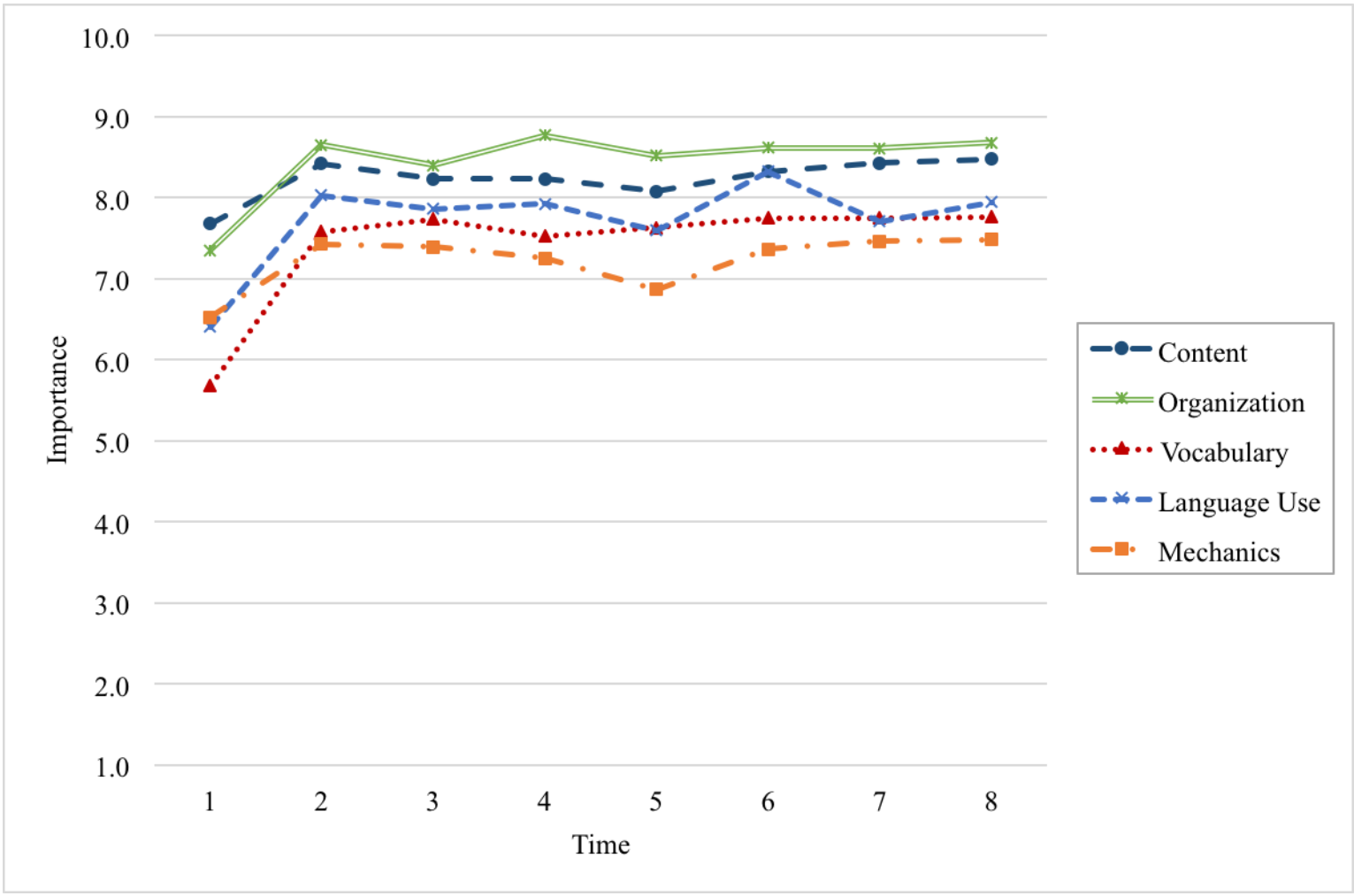


Figure 2. Criteria Importance Survey means for Group A.

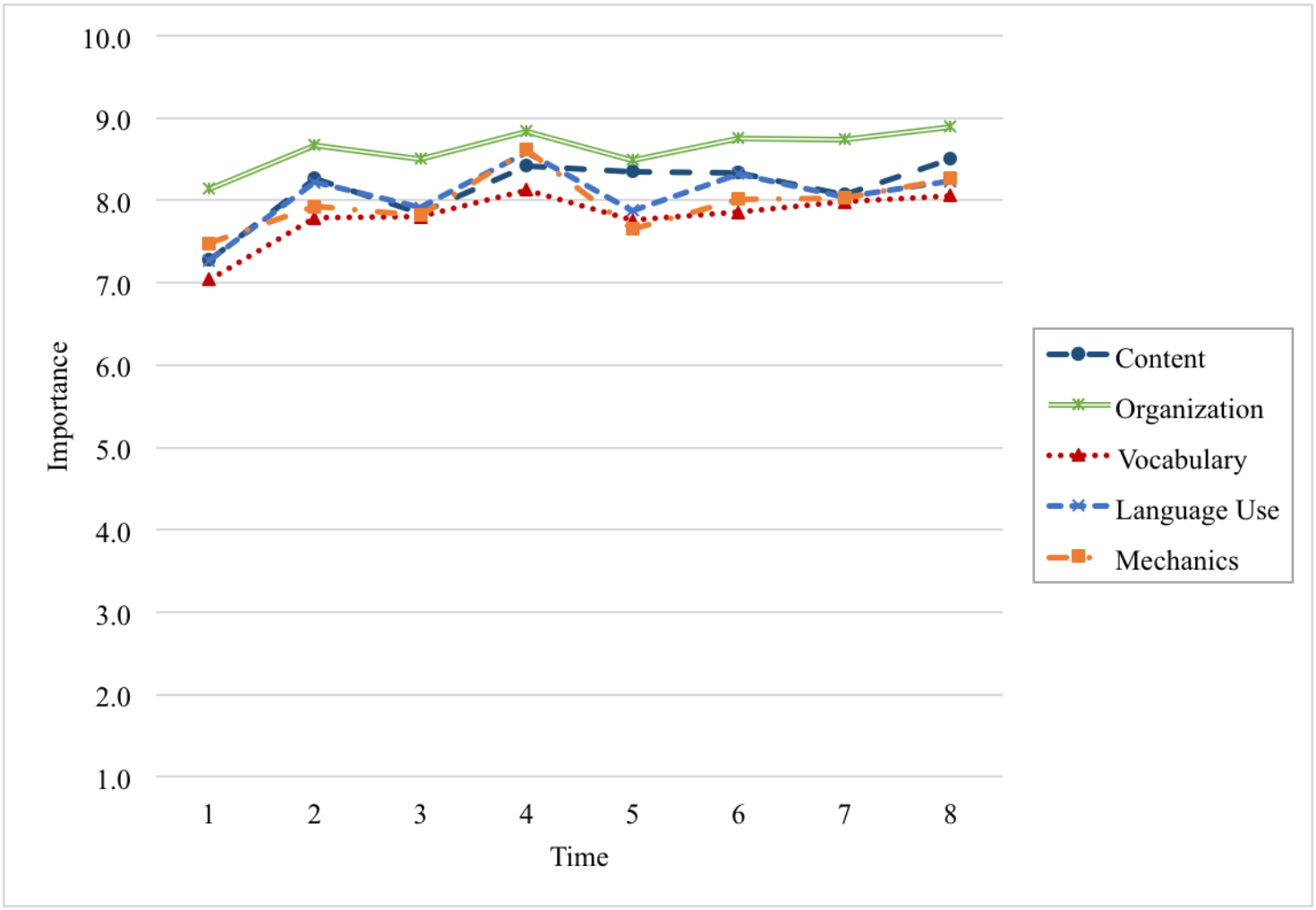


Figure 3. Criteria Importance Survey means for Group B.

Table 8

Category Pairwise Comparisons for Category Importance

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 3 (Round 1 Pre-rating)	Con.	Org.	-0.17	0.20	1.00	-0.83	0.50	-0.66*	0.19	0.02	-1.29	-0.02
		Vocab.	0.49	0.17	0.07	-0.07	1.06	0.06	0.16	1.00	-0.49	0.60
		Lang. Use	0.37	0.15	0.24	-0.15	0.89	-0.06	0.15	1.00	-0.56	0.44
		Mech.	0.85	0.26	0.03	-0.03	1.72	0.04	0.25	1.00	-0.81	0.88
	Org.	Con.	0.17	0.20	1.00	-0.50	0.83	0.66*	0.19	0.02	0.02	1.29
		Vocab.	0.66	0.25	0.16	-0.20	1.51	0.71	0.25	0.08	-0.11	1.54
		Lang. Use	0.54	0.19	0.10	-0.11	1.18	0.60	0.19	0.04	-0.03	1.22
		Mech.	1.01*	0.26	0.01	0.15	1.88	0.69	0.25	0.10	-0.14	1.53
	Vocab.	Con.	-0.49	0.17	0.07	-1.06	0.07	-0.06	0.16	1.00	-0.60	0.49
		Org.	-0.66	0.25	0.16	-1.51	0.20	-0.71	0.25	0.08	-1.54	0.11
		Lang. Use	-0.12	0.18	1.00	-0.72	0.47	-0.11	0.17	1.00	-0.69	0.46
		Mech.	0.35	0.30	1.00	-0.65	1.36	-0.02	0.29	1.00	-0.99	0.94
	Lang. Use	Con.	-0.37	0.15	0.24	-0.89	0.15	0.06	0.15	1.00	-0.44	0.56
		Org.	-0.54	0.19	0.10	-1.18	0.11	-0.60	0.19	0.04	-1.22	0.03
		Vocab.	0.12	0.18	1.00	-0.47	0.72	0.11	0.17	1.00	-0.46	0.69
		Mech.	0.48	0.27	0.94	-0.44	1.40	0.09	0.26	1.00	-0.79	0.98
	Mech.	Con.	-0.85	0.26	0.03	-1.72	0.03	-0.04	0.25	1.00	-0.88	0.81
		Org.	-1.01*	0.26	0.01	-1.88	-0.15	-0.69	0.25	0.10	-1.53	0.14
		Vocab.	-0.35	0.30	1.00	-1.36	0.65	0.02	0.29	1.00	-0.94	0.99
		Lang. Use	-0.48	0.27	0.94	-1.40	0.44	-0.09	0.26	1.00	-0.98	0.79

Table 8 (cont'd)

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 5 (Round 2 Pre-training)	Con.	Org.	-0.44	0.18	0.26	-1.06	0.18	-0.14	0.18	1.00	-0.74	0.45
		Vocab.	0.45	0.28	1.00	-0.49	1.39	0.60	0.27	0.35	-0.31	1.51
		Lang. Use	0.48	0.31	1.00	-0.56	1.52	0.48	0.30	1.00	-0.52	1.49
		Mech.	1.21	0.40	0.06	-0.15	2.57	0.70	0.39	0.86	-0.61	2.00
	Org.	Con.	0.44	0.18	0.26	-0.18	1.06	0.14	0.18	1.00	-0.45	0.74
		Vocab.	0.88*	0.25	0.01	0.05	1.71	0.74	0.24	0.04	-0.06	1.54
		Lang. Use	0.92	0.28	0.03	-0.02	1.86	0.62	0.27	0.29	-0.28	1.53
		Mech.	1.65*	0.38	0.00	0.39	2.91	0.84	0.36	0.29	-0.38	2.05
	Vocab.	Con.	-0.45	0.28	1.00	-1.39	0.49	-0.60	0.27	0.35	-1.51	0.31
		Org.	-0.88*	0.25	0.01	-1.71	-0.05	-0.74	0.24	0.04	-1.54	0.06
		Lang. Use	0.03	0.14	1.00	-0.43	0.50	-0.12	0.13	1.00	-0.56	0.33
		Mech.	0.77	0.27	0.08	-0.13	1.66	0.10	0.26	1.00	-0.76	0.96
	Lang. Use	Con.	-0.48	0.31	1.00	-1.52	0.56	-0.48	0.30	1.00	-1.49	0.52
		Org.	-0.92	0.28	0.03	-1.86	0.02	-0.62	0.27	0.29	-1.53	0.28
		Vocab.	-0.03	0.14	1.00	-0.50	0.43	0.12	0.13	1.00	-0.33	0.56
		Mech.	0.73	0.25	0.08	-0.12	1.58	0.21	0.24	1.00	-0.60	1.03
	Mech.	Con.	-1.21	0.40	0.06	-2.57	0.15	-0.70	0.39	0.86	-2.00	0.61
		Org.	-1.65*	0.38	0.00	-2.91	-0.39	-0.84	0.36	0.29	-2.05	0.38
		Vocab.	-0.77	0.27	0.08	-1.66	0.13	-0.10	0.26	1.00	-0.96	0.76
		Lang. Use	-0.73	0.25	0.08	-1.58	0.12	-0.21	0.24	1.00	-1.03	0.60

Table 8 (cont'd)

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 7 (Round 2 Pre-rating)	Con.	Org.	-0.18	0.19	1.00	-0.83	0.47	-0.68*	0.19	0.01	-1.30	-0.05
		Vocab.	0.67	0.21	0.03	-0.03	1.36	0.09	0.20	1.00	-0.58	0.75
		Lang. Use	0.72*	0.19	0.01	0.07	1.36	0.03	0.19	1.00	-0.59	0.65
		Mech.	0.96*	0.28	0.02	0.01	1.92	0.04	0.27	1.00	-0.88	0.96
	Org.	Con.	0.18	0.19	1.00	-0.47	0.83	0.68*	0.19	0.01	0.05	1.30
		Vocab.	0.85	0.28	0.05	-0.09	1.79	0.77	0.27	0.09	-0.14	1.67
		Lang. Use	0.90*	0.23	0.01	0.12	1.68	0.71	0.22	0.04	-0.04	1.46
		Mech.	1.14*	0.33	0.02	0.05	2.24	0.71	0.31	0.32	-0.34	1.77
	Vocab.	Con.	-0.67	0.21	0.03	-1.36	0.03	-0.09	0.20	1.00	-0.75	0.58
		Org.	-0.85	0.28	0.05	-1.79	0.09	-0.77	0.27	0.09	-1.67	0.14
		Lang. Use	0.05	0.18	1.00	-0.54	0.64	-0.06	0.17	1.00	-0.63	0.51
		Mech.	0.29	0.16	0.78	-0.24	0.83	-0.05	0.15	1.00	-0.57	0.47
	Lang. Use	Con.	-0.72*	0.19	0.01	-1.36	-0.07	-0.03	0.19	1.00	-0.65	0.59
		Org.	-0.90*	0.23	0.01	-1.68	-0.12	-0.71	0.22	0.04	-1.46	0.04
		Vocab.	-0.05	0.18	1.00	-0.64	0.54	0.06	0.17	1.00	-0.51	0.63
		Mech.	0.25	0.26	1.00	-0.64	1.13	0.01	0.25	1.00	-0.85	0.86
	Mech.	Con.	-0.96*	0.28	0.02	-1.92	-0.01	-0.04	0.27	1.00	-0.96	0.88
		Org.	-1.14*	0.33	0.02	-2.24	-0.05	-0.71	0.31	0.32	-1.77	0.34
		Vocab.	-0.29	0.16	0.78	-0.83	0.24	0.05	0.15	1.00	-0.47	0.57
		Lang. Use	-0.25	0.26	1.00	-1.13	0.64	-0.01	0.25	1.00	-0.86	0.85

Note. Based on estimated marginal means. *The mean difference is significant at the .025 level. b= Adjustment for multiple comparisons: Bonferroni.

Group B's ratings of importance for the five categories were much more similar to one another, with the only statistical difference being between Organization and Content. Interestingly, at both Time 3 and Time 7, there were statistical differences between Organization and the category that appeared last on rubric of initial training (Group A: Mechanics, and Group B: Content).

Criteria Recall

Descriptive statistics for all CRT data are in Figure 4, Figure 5, and Table 9. The general trends show that within groups, the groups performed similarly between Time 1 (Round 1 Post-training) and Time 2 (Round1 Pre-rating) and between Time 4 (Round 2 Post-rating) and Time 5 (Round 2 Pre-rating). Both groups were less accurate in their recall at Time 3 (Round 2 Pre-training, after a 5-week lapse in training), and demonstrated higher accuracy in Round 2 (Time 4 and 5) than Round 1 (Times 1 and 2) for each individual category. In general, for both groups, the descriptors in the Mechanics category were the easiest to remember (showing the highest scores), and Language Use, Content, and Organization were the most difficult to remember (showing the lowest scores), while Vocabulary fell toward the middle.

For the first repeated measures ANOVA model (including Round 1 Pre-rating and Round 2 Pre-training), the three-way interaction between Group, Time, and Category was not statistically significant, per the Greenhouse-Geisser statistic ($F_{3, 84} = 0.451$, $p = .738$, $\eta^2_P = .018$). The two-way interaction between Group and Category was statistically significant per the Greenhouse-Geisser statistic ($F_{3, 76} = 3.477$, $p = .020$, $\eta^2_P = .122$), showing that the groups remembered different amounts of certain categories. Pairwise comparisons for the Group*Category interaction revealed that Group A outperformed Group B at Time 3 for Content (mean difference= .18, $p = .002$, 97.5% CI [.058, .302]). No other comparisons were statistically

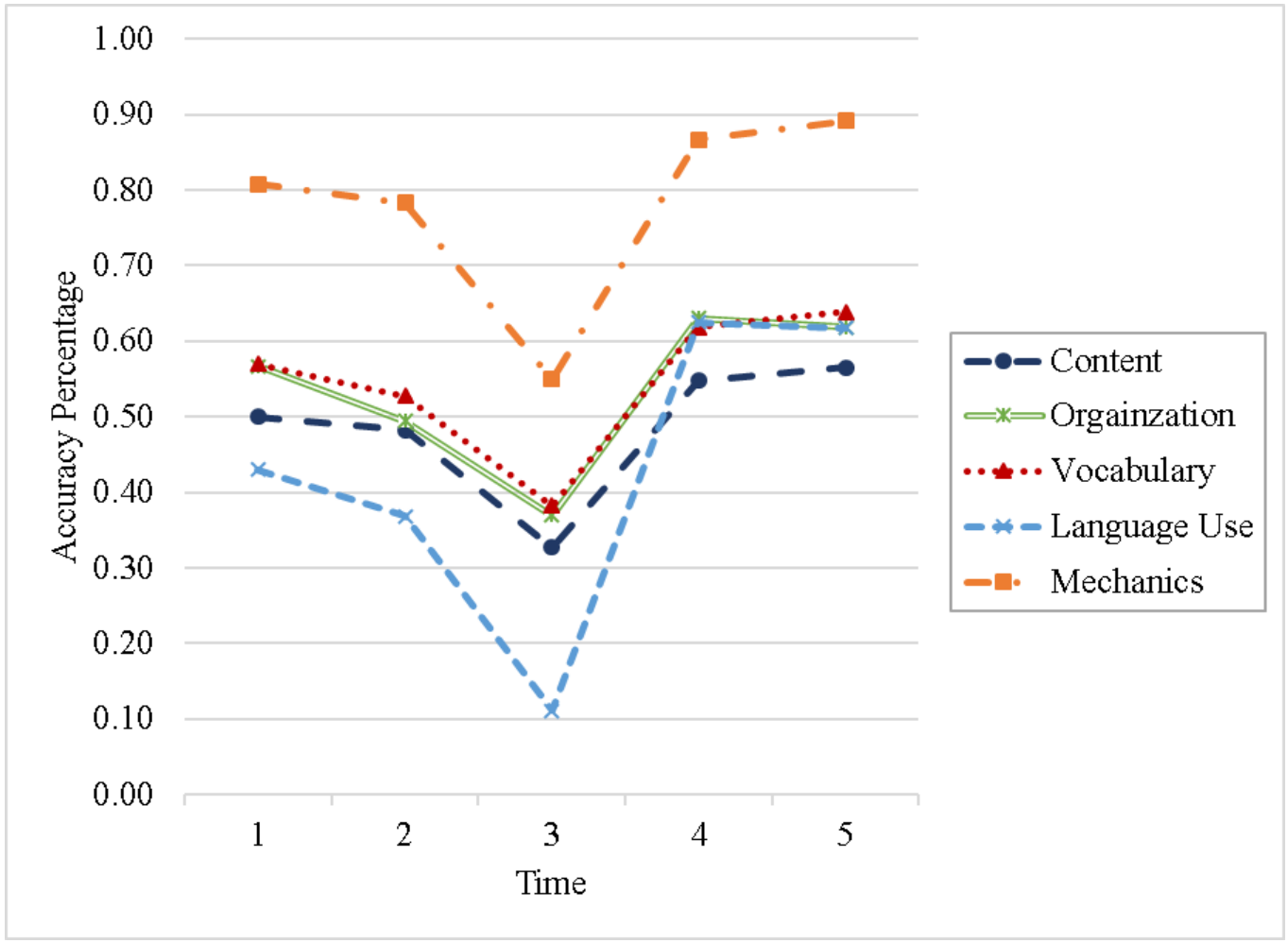


Figure 4. Criteria Recall Task means for Group A.

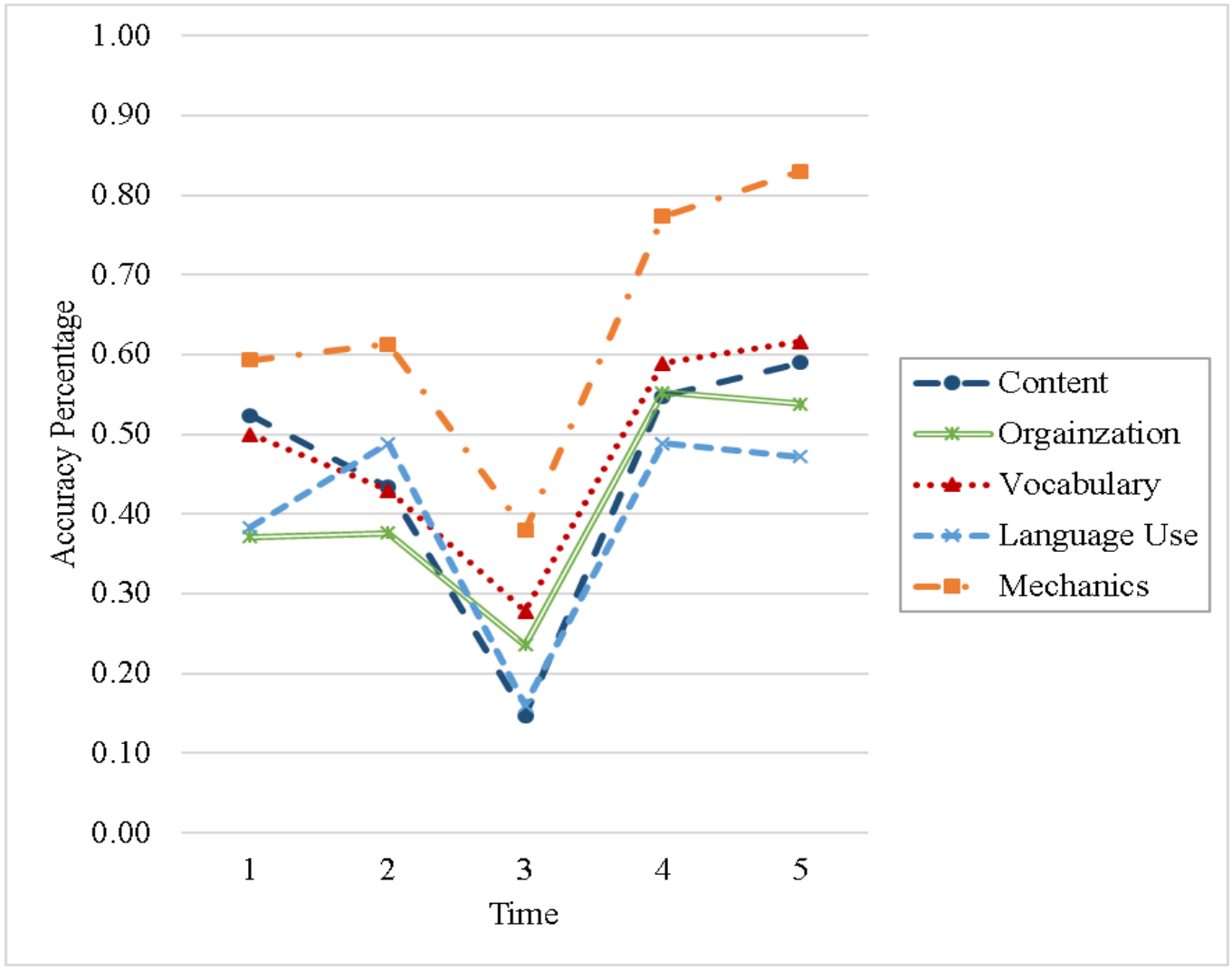


Figure 5. Criteria Recall Task means for Group B.

Table 9

Criteria Recall Task (CRT) Mean Scores

Round	Time	Group	Content	Organization	Vocabulary	Language Use	Mechanics
1	1 (post-training)	A	.50 (.17)	.57 (.19)	.57 (.16)	.43 (.28)	.81 (.14)
		B	.52 (.15)	.37 (.18)	.50 (.21)	.38 (.22)	.59 (.30)
	2 (pre-rating)	A	.48 (.26)	.49 (.14)	.53 (.19)	.37 (.23)	.78 (.20)
		B	.43 (.18)	.38 (.22)	.44 (.22)	.49 (.24)	.61 (.32)
	3 (pre-training)	A	.33 (.13)	.37 (.18)	.38 (.16)	.11 (.12)	.55 (.28)
		B	.15 (.14)	.24 (.14)	.28 (.24)	.17 (.20)	.38 (.25)
2	4 (post-training)	A	.55 (.12)	.63 (.10)	.62 (.15)	.63 (.13)	.87 (.15)
		B	.55 (.28)	.55 (.14)	.59 (.19)	.49 (.21)	.77 (.22)
	5 (pre-rating)	A	.57 (.19)	.62 (.13)	.64 (.21)	.62 (.18)	.89 (.11)
		B	.58 (.15)	.54 (.22)	.62 (.18)	.47 (.20)	.81 (.24)

Note. Standard deviations are in parentheses. Group A: n=13. Group B: n=14. The rubric changed between Time 4 and 5.

significant. For main effects, Time ($F1_{, 25} = 42.128$, $p > .001$, $\eta^2_P = .628$) and Category ($F_{3,76} = 16.906$, $p > .001$, $\eta^2_P = .403$) were statistically significant, but group was not ($F_{1, 25} = 3.958$, $p = .058$, $\eta^2_P = .137$).

Pairwise comparisons of Category at Time 2 (Round 1 Pre-rating) revealed that Mechanics was statistically and significantly different than every other category for Group A:

Content: mean difference= .301, $p = .017$, 97.5% CI [.012, .590]
Organization: mean difference= .290, $p = .008$, 97.5% CI [.034, .546]
Vocabulary: mean difference= .256, $p = .023$, 97.5% CI [.003, .509]
Language Use: mean difference= .415, $p = .001$, 97.5% CI [.109, .721]

At Time 3 (Round 2 Pre-training) for Group A, comparisons revealed that Language Use was statistically and significantly different than every other category:

Content: mean difference= -.216, $p = .009$, 97.5% CI [-.409, -.023]
Organization: mean difference= -.258, $p = .002$, 97.5% CI [-.458, -.058]
Vocabulary: mean difference= -.271, $p = .010$, 97.5% CI [-.514, -.028]
Mechanics: mean difference= -.439, $p > .001$, 97.5% CI [-.730, -.148]

Also, Content and Mechanics were statistically significantly different (mean difference= -.223, *p*

= .013, 97.5% CI [-.428, -.017]).

For Group B at Time 2, only Organization and Mechanics were statistically and significantly different (mean difference= -.237, p = .019, 97.5% CI [-.466, -.008]), and at Time 3, only Content and Mechanics were statistically and significantly different (mean difference= -.232, p = .003, 97.5% CI [-.416, -.048]).

For the second RM ANOVA model (including Time 4, Round 2 Pre-rating), the two-way interaction between Group and Category was not statistically significant ($F_{4, 90}$ = 1.218, p = .309, η^2_P = .046). The main effect of Category was statistically significant ($F_{3, 90}$ = 20597, p > .001, η^2_P = .452), but the main effect of Group was not ($F_{1, 25}$ = 2.066, p = .163, η^2_P = .076).

For Group A at Time 5 (Round 2 Pre-rating), pairwise comparisons between Categories revealed that Mechanics was significantly different from every other category:

Content: mean difference= .319, p > .001, 97.5% CI [.150, .489]
Organization: mean difference= .236, p = .003, 97.5% CI [.048, .424]
Vocabulary: mean difference= .249, p > .001, 97.5% CI [.081, .417]
Language Use: mean difference= .242, p = .005, 97.5% CI [.040, .443]

Similarly, for Group B at Time 5 (Round 2 Pre-rating), pairwise comparisons between Categories revealed that Mechanics was statistically significantly different from every other category:

Content: mean difference= .226, p > .001, 97.5% CI [.074, .377]
Organization: mean difference= .221, p = .002, 97.5% CI [.053, .389]
Vocabulary: mean difference= .180, p = .005, 97.5% CI [.030, .330]
Language Use: mean difference= .284, p > .001, 97.5% CI [.104, .465]

See Table 10 and Table 11 for all pairwise comparisons for RM ANOVA Model 1 and Model 2.

In summary, analyses of the CRT data showed that both groups found Mechanics to be the easiest to remember. At Time 2 (Round 1 Pre-rating), the groups remembered similar amounts of each category, though with slightly different within-group differences: Group A's

memory of the Mechanics category was statistically and significantly better than every other category, whereas Group B's memory of Mechanics was only statistically significantly better than their memory of Organization. After five weeks had passed, at Time 3 (Round 2 Pre-training), the two groups remembered similar amounts of each category except for Content, in which case Group A remembered a statistically significantly larger amount than Group B. There were also slight within-group differences: Group A's memory of Language Use was statistically significantly worse than every other category, and the only significant difference found in Group B's memory was between Mechanics and Content. At Time 5, after the groups had trained on the opposite rubric, the groups demonstrated very similar memory of the categories; both groups remembered significantly more of the Mechanics category than any other, and their memory of all other categories was similar (not significantly different).

Table 10

Category Pairwise Comparisons for Category Memory

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 2 (Round 1 Pre-rating)	Con.	Org.	-0.01	0.06	1.00	-0.21	0.19	0.06	0.05	1.00	-0.12	0.24
		Vocab.	-0.05	0.07	1.00	-0.29	0.20	0.01	0.07	1.00	-0.21	0.23
		Lang. Use	0.11	0.05	0.43	-0.07	0.29	-0.06	0.05	1.00	-0.22	0.11
		Mech.	-0.30*	0.09	0.02	-0.59	-0.01	-0.18	0.08	0.28	-0.44	0.08
	Org.	Con.	0.01	0.06	1.00	-0.19	0.21	-0.06	0.05	1.00	-0.24	0.12
		Vocab.	-0.03	0.06	1.00	-0.23	0.17	-0.05	0.05	1.00	-0.23	0.13
		Lang. Use	0.13	0.06	0.45	-0.07	0.32	-0.11	0.05	0.44	-0.29	0.07
		Mech.	-0.29*	0.08	0.01	-0.55	-0.03	-0.24*	0.07	0.02	-0.47	-0.01
	Vocab.	Con.	0.05	0.07	1.00	-0.20	0.29	-0.01	0.07	1.00	-0.23	0.21
		Org.	0.03	0.06	1.00	-0.17	0.23	0.05	0.05	1.00	-0.13	0.23
		Lang. Use	0.16	0.06	0.20	-0.06	0.38	-0.06	0.06	1.00	-0.25	0.13
		Mech.	-0.26*	0.08	0.02	-0.51	0.00	-0.19	0.07	0.11	-0.41	0.04
	Lang. Use	Con.	-0.11	0.05	0.43	-0.29	0.07	0.06	0.05	1.00	-0.11	0.22
		Org.	-0.13	0.06	0.45	-0.32	0.07	0.11	0.05	0.44	-0.07	0.29
		Vocab.	-0.16	0.06	0.20	-0.38	0.06	0.06	0.06	1.00	-0.13	0.25
		Mech.	-0.42*	0.09	0.00	-0.72	-0.11	-0.12	0.08	1.00	-0.40	0.15
	Mech.	Con.	0.30*	0.09	0.02	0.01	0.59	0.18	0.08	0.28	-0.08	0.44
		Org.	0.29*	0.08	0.01	0.03	0.55	0.24*	0.07	0.02	0.01	0.47
		Vocab.	0.26*	0.08	0.02	0.00	0.51	0.19	0.07	0.11	-0.04	0.41
		Lang. Use	0.42*	0.09	0.00	0.11	0.72	0.12	0.08	1.00	-0.15	0.40

Table 10 (cont'd)

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 3 (Round 2 Pre-training)	Con.	Org.	-0.04	0.06	1.00	-0.26	0.18	-0.09	0.06	1.00	-0.28	0.11
		Vocab.	-0.05	0.06	1.00	-0.26	0.15	-0.13	0.05	0.24	-0.31	0.05
		Lang. Use	0.22*	0.06	0.01	0.02	0.41	-0.01	0.05	1.00	-0.19	0.16
		Mech.	-0.22*	0.06	0.01	-0.43	-0.02	-0.23*	0.05	0.00	-0.42	-0.05
	Org.	Con.	0.04	0.06	1.00	-0.18	0.26	0.09	0.06	1.00	-0.11	0.28
		Vocab.	-0.01	0.07	1.00	-0.25	0.22	-0.04	0.06	1.00	-0.25	0.17
		Lang. Use	0.26*	0.06	0.00	0.06	0.46	0.07	0.05	1.00	-0.10	0.25
		Mech.	-0.18	0.08	0.41	-0.46	0.10	-0.14	0.07	0.67	-0.40	0.11
	Vocab.	Con.	0.05	0.06	1.00	-0.15	0.26	0.13	0.05	0.24	-0.05	0.31
		Org.	0.01	0.07	1.00	-0.22	0.25	0.04	0.06	1.00	-0.17	0.25
		Lang. Use	0.27*	0.07	0.01	0.03	0.51	0.12	0.06	0.83	-0.10	0.33
		Mech.	-0.17	0.09	0.69	-0.46	0.13	-0.10	0.08	1.00	-0.37	0.16
	Lang. Use	Con.	-0.22*	0.06	0.01	-0.41	-0.02	0.01	0.05	1.00	-0.16	0.19
		Org.	-0.26*	0.06	0.00	-0.46	-0.06	-0.07	0.05	1.00	-0.25	0.10
		Vocab.	-0.27*	0.07	0.01	-0.51	-0.03	-0.12	0.06	0.83	-0.33	0.10
		Mech.	-0.44*	0.09	0.00	-0.73	-0.15	-0.22	0.08	0.09	-0.48	0.04
	Mech.	Con.	0.22*	0.06	0.01	0.02	0.43	0.23*	0.05	0.00	0.05	0.42
		Org.	0.18	0.08	0.41	-0.10	0.46	0.14	0.07	0.67	-0.11	0.40
		Vocab.	0.17	0.09	0.69	-0.13	0.46	0.10	0.08	1.00	-0.16	0.37
		Lang. Use	0.44*	0.09	0.00	0.15	0.73	0.22	0.08	0.09	-0.04	0.48

Table 10 (cont'd)

			Group A					Group B				
						97.5% CI for Difference[b]					97.5% CI for Difference[b]	
Time	Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Time 4 (Round 2 Pre-rating)	Con.	Org.	-0.08	0.05	1.00	-0.26	0.09	0.00	0.05	1.00	-0.16	0.15
		Vocab.	-0.07	0.04	0.96	-0.21	0.07	-0.05	0.04	1.00	-0.17	0.08
		Lang. Use	-0.08	0.06	1.00	-0.26	0.11	0.06	0.05	1.00	-0.11	0.23
		Mech.	0.32*	0.05	0.00	-0.49	-0.15	-0.23*	0.05	0.00	-0.38	-0.07
	Org.	Con.	0.08	0.05	1.00	-0.09	0.26	0.00	0.05	1.00	-0.15	0.16
		Vocab.	0.01	0.05	1.00	-0.15	0.18	-0.04	0.04	1.00	-0.19	0.11
		Lang. Use	0.01	0.06	1.00	-0.19	0.20	0.06	0.05	1.00	-0.11	0.24
		Mech.	-0.24*	0.06	0.00	-0.42	-0.05	-0.22*	0.05	0.00	-0.39	-0.05
	Vocab.	Con.	0.07	0.04	0.96	-0.07	0.21	0.05	0.04	1.00	-0.08	0.17
		Org.	-0.01	0.05	1.00	-0.18	0.15	0.04	0.04	1.00	-0.11	0.19
		Lang. Use	-0.01	0.04	1.00	-0.15	0.14	0.10	0.04	0.13	-0.03	0.24
		Mech.	-0.25*	0.05	0.00	-0.42	-0.08	-0.18*	0.04	0.00	-0.33	-0.03
	Lang. Use	Con.	0.08	0.06	1.00	-0.11	0.26	-0.06	0.05	1.00	-0.23	0.11
		Org.	-0.01	0.06	1.00	-0.20	0.19	-0.06	0.05	1.00	-0.24	0.11
		Vocab.	0.01	0.04	1.00	-0.14	0.15	-0.10	0.04	0.13	-0.24	0.03
		Mech.	-0.24*	0.06	0.00	-0.44	-0.04	-0.28*	0.05	0.00	-0.46	-0.10
	Mech.	Con.	0.32*	0.05	0.00	0.15	0.49	0.23*	0.05	0.00	0.07	0.38
		Org.	0.24*	0.06	0.00	0.05	0.42	0.22*	0.05	0.00	0.05	0.39
		Vocab.	0.25*	0.05	0.00	0.08	0.42	0.18*	0.04	0.00	0.03	0.33
		Lang. Use	.242*	0.06	0.00	0.04	0.44	.284*	0.05	0.00	0.10	0.46

Note. Based on estimated marginal means. *The mean difference is significant at the .025 level. b= Adjustment for multiple comparisons: Bonferroni.

Table 11

Group Pairwise Comparisons for Category Memory

Time	Category	Mean Difference between Groups	Std. Error	Sig.[b]	97.5% CI for Difference[b] Lower Bound	97.5% CI for Difference[b] Upper Bound
Time 2 (Round 1 Pre-rating)	Content	0.05	0.08	0.56	-0.15	0.25
	Organization	0.12	0.08	0.15	-0.07	0.30
	Vocabulary	0.10	0.07	0.19	-0.08	0.28
	Language Use	-0.12	0.09	0.20	-0.34	0.10
	Mechanics	0.17	0.11	0.13	-0.09	0.43
Time 3 (Round 2 Pre-training)	Content	0.18*	0.05	0.00	0.06	0.30
	Organization	0.13	0.06	0.04	-0.02	0.28
	Vocabulary	0.10	0.08	0.20	-0.09	0.29
	Language Use	-0.05	0.06	0.41	-0.19	0.09
	Mechanics	0.17	0.10	0.10	-0.07	0.41
Time 4 (Round 2 Pre-rating)	Content	> 0.01	0.06	1.00	-0.14	0.14
	Organization	0.08	0.05	0.12	-0.04	0.19
	Vocabulary	0.02	0.07	0.71	-0.13	0.18
	Language Use	0.14	0.07	0.06	-0.03	0.30
	Mechanics	0.09	0.07	0.21	-0.08	0.27

CHAPTER 4

RESULTS: RUBRIC USAGE

To answer the second research question (To what extent do raters show evidence of ordering effects in their rubric usage?), I examined the eye-tracking data. While the participants scored essays, the eye tracker recorded data about the participants' fixations on the rubric, and specifically on each rubric category. Before analyzing the data, I set five areas of interest (AOI), which allows the eye-tracker to provide aggregated data based on fixations that fall within the AOI. I designated each rubric category as a separate AOI and could then investigate the participants' attention to the individual rubric categories (see Figure 6 for an example screen shot of the overlaid AOIs). Following Winke and Lim (2015), I examined three different eye-movement metrics: *time to first fixation* (TFF), *total fixation duration* (TFD), and *visit count* (VC). TFF measures how long it took a participant to fixate on a rubric category. In other words, it is the time from when the rubric appeared on screen and until the rater first looked at the text within the category. TFF can provide information on the order in which participants attended to the rubric categories. TFD measures the total sum time a participant fixated on text within a rubric category. VC measures how many separate visits a participant made to a rubric category. In other words, each time a person's eye gaze passes in and then out of the AOI, regardless of the amount of time spent gazing within the AOI, counts as one visit. These measures provide different information about the participants' attention to the rubric categories. Taken together, these three measures provide a snapshot of the raters' usage of the rubric and the individual categories during rating. From the raters' attention to the rubric categories, I can investigate primacy effects on the rater's real-time rubric-category usage.

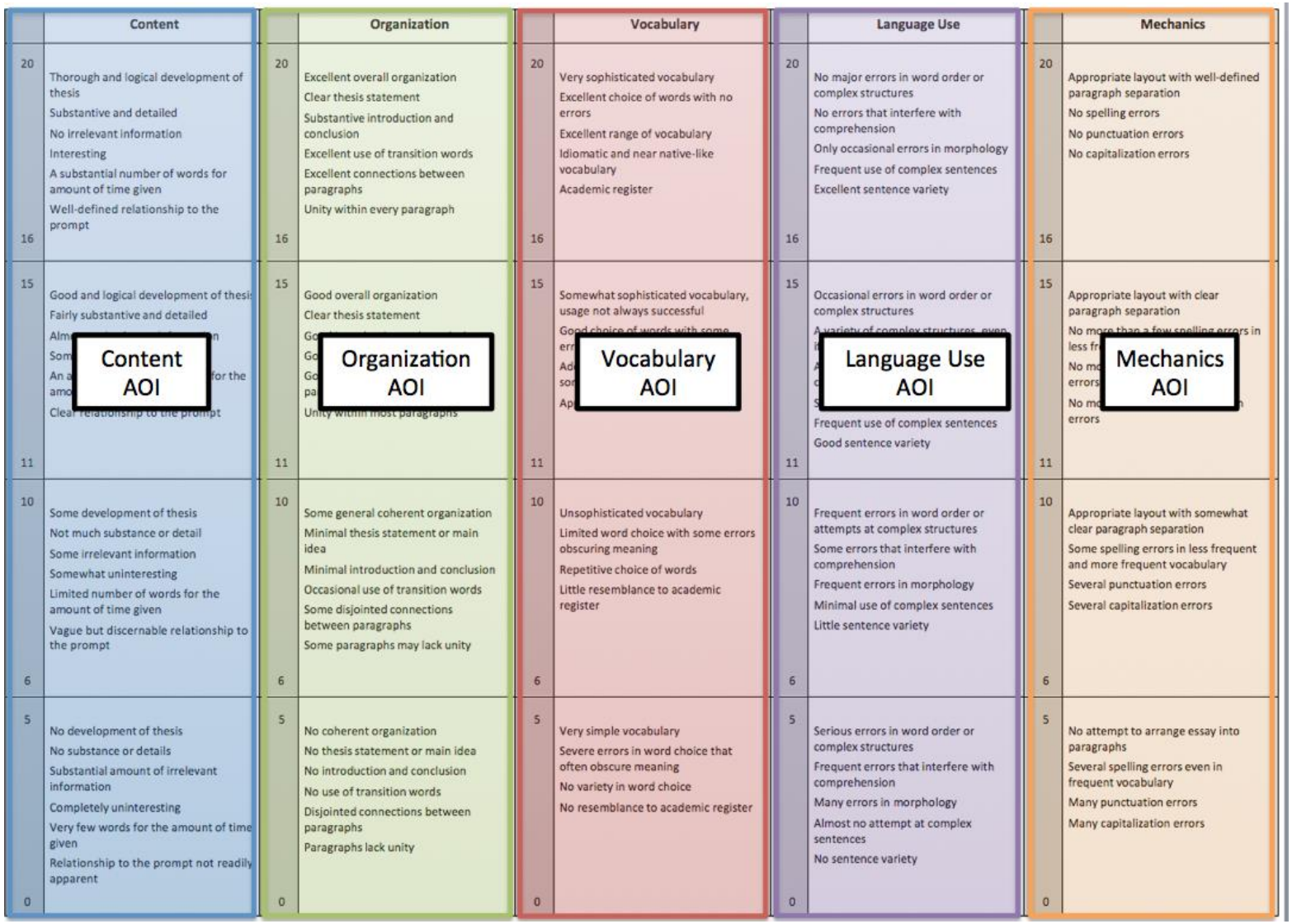


Figure 6. Example screenshot of the Standard Rubric with Areas of Interest (AOIs) superimposed.

Summary of the Rubric Usage Findings

Raters' Order of Attention

By using the time to first fixation (TFF) data, I examined the order in which raters fixated on categories and how this relates to the rubric the raters (1) first trained on and (2) used during the rating session. Table 12 displays mean TFF times, which are measured in seconds. Lower values indicate that raters fixated on a category earlier, and higher values indicate that it took raters longer to finally fixate on a rubric category. As shown in the descriptives, all raters were likely to look at the categories in the order that the categories appeared on the rubric used during the rating session. In other words, when using the SR rubric for rating, all raters tended to look at the categories in the following order: Content, Organization, Vocabulary, Language Use, and Mechanics. When using the RR for rating, raters also tended to look at the categories from left to right as they appeared: Mechanics, Language Use, Vocabulary, Organization, and Content. In other words, the raters' order of fixation aligned with the order in which the categories were presented from left to right. There was no evidence that raters' order of fixation was influenced differently by the rubric that a group trained on first (i.e., all raters demonstrated the same fixation behavior). The TFF values are plotted in Figure 7.

Table 12

Mean Time to First Fixation (TFF)

Rubric	Group	Content	Organization	Vocabulary	Language Use	Mechanics
SR	A	109.87 (27.97)	126.26 (34.76)	150.79 (41.97)	177.31 (50.70)	202.49 (60.61)
	B	108.46 (45.25)	117.19 (47.88)	138.09 (55.92)	154.88 (64.53)	173.50 (75.95)
RR	A	167.29 (66.09)	154.61 (57.51)	142.10 (44.10)	127.76 (40.31)	124.28 (40.23)
	B	206.50 (71.77)	182.06 (58.90)	158.85 (45.19)	135.69 (39.01)	125.08 (38.74)

Note. Standard deviations are in parentheses. Mean TFF and SD values are measured in seconds. Group A: n = 13. Group B: n = 14.

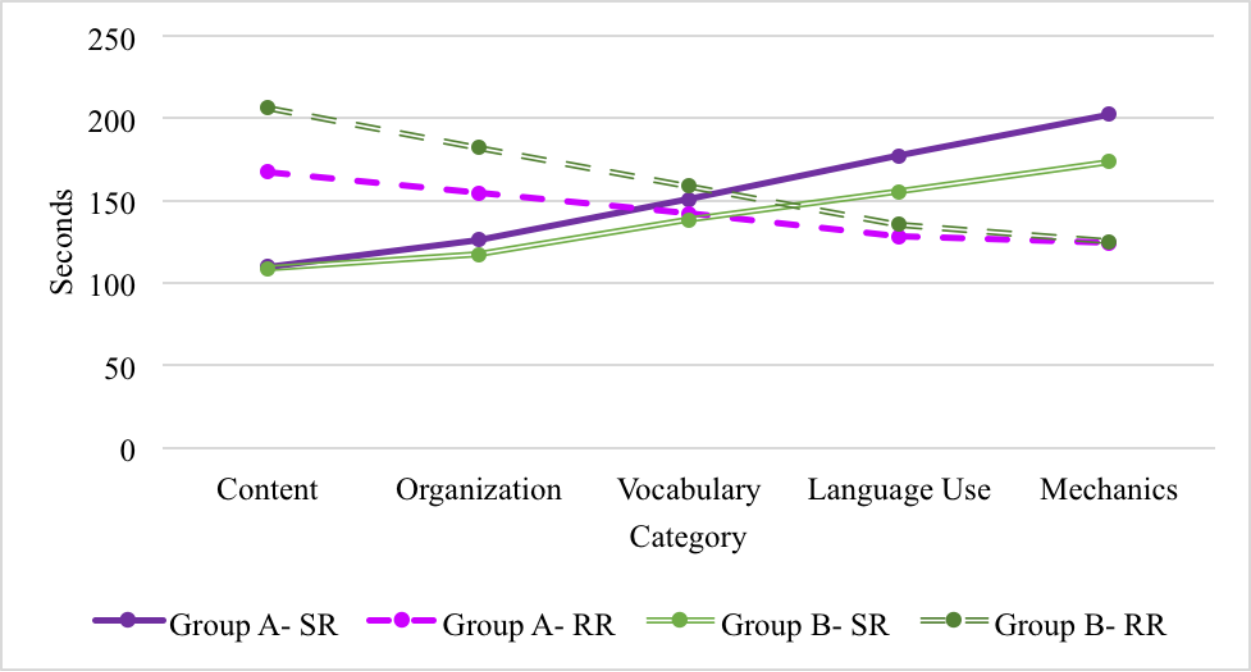


Figure 7. Mean time to first fixation (TFF). SR = Standard Rubric, RR = Reordered Rubric

Rater's Concentrated Attention to Rubric Categories

The descriptive statistics for controlled TFD are in Table 13, in Figure 2, and in Figure 3. The controlled TFD is the mean difference between SR and RR values controlled for the number of words (i.e., mean difference divided by number of words in the category). Positive values indicate that raters' TFD on a given category was longer for the SR than for the RR, and negative values indicate that raters fixated longer on a category when using the RR than when using the SR. A value of zero would indicate no difference in TFD between the two rubrics.

The controlled mean TDF values, which are plotted in Figure 2, show two general trends. The first is that SR values for both groups follow the linear trend that aligns with the order of the category presentation on the SR, with higher fixation values toward the left of the rubric and lower toward the right (with the exception of Content). The second is that for RR values for both groups, the four left-most categories have similar fixation values, and Content (the right-most category) is somewhat lower than the other four.

The controlled mean difference TFD values for both groups decrease unidirectionally from Content to Mechanics. For example, for Group A, the controlled TFD values decrease from Content (mean = 0.012, the largest change in TFD from SR to RR where raters fixated on the category more in the SR, resulting in a positive value) to Mechanics (-0.031, largest change in TFD from SR to RR where raters fixated on a category more in the RR, resulting in a negative value). This trend is shown in Figure 3.

There were also specific trends within each group. For Group A, the controlled TFD pattern shows that the raters: (1) spent more time fixating on Content and Organization when they appeared on the left of the rubric (i.e., on the SR); (2) spent about the same amount of time fixating on Vocabulary, which was in the middle on both rubrics; and (3) spent more time

fixating on Mechanics and Language Use when they appeared on the left of the rubric (i.e., on the RR). In Group B, however, (1) the raters fixated equal amounts of time on the Content category between the two rubrics, and (2) the raters fixated longer on the remaining four categories during Round 1. For these four categories, the farther left the category appeared on the RR, the bigger the difference in fixation time when the raters used the SR rubric.

Next, I computed an RM ANOVA to investigate whether and how the groups differed in their category fixations. The main effect of Category (Greenhouse-Geisser: $F_{3,60} = 9.227$, $p < .001$, $\eta^2_P = .278$) was statistically significant. However, neither the main effect of Group (Greenhouse-Geisser: $F_{1, 24} = 1.845$, $p = .187$, $\eta^2_P = .071$) nor the two-way interaction between Group and Category (Greenhouse-Geisser: $F_{3, 60} = 0.161$, $p = .894$, $\eta^2_P = .007$) were statistically significant. Post-hoc comparisons of the categories showed that, within groups (see Table 14), Group A's TFD values were statistically different between Mechanics and three other categories: Content, Organization, and Vocabulary. This shows that Group A's change in category attention (TFD) to Mechanics was statistically different from (i.e., larger than) the change in attention Group A showed for Content, Organization, and Vocabulary. For Group B, raters' TFD values were statistically different between Mechanics and all other categories, and between Content and Vocabulary. Similar to Group A, Group B's change in attention to the Mechanics category was the largest and was statically different from Group B's changes in attention for all other categories. Group B also showed similar statistical differences between Content (in which there was almost no change in attention between the two rubrics) and Vocabulary (in which there was a moderate change in attention between the two rubrics).

Finally, I computed a one-sample *t* test with the controlled TFD values for each category for each group to determine whether the fixation times between the two rounds for a Group on a

given category were statistically different (see Table 15). A controlled TFD value that is statistically different from zero indicates that the group's fixation behavior for the category was different between the two rounds. The results showed that the only category that was statically significant was Mechanics for Group B (p = .005), meaning that for all other categories, the groups fixated equally (within groups) on the categories between the two rubrics.

Table 13

Mean Total Fixation Duration (TFD)

Group	Category	SR	RR	Mean Difference TFD (sec.)	Number of Words	TFD (controlled) Mean	TFD (controlled) SD
Group A (n = 11)	Content	8.226	6.742	1.484	122	0.012	0.023
	Organization	8.653	8.046	0.608	100	0.006	0.024
	Vocabulary	6.958	6.865	0.093	90	0.001	0.032
	Language Use	7.662	8.971	-0.904	120	-0.008	0.035
	Mechanics	4.815	6.696	-1.770	89	-0.021	0.046
Group B (n = 15)	Content	6.451	6.412	0.039	122	0.000	0.026
	Organization	5.919	7.068	-1.149	100	-0.011	0.034
	Vocabulary	4.843	6.397	-1.553	90	-0.017	0.028
	Language Use	5.784	7.930	-2.146	120	-0.018	0.036
	Mechanics	3.230	6.512	-3.282	89	-0.037	0.044

Note. Mean difference is SR TFD value minus RR TFD value by category. Controlled TFD is mean difference TFD divided by number of words per category.

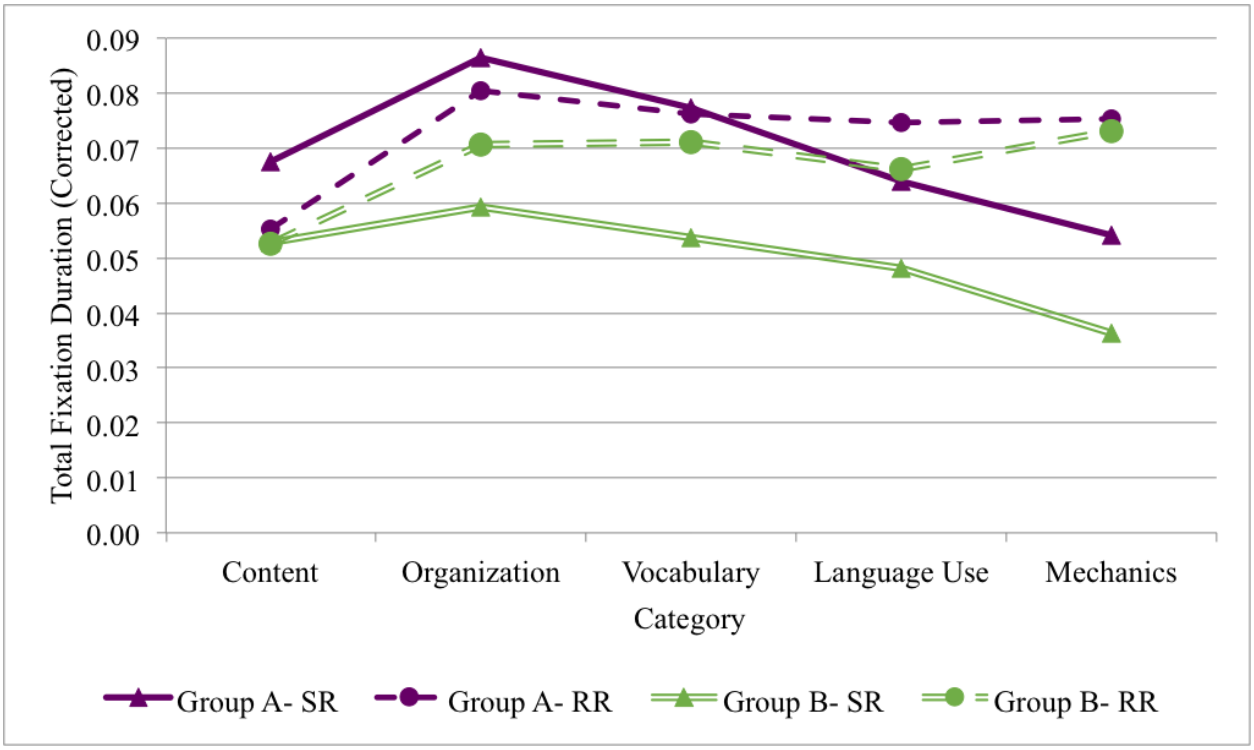


Figure 8. Mean Corrected Total Fixation Duration (TFD).

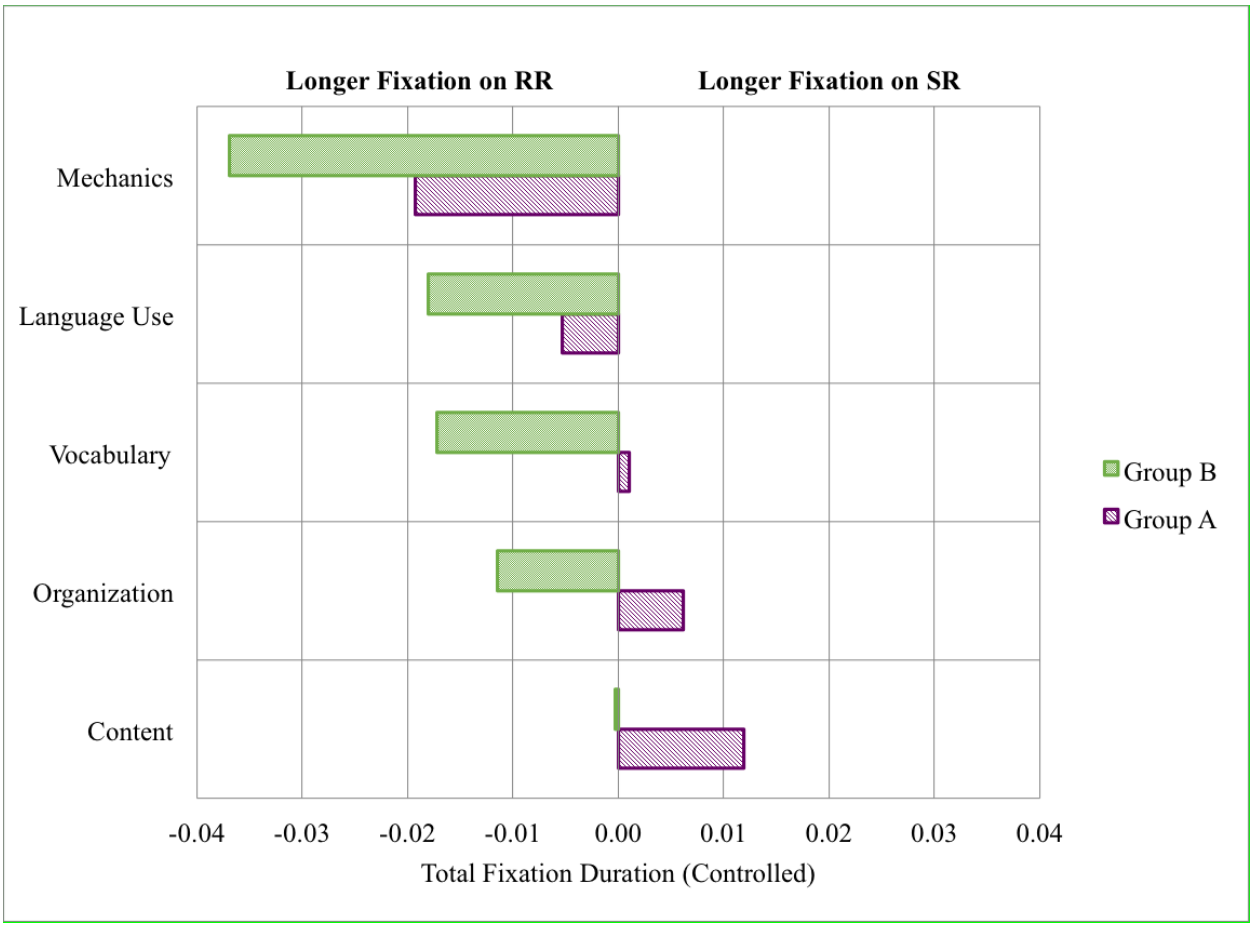


Figure 9. Mean Difference in Controlled Total Fixation Duration between the Standard Rubric and Reordered Rubric.

Table 14

Category Pairwise Comparison for Total Fixation Duration (TFD)

		Group A					Group B				
					95% CI for Difference[b]					95% CI for Difference[b]	
Category	Comparison Category	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound	Mean Difference	Std. Error	Sig.[b]	Lower Bound	Upper Bound
Con.	Org.	0.006	0.008	.471	-0.011	0.023	0.012	0.007	.110	-0.003	0.027
	Vocab.	0.011	0.009	.249	-0.008	0.031	0.018*	0.008	.039	0.001	0.034
	Lang. Use	0.020	0.011	.079	-0.002	0.042	0.018	0.009	.059	-0.001	0.037
	Mech.	0.033*	0.014	.022	0.005	0.061	0.037*	0.012	.004	0.013	0.061
Org.	Con.	-0.006	0.008	.471	-0.023	0.011	-0.012	0.007	.110	-0.027	0.003
	Vocab.	0.005	0.007	.481	-0.009	0.020	0.006	0.006	.348	-0.007	0.018
	Lang. Use	0.014	0.008	.103	-0.003	0.030	0.006	0.007	.362	-0.008	0.021
	Mech.	0.027*	0.010	.010	0.007	0.047	0.025*	0.008	.006	0.008	0.043
Vocab.	Con.	-0.011	0.009	.249	-0.031	0.008	-0.018*	0.008	.039	-0.034	-0.001
	Org.	-0.005	0.007	.481	-0.020	0.009	-0.006	0.006	.348	-0.018	0.007
	Lang. Use	0.009	0.009	.370	-0.011	0.028	0.001	0.008	.939	-0.016	0.017
	Mech.	0.022*	0.009	.016	0.005	0.040	0.020*	0.007	.013	0.005	0.035
Lang. Use	Con.	-0.020	0.011	.079	-0.042	0.002	-0.018	0.009	.059	-0.037	0.001
	Org.	-0.014	0.008	.103	-0.030	0.003	-0.006	0.007	.362	-0.021	0.008
	Vocab.	-0.009	0.009	.370	-0.028	0.011	-0.001	0.008	.939	-0.017	0.016
	Mech.	0.014	0.007	.081	-0.002	0.029	0.019*	0.006	.007	0.006	0.032
Mech.	Con.	-0.033*	0.014	.022	-0.061	-0.005	-0.037*	0.012	.004	-0.061	-0.013
	Org.	-0.027*	0.010	.010	-0.047	-0.007	-0.025*	0.008	.006	-0.043	-0.008
	Vocab.	-0.022*	0.009	.016	-0.040	-0.005	-0.020*	0.007	.013	-0.035	-0.005
	Lang. Use	-0.014	0.007	.081	-0.029	0.002	-.019*	0.006	0.007	-0.032	-0.006

Table 15

Total Fixation Duration (TFD) Mean Difference t Test

			Bootstrap[a]				
						95% CI	
Group	Category	Mean Difference	Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
Group A	Content	0.012	0.000	0.007	.111	-0.001	0.025
	Organization	0.006	0.000	0.007	.422	-0.007	0.019
	Vocabulary	0.001	0.000	0.009	.917	-0.017	0.019
	Language Use	-0.008	0.001	0.010	.517	-0.029	0.008
	Mechanics	-0.021	0.000	0.013	.147	-0.048	0.002
Group B	Content	0.000	0.000	0.007	.971	-0.013	0.012
	Organization	-0.011	0.000	0.008	.196	-0.027	0.004
	Vocabulary	-0.017	0.000	0.007	.043	-0.033	-0.004
	Language Use	-0.018	0.000	0.009	.072	-0.035	-0.001
	Mechanics	-0.037	0.000	0.011	.005*	-0.056	-0.016

Note. A. Unless otherwise noted, bootstrap results are based on 1,000 bootstrap samples. To adjust for multiple comparisons, mean differences were set to be significant at the .005 level.

Raters' Frequency of Attention to Rubric Categories

The descriptive statistics for visit count are in Table 16 and Figure 4. For the subsequent inferential statistics, I used the mean difference visit count, which is the difference between SR and RR values for a given category. Positive values indicate that raters had more visits on the SR than on the RR, and negative values indicate that raters had more visits for a given category on the RR than on the SR; a value of zero would indicate no difference in visits between the two rubrics.

Similar to the TFD results, the mean difference VC values (see Figure 5) for each group decrease unidirectionally from Content to Mechanics. For Group A, the largest mean differences (i.e., the absolute change between rubrics) was for the outer-most categories, Content and Mechanics. For Group B, however, the largest mean difference (absolute change) was for

Mechanics, and the smallest change was for Content, showing the most equal amount of visits for a category between the two rubrics.

Another trend in the mean VC data is that both groups on both rubrics had the highest visit counts on Organization, Vocabulary, and Language Use; Content and Mechanics had the fewest visits, which alternated by rubric. When Content was in the right-most positions (on the RR), raters made the fewest visits to the category. The same was true for Mechanics when it appeared in the right-most position (on the SR).

Looking at more specific, group-related trends, Group A paid more visits to Content, Organization, Vocabulary, in Round 1, but visited Language Use and Mechanics more frequently in Round 2. Group B, however, had more visits to every category in Round 1.

Next, I computed an RM ANOVA to investigate whether and how the groups differed in their category-visit (VC) behavior. Neither main effect of Group (Greenhouse-Geisser: $F_{1, 24}$ = 2.392, p = .135, η^2_P = .091), nor the main effect of Category (Greenhouse-Geisser: $F_{2, 51}$ = 1.942, p = .152, η^2_P = .075), nor the two-way interaction between Group and Category (Greenhouse-Geisser: $F_{2, 51}$ = 0.124, p = .894, η^2_P = .005) were statistically significantly different. Thus, I cannot reject the null hypothesis that the groups had similar category visit (VC) behavior.

Nonetheless, to test this hypothesis in another way, I computed a one-sample *t* test with the mean difference VC values for each category and for each group; I did this to determine whether the number of category visits between the two rounds for a Group on a given category were statistically different (see Table 17). A mean difference VC value that is statistically different from zero indicates that the group's fixation behavior (i.e., number of visits) for the category was different between the two rubrics. The results showed that only VC difference that was statistically significant was Content for Group A (p = .005). For every other category, the

data did not show statistical differences in visit frequency behavior.

Finally, I examined the number of instances that raters assigned a category score without reading the category on the rubric. The descriptive statistics for instances of category skipping are in Table 18 and Figure 6. Positive mean-difference values indicate that there were more instances of category skipping on the SR than on the RR for a given category; negative values indicate that there were more instances of category skipping on the RR than the SR. The data show three general trends. First, for Group A in Round 1, the number of category skips aligns with the rubric layout; that is, the farther to the right a category appeared, the more the raters skipped the category. Second, for Group B in Round 1 (RR), the instances of category skipping were relatively similar, meaning that a rater skipped each category an average of two times during the rating session. Finally, for both groups, raters were more likely to skip Content and Mechanics when they appeared in the right-most position, as demonstrated by lower, negative values for Content, and higher, positive values for Mechanics.

Table 16

Mean Visit Count (VC)

Group	Category	Mean SR	Mean RR	Mean Difference	Mean Difference SD
Group A (n = 11)	Content	3.24	2.52	0.61	0.52
	Organization	3.96	3.55	0.26	1.12
	Vocabulary	4.11	3.73	0.24	1.37
	Language Use	4.24	4.13	-0.03	2.05
	Mechanics	3.02	3.31	-0.30	1.97
Group B (n = 15)	Content	2.28	2.51	-0.23	0.94
	Organization	2.90	3.41	-0.52	1.31
	Vocabulary	2.82	3.37	-0.56	1.27
	Language Use	2.83	3.44	-0.61	1.84
	Mechanics	1.79	2.62	-0.84	1.45

Note. Mean difference is Standard Rubric (SR) category value minus Reordered Rubric (RR) category value.

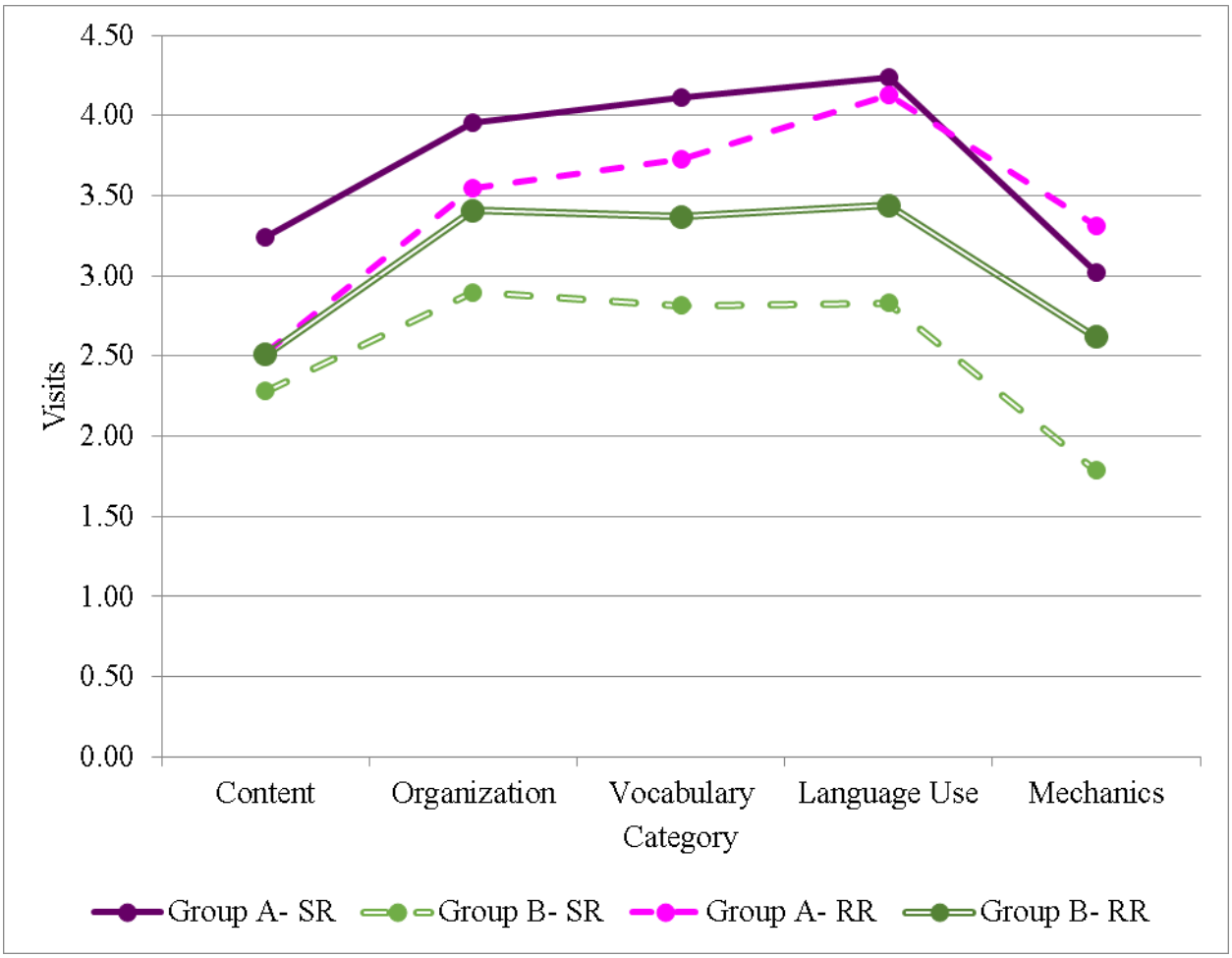


Figure 10. Mean visit count (VC).

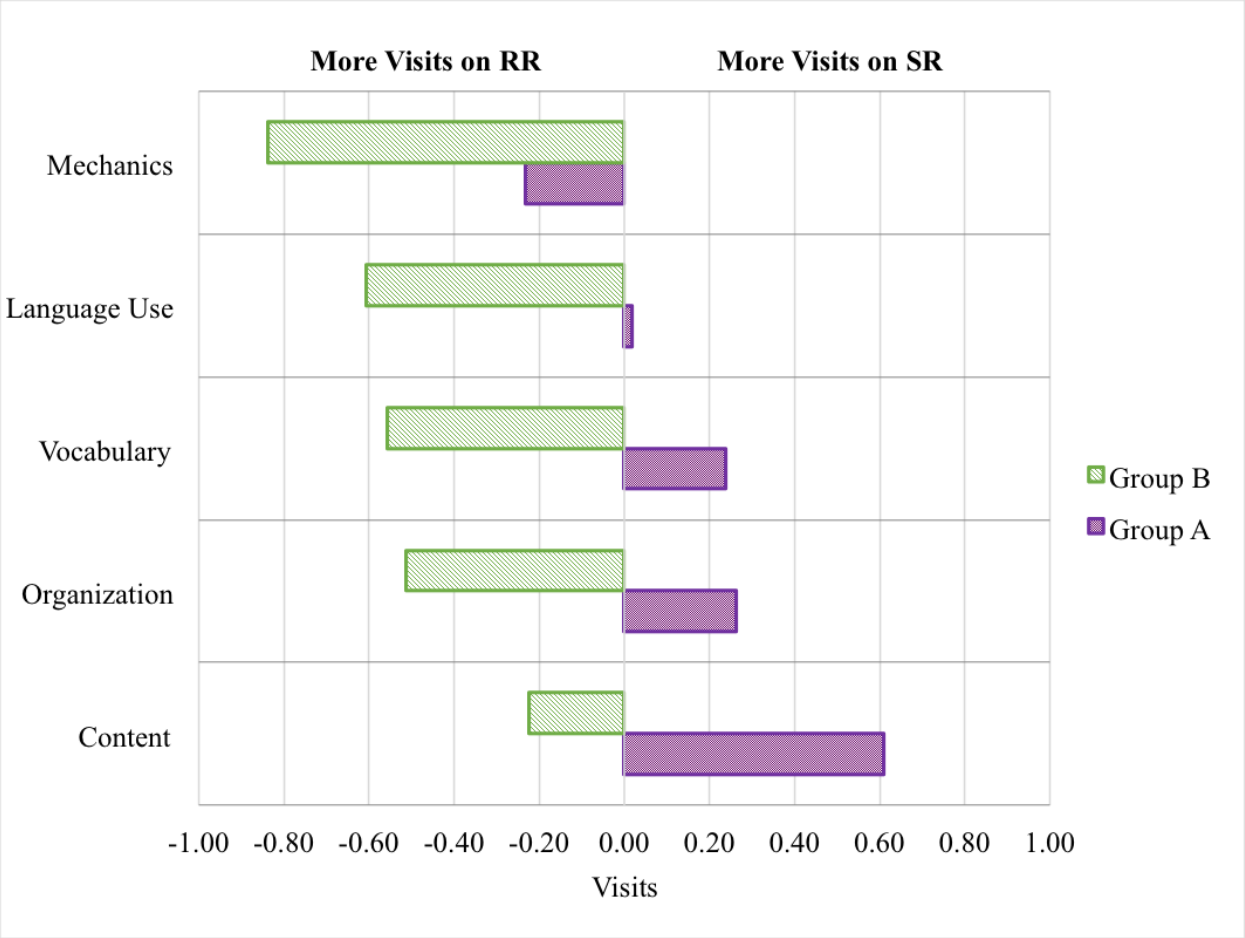


Figure 11. Mean difference Visit Count (VC).

Table 17

Visit Count (VC) Mean Difference t Test

			Bootstrap[a]				
						95%CI	
Group	Category	Mean Difference	Bias	Std. Error	Sig. (2-tailed)	Lower	Upper
Group A (n = 11)	Content	0.609	0.000	0.146	0.005*	0.341	0.909
	Organization	0.262	0.021	0.324	0.448	-0.363	0.918
	Vocabulary	0.238	-0.008	0.398	0.578	-0.530	1.036
	Language Use	-0.029	0.020	0.593	0.970	-1.287	1.041
	Mechanics	-0.295	-0.011	0.562	0.611	-1.486	0.695
Group B (n = 15)	Content	-0.230	-0.002	0.229	0.354	-0.680	0.233
	Organization	-0.517	0.001	0.344	0.195	-1.190	0.137
	Vocabulary	-0.557	-0.007	0.303	0.102	-1.203	0.003
	Language Use	-0.607	0.008	0.458	0.219	-1.503	0.343
	Mechanics	-0.837	0.000	0.357	0.037	-1.506	-0.117

Note. a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Table 18

Mean Frequency of Category Skipping

Group	Category	SR	RR	Mean Difference	SD (mean difference)
Group A (n = 11)	Content	0.45	1.27	-0.91	1.64
	Organization	0.82	0.27	0.55	0.69
	Vocabulary	0.91	0.18	0.73	1.68
	Language Use	0.91	0.27	0.64	0.81
	Mechanics	2.09	0.55	1.41	2.14
Group B (n = 15)	Content	1.47	2.07	-0.60	2.95
	Organization	1.40	1.80	-0.40	2.66
	Vocabulary	2.00	2.07	-0.07	3.71
	Language Use	1.47	1.80	-0.33	3.41
	Mechanics	3.60	2.13	1.47	5.32

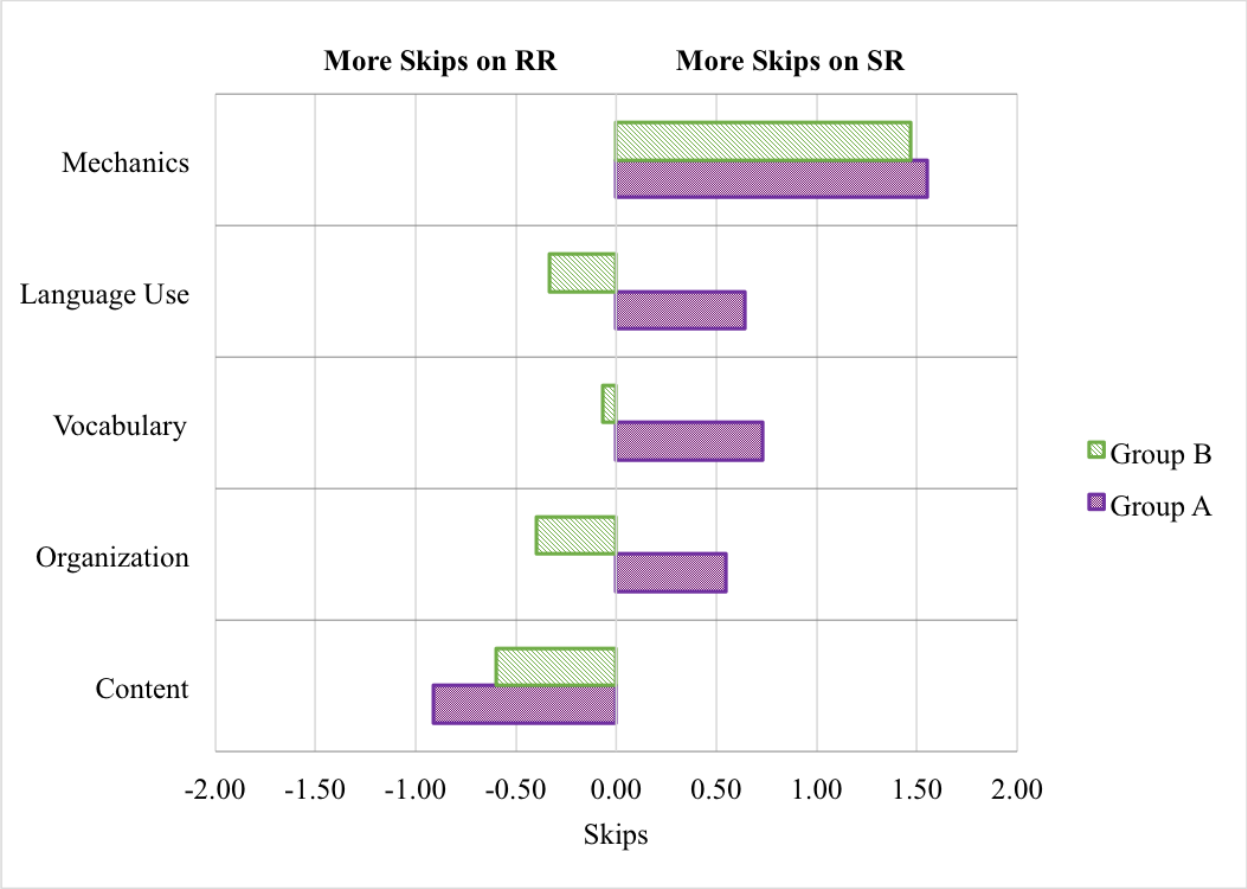


Figure 12. Category skipping (mean difference).

Eye Tracking Results Summary

To summarize the investigations into the primacy effect and raters' attention to the rubric categories, one strong trend emerged across the metrics I used to investigate these areas: raters demonstrated a strong left-to-right bias in their fixations. The order in which raters fixated on the categories, the amount of time they spent fixating on a category, and the number of times they fixated on a category all aligned with the left-to-right appearance of the categories. This was true for the SR and the RR, and for both groups. Additionally, the outer-most categories (i.e., left-most and right-most) seem most susceptible to ordering effects. For example, there

were significant differences in total fixation durations (TFD) for the Mechanics category between and within groups. Furthermore, Group A's fixations (measured via TFD) on Content was statistically different from Vocabulary and Language Use. That is, Group A fixated on Content more (for longer amounts of time) than they did on Vocabulary and Language Use. Another example is the number of times a category was skipped when it appeared in the right-most category: whether Content or Mechanics, the category was skipped more when it appeared in the right-most position.

One other trend that emerged was that Group B's fixations behavior provided some evidence for category equality when using the RR, meaning that they used categories somewhat more equally than Group A. One example of this is that Group B had equal fixation durations (TFD) on Content between the two rounds. Typically, when Content was in the right-most position, raters paid less attention to it than when it appeared in the left-most position, but this was not the case for Group B; they spent equal amounts of time on the category between the two rubrics. Another example of evidence for category equality for Group B is that they demonstrated relatively equal amounts of category skips in their rating using the RR.

CHAPTER 5

RESULTS: SCORE IMPACT

To answer the third research question (To what extent are raters' scores impacted by ordering effects?), I employ intra-rater reliability measures to examine rater reliability for each rubric category and Rasch analysis to examine relative rater severity in relation to rubric category for each rubric.

Summary of the Score Impact Findings

Descriptive statistics on raters' raw scores are in Table 19, and Table 20 summarizes the descriptive statistics for the collapsed data set in which I collapsed the raters' scores into five score points for the Rasch analysis. These descriptives show that raters assigned higher scores in Round 2, which was the Reordered Rubric (RR) for Group A and Standard Rubric (SR) for Group B. Across all raters and rubrics, raters generally assigned the highest scores in the Vocabulary category. Finally, the largest differences in scoring on the individual categories is *within* groups (e.g., Group A SR vs. Group A RR) rather than *between* groups (e.g., Group A SR vs. Group B SR), and this is true in both the uncollapsed and collapsed raw scores.

Rater Scoring Consistency

As presented in Table 21, raters were highly consistent in their scoring. For every category, no matter the rubric, the analysis yielded a coefficient of .884 (RR - Round 2 - Language Use) or higher, and the highest coefficient was .950 (RR – Round 1 – Total). A general pattern that emerged in Round 1 is that raters agreed most on the category presented first on the rubric, Content for the SR rubric (.946) and Mechanics for the RR rubric (.946). For Round 2, once both groups had been exposed to both rubrics, the general pattern that emerged is

that raters agreed the most on Content, Organization, and Mechanics.

Table 19

Descriptive Statistics for Raters' Essay Scores

		Group A					
		Cont.	Org.	Vocab.	Lang.	Mech.	Total
SR	Mean	11.13	10.79	11.67	11.44	11.55	57.39
(*n* = 13)	SD	3.09	2.82	2.78	2.56	2.92	12.48
	Min	4	5	4	4	5	24
	Max	20	20	20	20	19	94
RR	Mean	12.16	11.87	12.31	12.15	12.44	61.66
(*n* = 15)	SD	2.62	2.41	2.27	2.21	2.67	10.24
	Min	1	1	3	3	4	15
	Max	19	19	18	19	20	91
		Group B					
		Cont.	Org.	Vocab.	Lang.	Mech.	Total
SR	Mean	11.72	11.78	12.36	12.16	11.84	60.85
(*n* = 15)	SD	2.58	2.59	2.36	2.30	2.52	10.78
	Min	2	2	5	5	4	23
	Max	19	18	19	19	19	91
RR	Mean	11.28	11.39	12.19	11.84	11.72	59.02
(*n* = 15)	SD	3.19	3.28	2.90	2.72	3.18	12.81
	Min	1	3	3	4	3	18
	Max	20	20	20	20	19	94

Table 20

Descriptive Statistics for Raters' Essay Scores (Collapsed)

		Group A				
		Cont.	Org.	Vocab.	Lang.	Mech.
SR	Mean	3.01	2.85	3.18	3.05	3.16
(*n* = 13)	SD	1.05	0.98	0.95	0.92	1.02
	Min	1	1	1	1	1
	Max	5	5	5	5	5
RR	Mean	3.36	3.21	3.41	3.37	3.41
(*n* = 15)	SD	0.91	0.88	0.81	0.83	0.87
	Min	2	2	1	1	1
	Max	5	5	5	5	5
		Group B				
		Cont.	Org.	Vocab.	Lang.	Mech.
SR	Mean	3.19	3.17	3.39	3.33	3.24
(*n* = 15)	SD	0.90	0.91	0.85	0.89	0.89
	Min	1	1	1	1	1
	Max	5	5	5	5	5
RR	Mean	3.04	3.03	3.31	3.18	3.19
(*n* = 15)	SD	1.05	1.10	1.02	0.96	1.05
	Min	1	1	1	1	1
	Max	5	5	5	5	5

Table 21

Intraclass Correlations

		ICC Coefficient	
Rubric	Category	Round 1	Round 2
SR	Content	.946	.927
	Organization	.931	.927
	Vocabulary	.927	.914
	Language Use	.900	.913
	Mechanics	.919	.925
	Total	.941	.942
RR	Content	.938	.919
	Organization	.938	.922
	Vocabulary	.922	.898
	Language Use	.896	.884
	Mechanics	.946	.919
	Total	.950	.934

Rater Severity

In Table 22 are the results of the interaction between Group and Category for Model 1. The Partial Credit Model (PCM) allows inspection of raters' relative severity on each rubric category through the Group by Category interaction. In the target measure column are the logit (i.e., severity) values for each category. A value higher than zero indicates rater severity (across raters) on a given category, and a value below zero indicates rater leniency (across raters) on a given category. Between the raters who scored using the SR in Round 1 and those who used the RR in Round 1, there was no statistical difference in their severity for each individual rubric category. This information is in the target contrast value (i.e., the difference in logit measure between Group A and Group B), the *t* statistic, and the *p* value. The logit measures from this analysis are graphed in Figure 7. The individual lines represent the differences in logit values between the two rubrics. In Round 1, all raters were rating Content and Organization most severely (with logit values above zero), and all raters were rating vocabulary most leniently (with logit values below zero).

Table 22

*Round 1 Group*Category Bias Interaction Table: Group A (SR) vs. Group B (RR)*

	Group A (SR)			Group B (RR)							
Category	Target Measure	S.E.	Obs-Exp Average	Target Measure	S.E.	Obs-Exp Average	Target Contrast	Joint S.E.	t	Welch d.f.	Prob.
Content	0.06	0.07	0.04	0.19	0.07	-0.04	-0.13	0.10	-1.31	617	0.19
Organization	0.30	0.07	-0.03	0.19	0.07	0.03	0.11	0.10	1.08	616	0.28
Vocabulary	-0.21	0.08	-0.02	-0.28	0.08	0.02	0.07	0.11	0.67	617	0.50
Language Use	-0.03	0.08	-0.02	-0.11	0.08	0.02	0.08	0.11	0.69	616	0.49
Mechanics	-0.10	0.07	0.03	0.01	0.07	-0.03	-0.11	0.10	-1.06	617	0.29

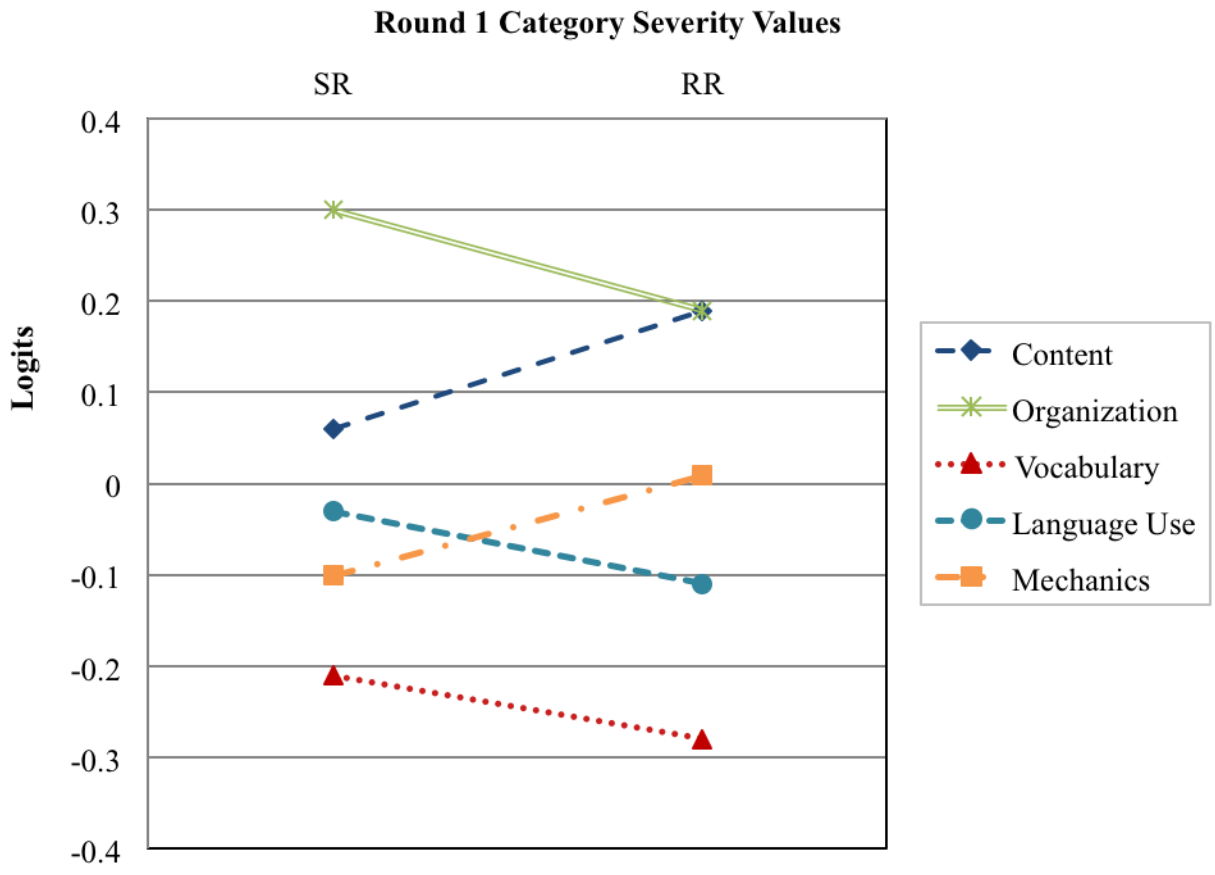


Figure 13. Round 1 category severity values by rubric. SR = Group A. RR = Group B.

For the Facets analysis of Model 2, results of the interaction between Group and Category are in Table 23. Between the raters who scored using the SR in Round 2 and those who used the RR in Round 2, there was no statistical difference in the group's relative severity for each individual rubric category. These results appear in the target contrast value (i.e., the difference in logit measure between Group A and Group B), the *t* statistic, and the *p* value. The logit measures from this analysis are in Figure 8. Similar to Round 1, in Round 2, all raters were rating Content and Organization most severely, and all raters were rating vocabulary most leniently. From Round 1 to Round 2, Group A became less severe in their rating of Content and Organization and less lenient in their rating of Vocabulary, Language Use and Mechanics. Between Round 1 and Round 2, Group B became less severe in their rating of Content, Organization, and Vocabulary, but less lenient in their rating of Language Use and Mechanics.

Table 23

*Round 2 Group*Category Bias Interaction Table: Group A (RR) vs. Group B (SR)*

	Group A (RR)			Group B (SR)							
Category	Target Measure	S.E.	Obs-Exp Average	Target Measure	S.E.	Obs-Exp Average	Target Contrast	Joint S.E.	t	Welch d.f.	Prob.
Content	-0.02	0.09	0.04	0.16	0.09	-0.04	-0.18	0.13	-1.38	551	0.17
Organization	0.24	0.09	-0.02	0.14	0.09	0.02	0.10	0.13	0.77	552	0.44
Vocabulary	-0.19	0.10	-0.04	-0.37	0.09	0.03	0.17	0.14	1.26	551	0.21
Language Use	0.03	0.10	-0.02	-0.08	0.09	0.02	0.11	0.13	0.86	551	0.39
Mechanics	-0.05	0.10	0.04	0.13	0.09	-0.04	-0.19	0.13	-1.43	550	0.15

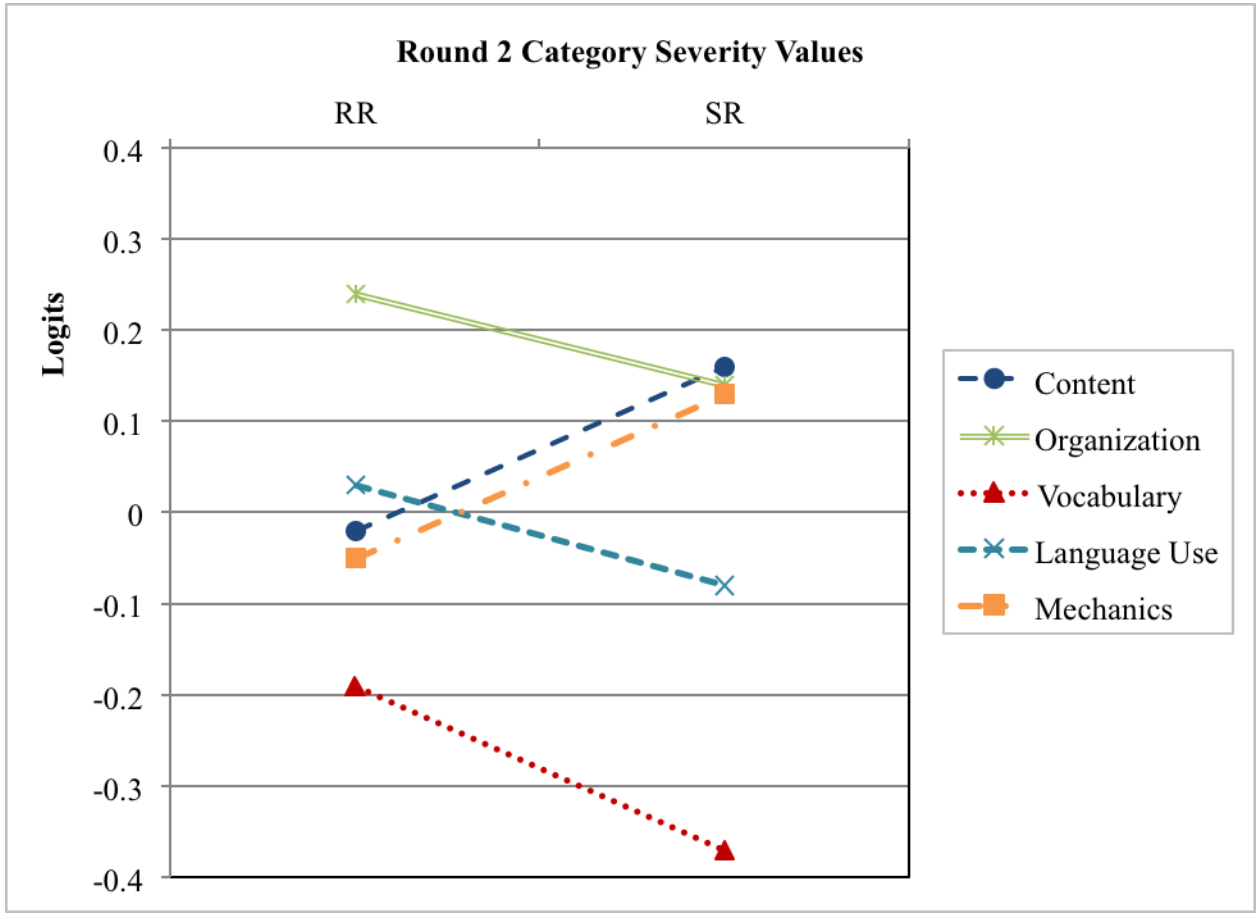


Figure 14. Round 2 category severity values by rubric. RR = Group A. SR = Group B.

Next, I submitted Models 3 and 4 to Facets to examine the role of rubric presentation order (i.e., the order in which raters were trained on the two rubrics) in rater scoring. For the Facets analysis of Model 3, results of the interaction between Group and Category are in Table 24. Between the raters who scored using the SR in Round 1 (Group A) and those who used the SR in Round 2 (Group B), the only statistical difference was for the Mechanics category (-0.27, t = -2.33, p = .02). For this measure, raters who were first trained on the SR rated Mechanics (the last category presented on the SR rubric) much more leniently than Group B, who first encountered the Mechanics category as the first rubric category presented on the RR rubric. Additionally, the contrast values for the Organization category approached significance[2] (.19, t = 1.68, p = .09). The logit contrast values from this analysis are in Figure 9. The horizontal bars represent the target contrast values of relative severity for each rubric category. Bars extending to the left show negative target contrast values, and bars extending to the right indicate positive target contrast values.

For the Facets analysis of Model 4, results of the interaction between Group and Category are shown in Table 25. Between the raters who scored using the RR in Round 1 (Group B) and those who used the RR in Round 2 (Group A), no statistical differences were found. However, both Content (-0.21, t = -1.83, p = .07) and Vocabulary (0.21, t = 1.73, p = .08) approached significance. The logit contrast values from this analysis are in Figure 10.

[2] I use the term "approaching significance" to signify that the p value for a given statistical test is relatively close to the established alpha level that is deemed "significant." In this case, the alpha level is set to .05. Since alpha levels are somewhat arbitrary, Klein (2004) and Larson-Hall (2009) recommended raising the accepted alpha level for exploratory behavioral research in the social sciences to .10, increasing the potential for Type 1 error, but simultaneously expanding the opportunity to find real effects that may otherwise go unnoticed. Hence, in favor of using a more conservative alpha level (.05), I opted to recognize p values that are relatively near the alpha level; I do this because if I had a larger sample, most likely statistical significance would have been reached.

Table 24

*Group*Category Bias Interaction Table: Group A (SR) vs. Group B (SR)*

	Group A (SR)			Group B (SR)							
Category	Target Measure	S.E.	Obs-Exp Average	Target Measure	S.E.	Obs-Exp Average	Target Contrast	Joint S.E.	*t*	Welch d.f.	Prob.
Content	0.12	0.08	0.02	0.21	0.08	-0.02	-0.09	0.11	-0.81	597	0.42
Organization	0.38	0.08	-0.05	0.19	0.08	0.05	0.19	0.11	1.68	597	0.09
Vocabulary	-0.31	0.08	-0.01	-0.33	0.08	0.01	0.02	0.12	0.21	597	0.84
Language Use	0.02	0.08	-0.04	-0.13	0.08	0.04	0.15	0.12	1.31	597	0.19
Mechanics	-0.21	0.08	0.07	0.06	0.08	-0.07	-0.27	0.11	-2.33	597	0.02

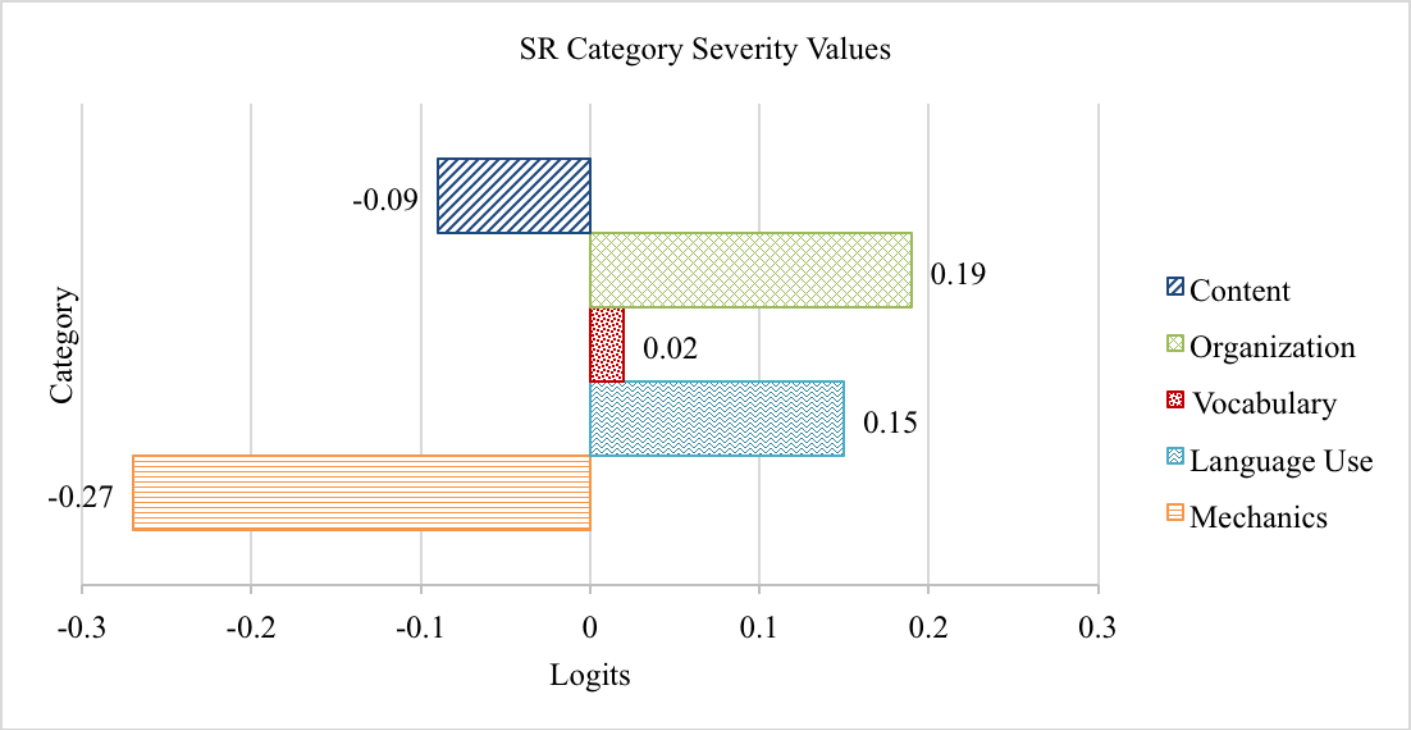


Figure 15. SR category severity contrast values (Group A severity - Group B severity).

Table 25

*Group*Category Bias Interaction Table: Group A (RR) vs. Group B (RR)*

	Group A (RR)			Group B (RR)							
Category	Target Measure	S.E.	Obs-Exp Average	Target Measure	S.E.	Obs-Exp Average	Target Contrast	Joint S.E.	t	Welch d.f.	Prob.
Content	-0.04	0.08	0.06	0.17	0.08	-0.05	-0.21	0.11	-1.83	566	0.07
Organization	0.21	0.08	-0.01	0.16	0.08	0.01	0.04	0.11	0.36	566	0.72
Vocabulary	-0.08	0.09	-0.06	-0.29	0.08	0.05	0.21	0.12	1.73	565	0.08
Language Use	-0.09	0.09	0.00	-0.09	0.08	0.00	-0.01	0.12	-0.05	564	0.96
Mechanics	0.02	0.09	0.00	0.03	0.08	0.00	-0.02	0.12	-0.14	564	0.89

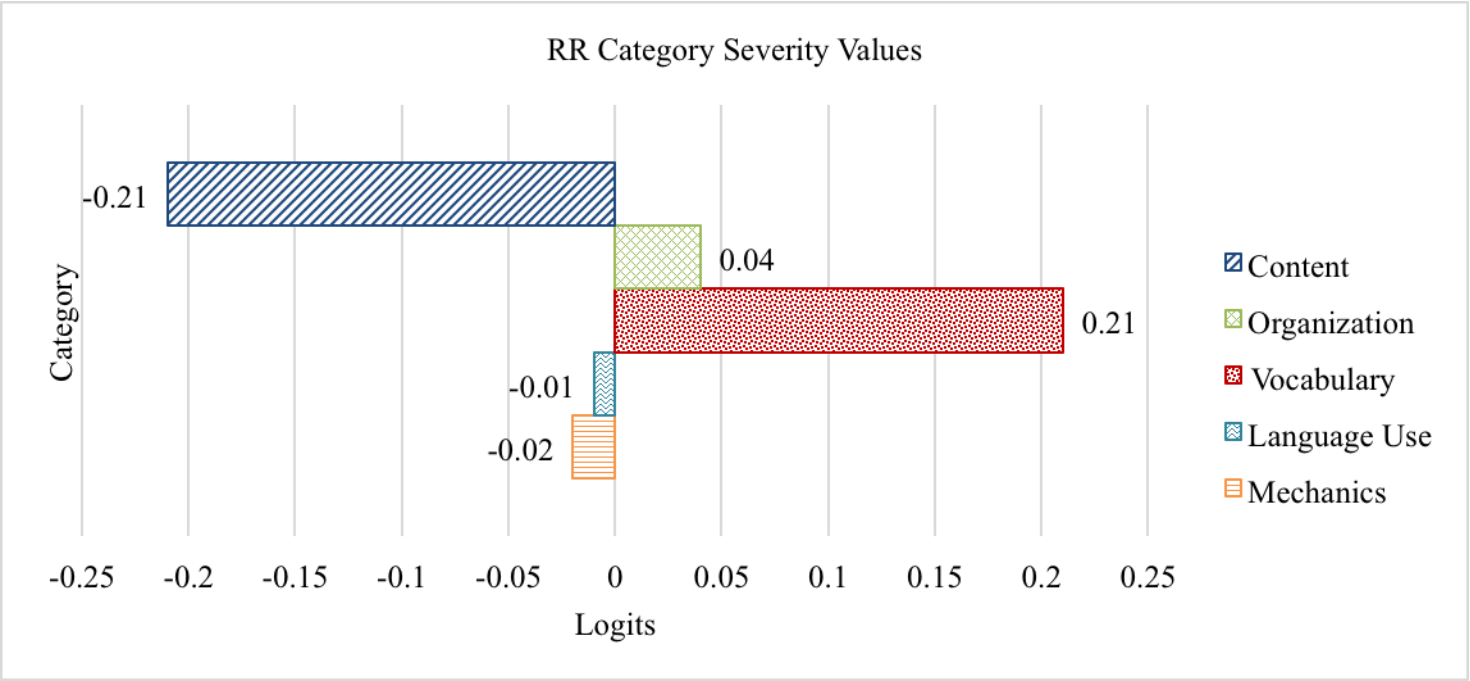


Figure 16. RR category severity contrast values (Group A severity - Group B severity).

Finally, I submitted Model 5 to Facets to examine the relative overall category differences between rater scoring on the SR and RR rubric. Results of the interaction between Group and Category are in Table 26. Between all raters' scoring on the SR and RR, raters' severity on Content (0.16, t = 2.10, p = .04), Organization (0.20, t = 2.59, p = .01), Language Use (0.17, t = 2.04, p = .04), and Mechanics (0.18, t = 2.25, p = .03) were all statistically different. Additionally, raters' severity on Vocabulary (0.15, t = 1.86, p = .06) approached significance. In every category, raters scored more severely on the SR than on the RR. The logit contrast values from this analysis are in Figure 11.

Table 26

*Rubric*Category Bias Interaction Table: Overall SR vs. Overall RR*

	SR			RR							
Category	Target Measure	S.E.	Obs-Exp Average	Target Measure	S.E.	Obs-Exp Average	Target Contrast	Joint S.E.	t	Welch d.f.	Prob.
Content	0.18	0.05	-0.05	0.02	0.05	0.05	0.16	0.08	2.10	1177	0.04
Organization	0.31	0.05	-0.06	0.11	0.05	0.06	0.20	0.08	2.59	1177	0.01
Vocabulary	-0.16	0.06	-0.04	-0.31	0.06	0.04	0.15	0.08	1.86	1177	0.06
Language Use	0.01	0.06	-0.04	-0.15	0.06	0.04	0.17	0.08	2.04	1177	0.04
Mechanics	0.07	0.05	-0.05	-0.1	0.06	0.05	0.18	0.08	2.25	1177	0.03

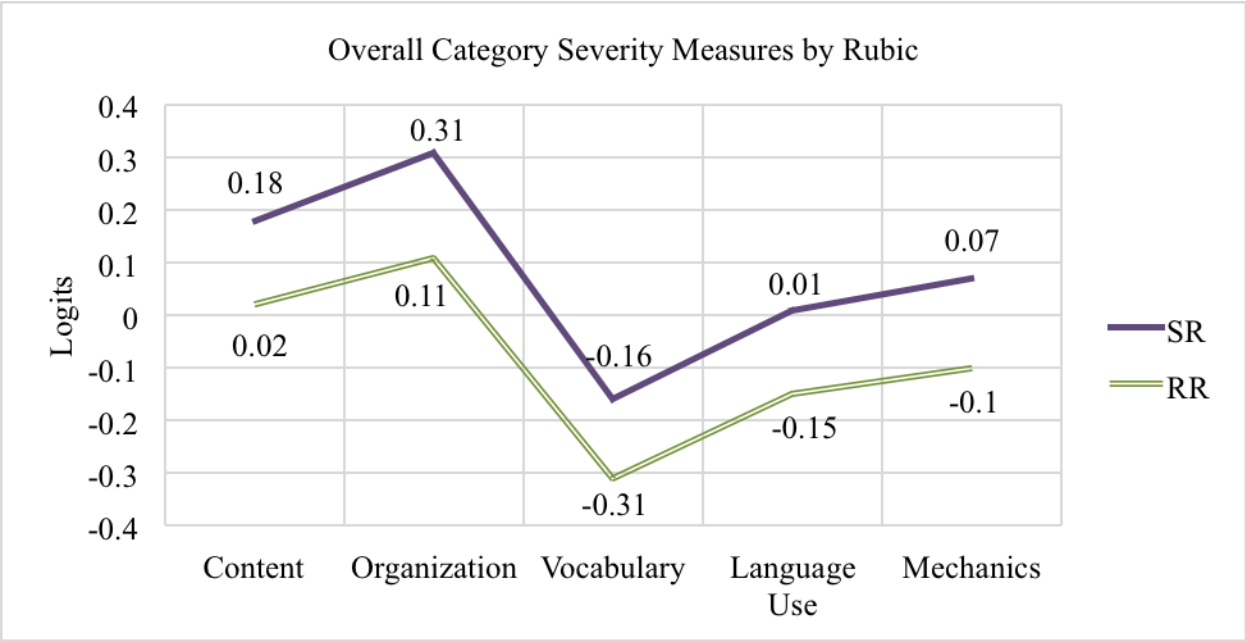


Figure 17. Overall category severity contrast values.

CHAPTER 6

DISCUSSION

A principal underpinning of performance-score validity is that raters score in a manner consistent with the construct being assessed and with the defined measurement goals of the assessment (Bejar, 2012). In writing performance, a crucial step toward upholding the assessment construct and meeting assessment goals is by establishing rater adherence to the defined scoring procedure. Raters undergo training with the goal of bringing all raters to a common understanding of the scoring process, the scoring criteria, and the score points themselves (Athey & McInterye, 1987; Bejar, 2012; Charney, 1984; Follman & Anderson, 1967; Weigle, 1998; Woehr, 1994). While research has shown that rater training improves rater agreement, scoring differences that stem from rater differences or from assessment-related factors still persist (Knoch, 2011; McNamara, 1996; Stuhlmann, Daniel, Dellinger, Denny, & Powers, 1999; Vaughan, 1991; Weigle, 1998, 1999; Weir, 2005). Thus, more research is needed to better understand the great diversity of factors that affect rater scoring so test designers can ensure fair and reliable scores for test takers (Crusan, 2015; Myford, 2012).

Barkaoui (2010) wrote that, in the rating process, raters interact with three texts: the prompt, the essay, and the rating scale. He argued that little is known about how rating scale variation affects raters and rating processes, and that "such information is crucial for designing, selecting, and improving rating scales and rater training as well as for the validation of ESL writing assessments" (p. 56). Though the rubric is only one component of the rating process, the rubric is particularly important because it specifies what raters should attend to, and it ultimately influences the validity of score interpretation and the fairness of decisions that educators make

about students based on the resulting scores (Weigle, 2002). Thus, in the current study I saught to more fully understand how rating scale (i.e., rubric) variation affects raters and rating processes. I investigated how variation in the format of one of the most commonly-used analytic ESL-writing rubrics (i.e., the Jacobs et al. [1981] rubric, revised by Polio [2013]) impacts rater cognition and behavior.

In educational measurement, analytic rubrics have been investigated by test developers and researchers (Cumming, 1990; Lumley, 2005; Smith, 2000; Weigle, 1999) and have been lauded because they draw raters' attention to separate criteria (Barkaoui, 2010; Cumming, 1990; Li & He, 2015). They offer more precise descriptors than holistic rubrics (Knoch, 2009; Smith, 2000), and they can enhance the reliability of scoring performance assessments (Jonsson & Svingby, 2007). However, analytic rubrics can be text-dense and potentially overwhelming for raters. Lumley (2005) found that raters may understand the criteria that are present on a rubric, but may treat the criteria differently in terms of importance. Raters may emphasize some criteria over others. One potential reason for this unequal treatment of rubric criteria may be due to the way that criteria on an analytic scare are organized (Barkaoui, 2007, 2010). One study that provided evidence for this supposition is a study on rater cognition by Winke and Lim (2015). While investigating raters' use of rubric categories, the researchers found that raters' attention to rubric categories and raters' scoring consistency were related to a category's position on the rubric. Winke and Lim speculated that the behavior and scoring that they observed from the raters was due to ordering effects, or more specifically, primacy effects. If this speculation is true, then test developers and researchers should be concerned; research on primacy effects has shown that the order in which information is presented affects one's perceptions of how important that information is, and that initial impressions affect later decisions that are based on

the initial information (Forgas, 2011; Hendrick & Costantini, 1970; Luchins & Luchins, 1970; Rebitschek, Krems, & Jahn, 2015). Considering how these effects may influence raters who train on and use an analytic rubric, the concern is that the rubric format (i.e., order of rubric categories) may elicit a consistently biased pattern of beliefs and responses from a rater, thus impacting score outcomes.

Therefore, given the concern about rubric variation, ordering effects, and potential score impact, in this study I investigated how the ordering of categories in an analytic rubric affects raters' cognition and subsequent behavior. Using a modified version of the Jacobs et al. (1981) rubric (modified by Polio [2013] and also by me for this study, as described in the methods section), in two rounds, I trained two groups of raters on two rubrics that were reverse ordered (i.e., had the same categories presented in mirrored orders). I then had the raters rate a batch of 20 essays. During the training and rating phases, the raters completed a battery of tasks that provided information about their beliefs about the rubric categories, their ability to remember the rubric descriptors, their use of the rubric categories during rating, and their scores for each category and for each essay. This information provided a comprehensive view of the rater's cognition and behavior throughout the training and scoring process. With the data, I investigated how category ordering impacted raters' mental-rubric formation, their rubric usage, and their scoring behavior. What follows is a discussion of the results.

Ordering Effects in Mental-rubric Formation

The goal of rater training is to bring raters onto the same page, that is, to instill in them the same "mental scoring rubric" (Bejar, 2010, p. 4; Cumming, 1990) by which they will score essays. After developing a common mental rubric during rater training, raters should produce scores that are the same regardless of who they are. Thus, the raters should be interchangeable,

meaning that they each score in the same way and produce the same results. To investigate this mental-rubric representation in raters' minds, in this study, raters completed Criteria Importance Surveys (CIS) and the Criteria Recall Tasks (CRT) at several points during training and rating over time. These data gave a snapshot of which descriptors and categories raters considered to be important and which they could recall from the rubric. Within this study, I trained the raters on rubrics that differed as to how the analytic categories on the rubric were ordered. Thus, with this design, I was able to investigate any ordering effects on the raters' beliefs of category importance. In other words, I was able to see if there was a primacy effect in rubric design, which would dictate that the categories that came *first* on the rubric would be viewed by the raters as more important. Conversely, I was also able to see if there is a recency effect in rubric design, which would dictate that the categories that came *last* (that is, those that were most recently reviewed) would be deemed most important.

In the CIS data, one overarching trend emerged in the data. After two groups of raters were trained on the rubric, regardless of the order in which the analytic categories were presented on the rubric, both groups of raters indicated that Organization was the most important category. Raters maintained this belief across rounds, no matter on which rubric they trained. If there had been a clear and direct ordering effect, the order of category presentation would have directly affected the raters' beliefs about what is important. If primacy effects were strongly at work, raters would have considered the left-most categories to be most important during initial training (e.g., Group A- Content and Organization; Group B- Mechanics and Language Use), or if recency effects were at play, raters would have believed the right-most categories to be most important (e.g., Group A- Mechanics and Language Use; Group B- Content and Organization). However, Organization was the prevailing category for both groups (regardless of which rubric

the raters first trained on), suggesting that this group of raters (though novice raters) were bringing with them or picked up during the rating process beliefs about important aspects of quality writing, perhaps learned from their collegiate writing courses or perceived while applying the rubric itself. And this importance criteria transcended primacy or recency, showing that even if there are primacy or recency effects, they are (or can be) softened or mediated by other, essay-quality-intoned importance factors.

There are two other notable trends that pervade the CIS data. The first is that Group A, who trained on the standard rubric (SR) first, consistently demonstrated that they believed that Content was the next most important category (after Organization) and that Mechanics was the least important category, with Mechanics being statistically less important (in the data) than Organization. Not only did these beliefs occur when Group A was trained on and was using the SR, the beliefs about Content and Mechanics also persisted into the phase when the raters were using the reordered rubric (RR). The second trend is that, for Group B, the Mechanics category (appearing first on the RR) was believed to be equally important as all other categories. Content (appearing last on the RR), however, was believed to be less important than Organization, and statistically so. This shows that the position in which Group B raters first encountered the Content category may have influenced their belief about the importance of the Content category, especially because these beliefs were demonstrated at both the Round 1 and Round 2 pre-rating time points. The raters' behaviors suggest a sustained effect of primacy for the Content category, indicated by the raters' similar beliefs about criteria importance, even when the same group has just trained on the two different rubrics.

The primacy effect predicts that initial ordering of information shapes one's impressions about information importance, and these impressions persist over time even when contradictory

information is presented (Hendrick & Costantini, 1970; Luchins & Luchins, 1970; Forgas, 2011). Both trends in the CIS suggest that raters' initial impressions were affected by category ordering and that these impressions persisted even when new category ordering was introduced. Group A initially trained on the SR, in which Content appeared first and Mechanics appeared last. The ordering may have subconsciously affected the raters' beliefs that Content is more important and Mechanics is least important. These beliefs carried through to five weeks after the initial training, and then persisted even after the raters trained on a new rubric and had the opportunity to reconsider their beliefs about category importance. In other words, their beliefs about category importance were consistent with their initial impression and did not seem to change even after receiving different (i.e., somewhat contrary, opposite-ordered) input.

The effects of primacy for Group B, however, were likely also impacted by the pairing of certain categories in certain positions. Raters in Group B indicated that they believed Mechanics (left-most on their rubric of initial exposure) to be equally important as all other categories, yet they believed content (right-most on their rubric of initial exposure) to be less important than any other category. This suggests that the appearance of Content in the right-most position had a negative effect on its perceived importance, which also carried over to Round 2. At the five-week mark, when primacy effects may be strongest, Group B did not have any statistical differences in their category-importance beliefs, suggesting that the order in which the categories were presented had a leveling effect (i.e., indicating equal importance across categories). This was contrary to Group A, who showed a distinct difference at Time 5 in their beliefs about Mechanics. Thus, while both groups demonstrated evidence of primacy effects in their beliefs about criteria importance, there were distinct patterns in the groups' beliefs which related to the position of particular categories.

While the CIS data provided information about raters' beliefs about rubric criteria, the CRTs provided a window into the raters' mental rubric representation. In other words, the CRTs gave raters the opportunity to demonstrate how many descriptors and category titles they held in memory at given time points throughout the study. The raters' recall of the category descriptors serves as evidence of their mental category representation (what the raters hold in their minds) and how this relates to the rubric on which the raters trained. In the CRT data, the predominant trend of all raters was that Mechanics was the easiest to reproduce, suggesting that it was the easiest to remember. Overall, Mechanics had the fewest descriptors within the rubric category, and arguably it was the most similar (in terms of wording) across the levels of proficiency indicated by the score bands and in terms of the concrete descriptors. For example, moving vertically up and down the scale, the quantifiers change from *many* to *several* to *no more than a few* to *no* for both punctuation and capitalization errors. This would make the descriptors easier to recall since they appear with the same language and format multiple times. Additionally, the descriptors refer to more concrete essay characteristics (*spelling*, *punctuation*, *capitalization*, *paragraph separation*), as contrasted with more abstract characteristics like *academic register* or *adequate range of vocabulary* in the vocabulary category (see Appendix C for the full set of rubric descriptors). Previous research on rater-rubric interactions has indicated that raters find concrete, specific descriptors easier to apply, and that vague, abstract descriptors are more difficult to use (Knoch, 2009; Smith, 2000). This may also be the case for raters' ability to recall category descriptors, where concrete descriptors are easier to remember, while abstract descriptors are less memorable.

Just before the first rating session, there was evidence that category order impacted raters' memory. The recall data at Round 1 Pre-rating (Time 2), which was one to two days after

the initial rater training, revealed that all raters remembered a statistically higher amount of the Mechanics category. In addition, Group B raters remembered statistically more for Mechanics than they did for Organization. In this short time span from training to descriptor recall (1 to 2 days), there appears to be short-term evidence that raters' memory of the descriptors in Organization (appearing toward the right of the RR) was affected by the order in which it appeared on the rubric. This may especially be the case since, just after completing this CRT, the same raters indicated that they believed Organization was the most important category. Thus, the fact that the raters believed that the Organization category was important, yet had difficulty recalling what comprised the category, provides evidence that the right-side position of the category may have negatively affected raters' memories.

Five weeks after initial rater training, both primacy effects and category word count may have impacted raters' memories. The recall data from Round 2 Pre-training showed that raters displayed similar behavior, in that both groups had a decrease in recall accuracy for each category, decreasing in accuracy anywhere from 12% (Group A- Organization) to 32% (Group B- Language Use) in their recall of category descriptors. After this long-term break, any evidence of primacy effects should (theoretically) be most salient. What the data show is that Group A's recall of Language Use (which appeared on the right side of the SR) was the lowest and was statistically lower than every other category. Additionally, their recall of Mechanics (which appeared on the right side of the SR) was statically better (higher) than their recall of Content (which appeared on the left side of the SR). While it fell in line with primacy theory that Group A would have lower recall of Language Use, it does not explain why their recall of Content was so poor compared to the other categories. What the primacy theory would predict is that Group A's recall of Content should be one of the highest. However, this effect appears to be

tempered by the Content category having the largest number of words (122), and those words are more abstract in nature. This combination of text density and abstractness may be negatively affecting the raters' recall of the Content category descriptors, making the descriptors more difficult to remember. One piece of evidence to support this notion is the fact that Group B performed significantly worse than Group A on Content recall, evidencing differences between the two groups that may be caused by both category ordering and word count. In other words, were there to only be an effect of word count, the groups should have performed similarly in their recall of the category. However, Group A reproduced much more from the Content category than did Group B. While Group A saw Content in the left-most position, Group B saw it in the right-most position. Thus, this statistical difference in recall provides evidence for primacy effects that persist despite a possible competing effect of category word count.

After training on a new rubric, word count and rubric exposure shaped rater's category memories. After all raters trained on the alternate rubric, the data from Round 2 Pre-rating showed that that all raters recalled similar amounts of descriptors from each individual rubric category. In fact, after training on both rubrics, raters' recall of the categories tended to align with the number or words in the categories, with Mechanics (89 words) and Vocabulary (90 words) being the easiest to recall, and Organization (100 words, Language Use (120 words), and Content (122) being more difficult to recall. Thus, after being trained on the initial rubrics in Round 1, raters' recall seemed to be somewhat affected by category order. However, in Round 2, after being trained on both rubrics, this primacy effect was washed out, and a more qualitative important factor played a role, one that was aligned with category word count. All categories were equal in the raters' memories after more exposure and training on a new, reorganized rubric.

Ordering Effects in Rubric Usage

Raters' Order of Attention

I first sought to uncover the order in which raters read each category to see whether raters consulted the rubric (and presumably scored) categories in a certain order. What the time to first fixation (TFF) data showed is that all raters fixated on the categories in a sequential left-to-right order. This was true regardless of rubric or group. This finding is not surprising, given the strong left-to-right bias in left-to-right written languages such as English. In their study on rater-rubric behavior, Winke and Lim (2015) also found that raters fixated on categories in a sequential left-to-right order. However, in this present study I found that raters followed this sequence no matter the order of the category presentation, meaning that it is not likely that the raters' fixations were due to the categories themselves; rather, these data provide evidence that it is the left-to-right position that influences fixation order. In other words, it did not matter that Content (in the standard rubric) or Mechanics (in the reordered rubric) were fixated on first; they were fixated first *because* they were present in the left-most position. This order of rater fixation is also corroborated by several other researchers who have investigated rater processes. Lumley (2005) and Barakaoui (2010), who investigated rating processes through think-aloud protocols, and Winke and Lim (2015), who utilized eye-tracking methodology, all found that raters utilized the rubric and scored categories in the order that the categories appeared (from left to right) on the rubric.

Rater Behavior and Category Importance

Raters' concentrated attention to rubric categories

One way to measure raters' perceptions about category importance is through their rubric-use behavior, specifically through their total fixation duration (TFD) on a given category.

What the controlled TFD data showed is that, on the SR, both groups had high attention to the left-most categories (Content and Organization), and raters paid the least attention to right-most categories (Language Use and Mechanics). On the RR, however, raters' attention to the four left-most categories (Mechanics, Language Use, Vocabulary, and Organization) were similar across categories, but raters' attention to Content (the right-most category) was lower. From these descriptives, it seems as though the primacy effect may be the strongest for the SR, in which there is a decline in attention (on the four right-most categories) as the raters read from left to right. Winke and Lim (2015), who used a version of the Jacobs et al. (1981) rubric similar to the SR (they only used one rubric in their study), also found similar attentional patterns: Raters paid the most attention to Organization and Content (the left-most categories) followed by Vocabulary, Language Use, and lastly, Mechanics (the right-most category). On the RR, however, there seems to be an equalizing effect for the first four categories, to which raters in both groups paid equal attention. As for Content on the RR, raters paid the least attention to this category, perhaps because of its right-most position.

Between rubrics, changes in raters' attention were most salient for the Mechanics category. I examined the differences in the Groups' attention to a category between the two rubrics in order to see whether there were any trends related to category movement. For Group A, the raters paid more attention to Content and Organization in the SR and more to Language Use and Mechanics in the RR. Additionally, their change in attention to the Mechanics category was statistically different (larger) than that for Content, Organization, and Vocabulary. This pattern of attention is predicted by the primacy theory, which states that whatever is presented first receives the most attention. This pattern is particularly clear in Group A's attentional behavior; the raters paid more attention to the two left-most categories on a given rubric than

when the same categories appeared in the right-most positions. Furthermore, Group A's change in attention was most marked for Mechanics, perhaps because it moved from the right-most positions (and was thought least important) to the right-most position, which indicates importance. For Group B, raters fixated equal amounts on the Content category between the two rubrics, and their pattern for attention to the remaining categories increased in alignment with the reordered rubric (i.e., the farther left the category appeared on the RR, the more attention the raters paid to it as compared to their attention on the SR). Additionally, the raters' change in attention to Mechanics was significantly larger than all other categories, and their attention to Mechanics was significantly different between the two rubrics, indicating a strong primacy effect on the Mechanics category. For all other categories for Group B and for all categories for Group A, their fixations on a given category between the two rubrics were not significantly different.

This data suggest that Group B was more susceptible to primacy effects when rating on the SR; while their attention to categories on the RR were relatively similar (suggesting category equality). There was a clear pattern of *attention decrement* (Crano, 1977; Forgas, 2011; Hendrick & Costantini, 1970; Tulving, 2008) as the raters moved farther right on the rubric. This would suggest that the ordering of the categories in the RR somehow draws more equal attention to the categories, and I posit that this is because on the RR, the technical, *local* categories (Mechanics and Language Use) appeared in the left-most positions, and the more holistic, *global* categories (Content and Organization) appeared in the right-most positions. Ordering aside, I suspect that novice raters would tend to think that the more global categories are more important due to the more holistic or impressionistic nature of the category, and that the categories that deal with more technical, local issues (which arguably take more training to attend to and decipher) would naturally be thought of as less important. A tendency to attend to more global categories was

demonstrated by raters in Barkaoui's (2010) study. While scoring using a holistic rubric (which does not as strictly constrain what constructs inform raters' scoring), raters focused most on organization and overall communicative quality as they assigned holistic essay scores. Thus, I suspect that, on an analytic rubric, fronting the local categories implicitly conveys that the local categories are important and, in a sense, raises them to equal importance with the more global categories. Conversely, when the local categories are placed at the right-side of the scale, according to the primacy effect, this would naturally cause raters to think that the categories are less important, and this effect compounds with (or is reinforced by) the novice raters' natural beliefs about the relative (un)importance of the categories. Because both groups had statistical differences in their attention to Mechanics (either between categories or between rubrics), there is evidence to support this conjecture. Thus, in the SR, the raters showed an attention deficit between the left and the right sides of the rubric, but in the RR, raters paid more equal attention to the categories, perhaps as a function of the nature of the categories themselves. Both Group A and Group B displayed these attentional patterns on the SR and RR.

The results demonstrate that raters do not pay equal attention to all rubric categories. This is corroborated by Winke and Lim (2015), who likewise found through eye-tracking metrics, that raters did not pay equal attention to the rubric categories while rating. Lumley (2005) and Barkaoui (2010), on the contrary, found that the raters in their studies did pay equal attention to the rubric categories, but Lumley and Barkaoui documented the raters' attention through think-aloud protocols (TAPs). As Winke and Lim posited in their study, these contradictory finding could be attributed to differences in data-collection methods. While TAPs have the benefit of offering moment-by-moment insight in to raters' thoughts and rating processes, this method has been criticized for being prone to reactivity (the methodology itself

causing a change in the rater's behavior) and veridicality (the methodology not being able to fully capture the raters' thoughts and processes). In TAPs, asking raters to talk though their rating processing may, by nature, cause them to behave differently (e.g., focus more equally on all rubric categories, potentially for the benefit of the researcher), thus leading to unnatural and atypical evidence on rating behavior. One benefit of eye-tracking methodology, on the other hand, is that it is unobtrusive, and thus is less likely to interfere with the natural rating process. Hence, it may be the case that the behaviors of the raters in the current study and in the study of Winke and Lim may represent more ecologically valid (true-to-life) rating behavior; thus, raters may not (in real-life) pay equal attention to all rubric categories.

Raters' frequency of attention to rubric categories

Another metric I used to quantify raters' attention to rubric categories was the number of times the raters visited a category (visit count; VC). Past researchers have used back-and-forth essay-to-rubric reading behavior in think-aloud protocols (Barkaoui, 2008) and visit count in eye tracking (Winke & Lim, 2015) as an indicator of attention and category importance, as raters read and return to a category to reread the descriptors. In the descriptive VC data in this current study, two main trends surfaced: (1) raters visited the Organization, Vocabulary, and Language Use categories the most; and (2) raters visited the Content and Mechanics categories the least, and the number of visits to these two categories alternated in correspondence with the category location. It is not surprising that the middle three categories had the highest visit counts, as the eye would likely fall toward the middle of the rubric when the rater is orienting back to the rubric in order to search for a specific category or descriptor. For example, raters moved back and forth between reading the rubric and reading the essay, and each time the raters return to the rubric, they must return to the category where they left off or locate the category they are moving to

next. Each time the eye falls anew on the rubric, it would likely be toward the middle of the screen (i.e., the middle of the rubric), thus inflating the visit counts for the middle three categories. The outer-most categories, however, would not likely be as susceptible to this type of visit count inflation, as most fixations that fall in the outer-most categories are more likely to be intentional. Given this, the alternation in visit counts for Content and Mechanics is telling. For both groups on the SR, raters visited the Content category the least, and it was in the right-most position. Conversely, in the RR, when the Mechanics category was in the right-most position, raters visited it the least. This is corroborated by Winke and Lim, who also found that raters visited the three middle categories most frequently, followed by Content (in the left-most position), and lastly Mechanics (in the right-most position). Thus, the data suggest that despite inflated visit counts on the middle three categories, the raters' visits to the outer-most categories provide evidence for primacy effects that align with category position.

The mean differences for visits between the two rounds follow similar patterns as the TFD data. Group A had more visits to a category relative to its fronting on a rubric. For Content and Organization, the raters visited these categories more when they appeared on the left side of the rubric (on the SR), and they visited Mechanics more when it was in the right-most position (on the RR). For Group B, they visited every category more often on the RR, but the pattern showed that the farther left a category was on the RR, the more visits raters made to the category, as compared to where it was located on the SR. The only statistical difference in visit behavior between rubrics was Group A's visits to Content, which were more frequent on the SR. This significant difference in Group A's attention to Content, paired with the alternating trend (dependent on category location) do suggest that primacy may be at play in the raters' attention (as measured through visit frequency) to the categories.

Since the VC metric can be prone to inflation from the "reorienting fixations," it is helpful to consider the VC findings in tandem with the TFD results to better understand the raters' attentional behavior. Similar to the VC findings, the TFD data corroborate the alternating attention to the outer-most categories, which seems to be position-dependent. Both metrics showed that raters paid less attention to a category when it was in the last (right-most position), suggesting a primacy effect.

Another way to look at the raters' visit behavior is through their category-skipping behavior. If raters skipped a category and did not consult that rubric area before assigning a score for that respective category, two things may have happened. First, the raters may have found the category less important and thus paid less attention to it (and rather may have just assigned a score that aligned with the score previously given on the other categories). Second, the raters may have internalized the category descriptors and, therefore, were able to assign a score without consulting the rubric. The category-skipping data show that raters most often skipped the Mechanics and Content categories, and this corresponded to the location of the category on the rubric; when Mechanics and Content were last on the rubric, raters skipped them more often. Winke and Lim (2015) observed this same pattern of behavior in their study; raters most frequently skipped the Mechanics category (in the right-most position on the rubric). Comparing the behavior of the raters on the two rubrics confirms that raters are most likely to skip a category when it is in the right-most position, and this was true whether the category was Content or Mechanics. Given the average number of skips for the Mechanics category as compared to the Content category, I would argue that there may be two forces as work in the raters' behavior. There is evidence that primacy is affecting the raters' attention to the Mechanics category, but I believe that this may also be concurrently at play with the raters'

internalization of the Mechanics category. In the CRT, raters had the best recall of the descriptors in the Mechanics category. If raters are able to hold these descriptors in their minds (which they consistently demonstrated on the CRTs), it seems plausible that they would utilize the rubric less when making scoring decisions for the Mechanics category. Conversely, in the CRT, raters found the descriptors in the Content category more difficult to recall, suggesting that they would need to rely more on the rubric (i.e., read it more often and perhaps for longer periods of time) when making scoring decisions. The fact that raters did skip this category suggests that primacy was at play, leading raters to pay less attention to and skip over the category.

To summarize the investigations into the primacy effect and raters' attention to the rubric categories, one trend emerged across the three metrics I used: the raters' attentional behavior regarding the outer-most categories (i.e., left-most and right-most) seem most susceptible to ordering effects. The rater's order of category fixation, the raters' amount of time spent fixating on the categories, the rater's number of visits to and skips of the outer-most categories all point toward an influence from the categories' positions on the rubric.

An additional trend that emerged in the data was a possible effect of category type on the raters' attentional behavior to the rubric during rating. Of the constructs represented on the rubric, the Content and Organization categories tended toward holistic, global, impressionistic constructs, whereas Language Use and Mechanics fall more toward the technical, local constructs. Novice raters may have a natural inclination toward considering global measures as more important (see Barkaoui, 2010). These categories may seem easier to gauge because they tend to be more impressionistic, potentially leading raters to feel more confident in their ability to score these categories, which in turn may lead them to think they are more important. Novice

raters may likewise consider the more technical, local measures as less important (and more difficult to score) because these measures take greater training and experience to be able to feel as though a rater reliably notices such characteristics in an essay (e.g., sentence variety, sentence complexity, and vocabulary frequency). One fact to support this claim is that raters in the current study had the lowest interrater reliability for the Language Use category. One rating study may corroborate this speculation; in Shi's (2001) study on rater beliefs and scoring, the researcher asked a group of native-English-speaking raters to score a batch of essays on a holistic ten-point scale. The raters were given no criteria on which to base their judgments; rather, the researcher asked the raters to assign a score on the scale and to document the top three reasons that they assigned a given score for an essay. Shi found that the raters focused primarily on content- and organization-related features, followed by linguistic features. Assuming a propensity for raters to bend toward content and organizational features in their decision-making, I speculate that the presentation of the RR (with the more technical, local categories appearing first) may have lead the raters in the current study to attend more to these categories, bringing the categories on equal grounding with the more global categories. One such exemplary piece of evidence for this speculation is that Group B, when rating on the RR, skipped each category about the same number of times. Another piece of evidence may be that Group B's concerted attention to each category was roughly equal, as compared to Group A's category attention on either the SR or the RR (after their beliefs had already been influenced by the ordering of the SR).

Scoring Impact and Category Importance

A crucial component of considering how ordering effects may influence rater behavior is in examining the end result, the scores. To this end, I sought to examine whether and how the

ordering of the categories affected (1) the raters' score agreement on each category and (2) the raters' scores on the five individual categories. I did this by first calculating Intracclass correlations (ICC), a measure of interrater reliability, and then by submitting the raters' scores to Rasch analysis to examine bias effects (lenience and severity) based on category and category position. The ICC showed that, in both rounds, interrater reliability was very high, the lowest being a .88 (.7 and higher being an acceptable reliability coefficient; Brown, Glasswell, & Harland, 2004), and there were very small differences between categories and groups, demonstrating that the raters were reliable (i.e., consistent) in their scoring behavior.

The interrater-reliability patterns align with category position. In Round 1, raters had the highest agreement on the categories that appeared in the left-most position. Winke and Lim found this same trend, that raters agreed the most on the left-most categories (Content and Organization). In Round 2, however, the highest agreement was for Content, Organization, and Mechanics. Thus, in Round 1, the data could support the notion that raters' behaviors and subsequent scores are affected by category order, with raters' attention to the left-most category resulting in greater rater agreement. In Round 2, there could be a combined effect where raters are agreeing most on the categories that appeared left-most in Round 1 *and* in Round 2. Another finding common to Winke and Lim's study was that raters were less reliable in their scoring of Language Use. Citing Smith (2000) work on rating scales, Winke and Lim argued that this category may be less reliable because its descriptors may be less precise and clear than other categories. Both Knoch (2009) and Smith (2000) argued that when rubric descriptors are detailed, precise, and clear, the descriptors and categories can lead raters to a better understanding of the rubric, and thus, to more reliable scores. In the current study, the Language Use category contained 120 words and six descriptors. The wordiness, and perhaps the

complexity of the descriptor propositions themselves, may have lead raters to be less reliable in their scoring of the category. Smith (2000) and Burrows (1994) also found that raters had difficulty in interpreting and applying the language-category descriptors, which lead to lower reliability among raters. Thus, it seems that in the current study, raters had the highest reliability on Content, Organization, and Mechanics, perhaps as a result of ordering effects, but Language Use did not benefit from any front positions in the RR due to raters' potential difficultly in applying the category descriptors.

Turning to the data on raters' scoring severity, the Rasch data showed one overarching trend: raters scored Organization and Content most severely, and they scored Vocabulary, Language Use, and Mechanics more leniently. This general pattern was true for both groups of raters and both rubrics. This greater severity on (primarily) the Organization category and secondly on the Content category is not surprising since both groups indicated that they thought each category was somewhere between *important* and *very important* on the CIS scale. Group A indicated that Vocabulary, Language Use, and Mechanics were less important, but Group B's indications of category importance were fairly similar between Content, Vocabulary, Language Use, and Mechanics. Eckes' (2012) research on raters' perceived importance and rating severity corroborate this finding. He also found that when raters perceived a category as being more important, they were more likely to rate the category more harshly.

Despite this general similarity in scoring between Group A and Group B, there was one significant difference in Group A and Group B's scores. In comparing the SR Round 1 (Group A) scores to the SR Round 2 (Group B) scores, the Rasch analysis uncovered that the raters who had first trained on the RR were much less lenient in their scoring of the Mechanics category. One explanation for this behavior is that Group B, who first trained on the RR, saw the

Mechanics category in a dominant (highly salient) position, thus leading to a perceived importance of the category. On the other hand, at this point, Group A has only been exposed to the SR, in which the Mechanics category appeared in the weak position (right-most position), potentially causing raters to believe Mechanics was less important and thus leading them to score the category more leniently. This difference in category perception, caused by a primacy effect, could have led to this difference in scoring severity between the two groups. In other words, the order in which the groups trained on the rubric seemed to play a role in the raters' subsequent scoring on the Mechanics category, which surfaced as a difference in relative category severity. Also noteworthy is that when comparing Group A and Group B's RR scores, the two groups scored the Mechanics category with very similar severity. At this point Group A had already been exposed to the SR (on which they leniently scored the Mechanics category), and this leniency from the SR did not seem to transfer to the RR. I posit that the fronting of the Mechanics category played a role in Group A's change in scoring, as it did with their concentrated attention and number of visits to the Mechanics category when it appeared at the beginning of the rubric.

Category ordering effects may have also caused halo effects in the rating process. A halo effect is when a score that a rater assigns for a given rubric category may be influenced by the scores the rater has just assigned on the previous category and by the general impression the rater is forming while assigning scores (Yorozuya & Oller, 1980, p. 145). Knoch (2009) argued that halo effects could be an artifact of a rubric itself, which may be the case in the current data set. It is possible that the score given on the first category or perhaps the first two categories (i.e., the left-most categories) influences the subsequent assignment of category scores. In the overall data from Model 5, I found that all raters scored every category more harshly on the SR than on

the RR. A potential explanation for this is that encountering Content first (on the SR) could have led raters to rate the subsequent categories more harshly because they were "conditioned" to rate more severely solely by encountering a "more important" category first. Conversely, starting with the Mechanics category (on the RR) could have set raters on a more lenient scoring path. This theory is supported by the CIS data, in which raters indicated that they thought Content and Organization to be more important and Mechanics to be less important. Since Eckes (2010) also found that there was a relationship between scoring severity and criteria importance, it would follow that the combination of category position and perceived category importance would lend itself to halo effects.

Main Findings and Implications

In this study, I examined various aspects of the rating process (e.g., mental-rubric formation, rubric usage during rating, and rater scores) in order to develop a preliminary understanding of how ordering effects may influence raters' cognitive processes during rating. The many facets of data seem to tell the same story: as novice raters train on a new rubric and assign scores using the individual categories on the rubric, the raters' behavior pertaining to the outer-most positions (e.g., left-most and right-most) seems most susceptible to ordering effects. That is, the findings of this study have provided some evidence that the position of a category affected the raters' beliefs about what criteria are the most and least important when scoring an essay, how many descriptors raters were able to recall from a category, how much attention raters paid to a category on the rubric while rating, and how severely raters scored a given category. While these effects were not always present for both groups, the data did suggest that the category itself and the position matters. That is, the nature of the category and the order in which the categories appeared on the rubric mattered. Overall, Group A's behavior

demonstrated a primacy effect on the Mechanics category. Group A initially encountered Mechanics in the right-most position on the rubric, thus indicating (subconsciously through the primacy phenomenon) that it was less important. This was compounded with the fact that Mechanics may be considered less important on its own because it has the fewest words and because it is a more technical, local construct. This quality of the Mechanics category, compounded with its placement in the final position on the rubric, produced negative patterns in rater behavior related to the Mechanics category. Group B, on the other hand, initially encountered Mechanics at the beginning of rubric, thus indicating (subconsciously through the primacy phenomenon) to the raters that it was more important. Here, the interplay between the primacy effect and raters' beliefs about Mechanics could explain why Group B considered Mechanics to be equally important to the other categories on the rubric. I describe this as a leveling effect, in which raters' thoughts and behaviors regarding the Mechanics category were similar to their behaviors regarding other categories. Additionally, there was evidence of a halo effect, in which the first category affected raters' scoring severity in the subsequent categories.

Primacy does not explain everything, though. Even though the participants in this study were all novice raters who had never scored essays or been trained on a rubric, these raters still brought their own ideas about quality writing to the task. These ideas could have come from their own experiences in collegiate writing courses, of which all of the participants had completed at least two. While raters do bring their own biases with them (Barkaoui, 2011; Brown, 1995; Cumming, 1990; Johnson & Lim, 2009; Kang, 2012; Lumley & McNamara, 1995; Weigle, Boldt, Valsecchi, 2003; Winke, Gass, & Myford, 2012), this study has shown that rater training and exposure to a rubric can effectively shape raters' beliefs and behaviors in the scoring process, as has been expounded upon before (Davis, 2015; Shohamy, Gordon, & Kraemer, 1992;

Weigle, 1994; Wolfe, Matthews, & Vickers, 2010). Though this study only involved two rounds of rating over the course of five weeks, many raters in rating programs score essays over a duration of long sequences of time. I posit that as raters continue to be exposed to and rate on a single rubric, the primacy effect could potentially cause a deep entrenchment of beliefs that shape raters' scoring behavior, as Luchins and Luchins (1970) showed that primacy effects grow stronger over time. This is problematic because, on many analytic rubrics, the categories are meant to be treated with equal importance and should be scored accordingly (Lumley, 2002). However, ordering effects could lead raters to stray from ideal rating behavior, thus compromising the essay score interpretation, and thus test validity.

Overall the results show that the psychological phenomenon described by psychologists like Underwood (1975)—that people assign more importance to information that comes first on a list—is important for foreign and second language performance testing programs that have raters who use analytic rubrics. Information ordering (on a rubric) matters in how raters process, retain, and make scoring decisions. In sum, the *attention decrement hypothesis* can be applied to the context of rating programs (that use analytic rubrics) in applied linguistics. The exact words from the psychologist Crano (1977, p. 9) should be heard and taken to heart by language testers, that there is a "progressive decline in attention to trait descriptors over the course of a complete list… [and later information is] less heavily weighed in the process of impression formation. The relative influence of a descriptor varies as a function of its serial position." I suspect that the findings in my study (that raters pay less attention to later-presented information on an analytic rubric) would be supported by biological evidence, as Tulving (2008) proposed. As I explained in the literature review, Tulving noted that the biological process called *camatosis* (a slowing of the neuron activity in the brain) is related to information retention. As people work through a list

of information, the theory is that neural activity decreases, which results in less attention being paid to information presented later on the list. My eye-tracking data showed that there is less visual focal attention on later-presented rubric information (as also shown by Winke & Lim, 2015), but I do not have direct evidence of a reduction in neural activity as raters work through information on a rubric. The eye-tracking data only showed that raters paid less visual (cognitive) attention to information presented last on the rubric: The eye-tracking data do not correspondingly show that *less* neural activity overall was occurring (the raters could have simply been paying attention to the essays *instead* of the rubric, for example). Whether raters pay less attention (neurologically) over time when using an analytic rubric could be investigated in the future by combining eye-tracking methods with electroencephalography (EEG), which is common method of tracking neurobiological dysregulation. Eye-tracking and event-related potentials (ERPs; a more nuanced metric used to quantify or segment EEG information) have been co-registered in neuroscience investigations into reading (see Henderson, Luke, Schmidt, and Richards, 2013, for an overview). And in a language testing context and in a study on rater behavior, such dual data collection methods (eye-tracking and ERP information) could be extremely revealing. This is because the order in which information is presented and the amount of attention paid to the information has a measurable impact on subsequent decisions (Forgas, 2011; Hendrick & Costantini, 1970; Luchins & Luchins, 1970; Rebitschek, Krems, & Jahn, 2015). Understanding the amount of visual attention paid to rubric categories over time in combination with the amount of neural activity employed in the overall rating process could help rating program designers better understand the importance of breaks, rater recalibration, and the need for recurrent training on the rubric categories.

Provided the findings of this study, it would be beneficial for test designers to carefully

consider the layout and ordering of analytic rubrics used in operational testing. Rubric designers could leverage ordering effects to their benefit by fronting any categories that are typically seen as less important or have lower interrater reliability scores. Test designers may also want to consider making word count similar across categories (as done by Polio [2013] in her paper on revising the Jacobs et al. [1981] rubric) and striving for clarity and precision in each individual descriptor in order to reduce the amount of rater interpretation needed for a descriptor, as requested by Knoch (2009).

Given that raters may become more and more entrenched in their beliefs and scoring patterns when rating over long periods of time, test designers could also consider creating a online rater-training and scoring platform (see Knoch, Read, & von Randow, 2007; Wolfe, Matthews, & Vickers, 2010) which would encourage raters to pay equal attention to each rubric category. One example may be a digital platform that presents raters with a randomized, forced order of training, norming, and scoring. For each essay, the platform could randomly prompt raters to score a given category, only allowing raters to score one category at a time and input scores for the category appears on the screen. This may reduce rater's conditioning to attend most to certain categories while least to others. Additionally, many researchers advise having two raters score each essay (Elder et al., 2007; Lumley & McNamara, 1995; Marzano, 2002; McNamara, 1996), and if raters trained and scored on categories in a random order, then pairs of raters would provide a more balanced scoring scheme and would be an additional step to mitigate any effects of primacy on scoring.

In the case that rating programs intend certain categories to be more important, those categories should be left-most, and training should indicate that the left-most categories are more important and explain why. I suspect that this is being done subconsciously in rating programs

that use analytic rubrics. The rater training most likely has the new raters learn about the categories in the order they are presented (from left to right on the rubric). The rater trainers most likely work through sample scoring scenarios using the rubric from left to right, and may even unintentionally spend more time explaining the left-most categories. This ordering may have an effect on mental rubric representation, how raters view the importance of the categories, and how well certain categories are used over time. This study shows that ordering effects are real. Rater training programs now need to use that information to better design rating programs such that any ordering effects are intentional and to the betterment of the program, or the category ordering needs to be controlled so that ordering effects will not take hold and be detrimental to the rating program over time.

CHAPTER 7

CONCLUSION

The nature of this study was exploratory, investigating possible effects of primacy on raters' cognition and behaviors. Because this study was the first of its kind, empirically investigating ordering effects in rubric format, there were a number of methodological areas that could be improved in future research. First and foremost, the sample size in this study is rather small with only 31 participants. Secondly, the participant population that I chose for this study was strategic. In order to limit the amount of category bias that participants might bring into the study, I opted to use novice raters that had no experience rating essays or applying rubrics to a piece of writing. This methodological decision provided the opportunity to have "blank slates" in the training process, thus offering cleaner data. Simultaneously, this decision to use novice raters, along with having a small sample size, leads to less generalizable findings since it is unclear how experienced raters may respond to ordering effects in a novel rubric.

An additional issue that surfaced in the data is that visit count (VC) may be a problematic indicator of rater attention to rubric categories. This measure seemed susceptible to reorienting behavior (i.e., after reading the essay, reaffixing on the middle of the rubric before moving to the intended category). In this study, I used the VC measure in tandem with other measures to clarify raters' behavior. However, a more methodologically rigorous technique may provide cleaner VC data. For example, the data could be hand-coded for true reading-based fixations, and this would limit the inflation of VC due to reorienting fixations. In the future, eye-tracking technology platforms may consider building features that are intended for research paradigms where participants intentionally look away from the screen and then refocus on the screen, such

as in rubric-use research.

In this study, I employed a cross-over design, in which both groups of raters trained and scored in both rubrics. While this design allowed for examination of retraining effects from being trained on the alternate rubric, it also made data modeling difficult. For future research, a more simple between-subject design with a large participant sample over a longer period of time would provide further evidence for long-term effects of primacy on rater cognition and behavior. In particular, this would allow for the investigation of the entrenchment of raters' beliefs and scoring behavior related to category-ordering effects.

Finally, future researchers could investigate primacy effects from a more nuanced framework. While the current study has shown that primacy effects plays an influential role in rater behavior, there remains the question of how individual raters may be impacted differently by ordering effects and to what extend these ordering effects impact raters' beliefs and behaviors on an individual basis. As I suggested in the discussion section, a study investigating rubric primacy effects using both eye-tracking and EEF methods could reveal find-grained insights into raters' internal cognitive processes involved in foreign and second language rating tasks.

The motivation for this investigation has been to examine how a widely-accepted psychological phenomenon (primacy effects; see Crano, 1977) could be covertly impacting the ways in which raters interpret and apply an analytic rubric. With evidence that primacy effects may be at play in the rating process, researchers and rater trainers alike must continue to improve testing and training in a fashion that upholds the fundamental propositions of fair testing (e.g., that raters "attend to the criteria included in the rubrics when making their judgments; raters do not let construct-irrelevant criteria enter into their judgments; raters continue to use the rubrics appropriately… with no evidence of changes in their application of the rubrics over time";

Myford, 2012, pp.47-48). As Jacobs et al. (1981) themselves posited, it the duty of assessors to 'ensure more consistent interpretation and application of the criteria and standards for determining the communicative effectiveness of writers' (p. 43)."

The next step is for assessors to consider the analytic rubrics they have, and to think of ways to redesign their training and rating processes to mitigate any primacy effects their rubrics may unintentionally be imposing.

APPENDICES

APPENDIX A

Variable Operationalizations

Table 27

Variable Operationalizations

Variable Name	Definition	Measurement	Variable Type	Variable Type
Rater attention	The amount of attention a participant pays to a given rubric category	TFD, TFF, VC		
Total fixation duration (TFD)	The total time a participant spends fixating on a rubric category while rating an essay	In milliseconds, the sum of the duration of all eye fixations within an area of interest	Dependent	Continuous
Time to first fixation (TFF)	How long it takes before a participant looks at a rubric category while rating an essay	In milliseconds, how long it takes before a participant fixates his or her eyes on an area of interest	Dependent	Continuous
Visit count (VC)	How many times a participant looks at a rubric category while rating an essay	the number of eye visits within an area of interest	Dependent	Continuous
Time	The point in the study at which the data is collected (e.g., before rater training, after rater training, before rating, etc.)	The time at which the data for a given task was collected, labelled ordinally, from the beginning of the study to the end	Independent	Ordinal
Category	A vertical subsection of the rubric which contains descriptors related to one topic	Rubric subsection: content, organization, vocabulary, language use, mechanics	Independent	Categorical
Category position	The position in which a given category appears on the rubric	The ordinal position in which a category appear on the rubric (i.e., 1, 2, 3, 4, 5)	Independent	Ordinal
Category score	The score participants assign to a given category (e.g., content, organization, etc,) during rating	The score assigned by raters to each category	Dependent	Continuous
Mental rubric	What participants hold in their minds from the rubric	Criteria retention and criteria importance (see below)		
Criteria retention	How well participants remember each rubric category	Category accuracy score from criteria recall task	Dependent	Continuous

Table 27 (cont'd)

Criteria importance	Participants' beliefs about how important a criterion is in the scoring process	Likert-scale ratings from the criteria importance survey	Dependent	Ordinal
Order of rubric exposure	The order in which participants are exposed to the rubrics	The order in which the participants are exposed to a rubric (i.e., 1 or 2)	Independent	Ordinal
Rater Decision Making Process	The process by which a rater arrives at a final score for an essay, including what rubric-based criteria are considered when deciding on a score	The outline produced from the decision-making-process outline (DMPO) task and a subsequent interview about the DMPO	Dependent	Qualitative
Group	The experimental group to which the participant belongs, which is defined by the order in which the participants are exposed to the rubrics (i.e., receiving the standard rubric or the reordered rubric first)	Group A: SR, RR Group B: RR, SR	Independent	Categorical

APPENDIX B

MSUELT Prompts

Write as much as you can, as well as you can, in an original, 35-minute composition on ONE of the topics below.

1. Many people dream of winning millions of dollars in a lottery or other contest. However, winning that much money can have a negative impact on someone's life. In your opinion, what are the advantages and disadvantages of winning a large amount of money? Be sure to support your ideas with specific explanations and details.

2. Think about a time in your life when you felt extremely proud or extremely disappointed in yourself (choose ONE). What did you learn from this experience that has changed your life? Be sure to support your ideas with specific explanations and details.

3. Most people agree that exercise, nutrition, and medicine are all important for the human body. Which of these, exercise, nutrition, or medicine, do you believe is the most important for people's health? Why? Be sure to support your ideas with specific explanations and details.

4. Parents often control many aspects of their children's lives, even after children reach high school and have some responsibility for themselves. Do you think parents today have too much control over the lives of high school students? Why or why not? Be sure to support your ideas with specific explanations and details.

APPENDIX C

Standard Rubric

	Content		Organization		Vocabulary		Language Use		Mechanics
20 16	Thorough and logical development of thesis Substantive and detailed No irrelevant information Interesting A substantial number of words for amount of time given Well-defined relationship to the prompt	20 16	Excellent overall organization Clear thesis statement Substantive introduction and conclusion Excellent use of transition words Excellent connections between paragraphs Unity within every paragraph	20 16	Very sophisticated vocabulary Excellent choice of words with no errors Excellent range of vocabulary Idiomatic and near native-like vocabulary Academic register	20 16	No major errors in word order or complex structures No errors that interfere with comprehension Only occasional errors in morphology Frequent use of complex sentences Excellent sentence variety	20 16	Appropriate layout with well-defined paragraph separation No spelling errors No punctuation errors No capitalization errors
15 11	Good and logical development of thesis Fairly substantive and detailed Almost no irrelevant information Somewhat interesting An adequate number of words for the amount of time given Clear relationship to the prompt	15 11	Good overall organization Clear thesis statement Good introduction and conclusion Good use of transition words Good connections between paragraphs Unity within most paragraphs	15 11	Somewhat sophisticated vocabulary, usage not always successful Good choice of words with some errors that don't obscure meaning Adequate range of vocabulary but some repetition Approaching academic register	15 11	Occasional errors in word order or complex structures A variety of complex structures, even if not completely successful Almost no errors that interfere with comprehension Some errors in morphology Frequent use of complex sentences Good sentence variety	15 11	Appropriate layout with clear paragraph separation No more than a few spelling errors in less frequent vocabulary No more than a few punctuation errors No more than a few capitalization errors
10 6	Some development of thesis Not much substance or detail Some irrelevant information Somewhat uninteresting Limited number of words for the amount of time given Vague but discernable relationship to the prompt	10 6	Some general coherent organization Minimal thesis statement or main idea Minimal introduction and conclusion Occasional use of transition words Some disjointed connections between paragraphs Some paragraphs may lack unity	10 6	Unsophisticated vocabulary Limited word choice with some errors obscuring meaning Repetitive choice of words Little resemblance to academic register	10 6	Frequent errors in word order or attempts at complex structures Some errors that interfere with comprehension Frequent errors in morphology Minimal use of complex sentences Little sentence variety	10 6	Appropriate layout with somewhat clear paragraph separation Some spelling errors in less frequent and more frequent vocabulary Several punctuation errors Several capitalization errors
5 0	No development of thesis No substance or details Substantial amount of irrelevant information Completely uninteresting Very few words for the amount of time given Relationship to the prompt not readily apparent	5 0	No coherent organization No thesis statement or main idea No introduction and conclusion No use of transition words Disjointed connections between paragraphs Paragraphs lack unity	5 0	Very simple vocabulary Severe errors in word choice that often obscure meaning No variety in word choice No resemblance to academic register	5 0	Serious errors in word order or complex structures Frequent errors that interfere with comprehension Many errors in morphology Almost no attempt at complex sentences No sentence variety	5 0	No attempt to arrange essay into paragraphs Several spelling errors even in frequent vocabulary Many punctuation errors Many capitalization errors

Figure 18. Standard Rubric

APPENDIX D

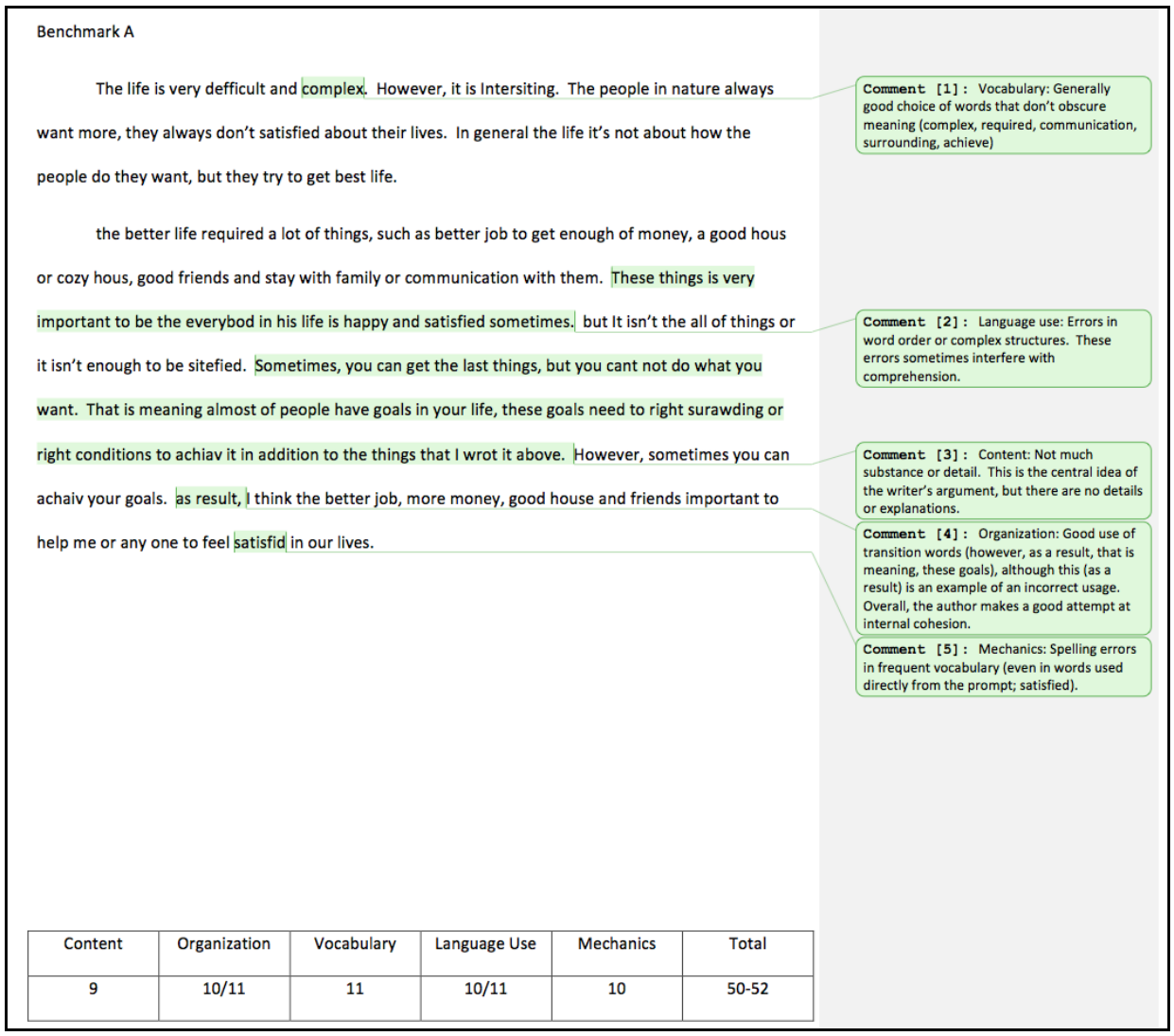
Reordered Rubric

	Mechanics		Language Use		Vocabulary		Organization		Content
20 16	Appropriate layout with well-defined paragraph separation No spelling errors No punctuation errors No capitalization errors	20 16	No major errors in word order or complex structures No errors that interfere with comprehension Only occasional errors in morphology Frequent use of complex sentences Excellent sentence variety	20 16	Very sophisticated vocabulary Excellent choice of words with no errors Excellent range of vocabulary Idiomatic and near native-like vocabulary Academic register	20 16	Excellent overall organization Clear thesis statement Substantive introduction and conclusion Excellent use of transition words Excellent connections between paragraphs Unity within every paragraph	20 16	Thorough and logical development of thesis Substantive and detailed No irrelevant information Interesting A substantial number of words for amount of time given Well-defined relationship to the prompt
15 11	Appropriate layout with clear paragraph separation No more than a few spelling errors in less frequent vocabulary No more than a few punctuation errors No more than a few capitalization errors	15 11	Occasional errors in word order or complex structures A variety of complex structures, even if not completely successful Almost no errors that interfere with comprehension Some errors in morphology Frequent use of complex sentences Good sentence variety	15 11	Somewhat sophisticated vocabulary, usage not always successful Good choice of words with some errors that don't obscure meaning Adequate range of vocabulary but some repetition Approaching academic register	15 11	Good overall organization Clear thesis statement Good introduction and conclusion Good use of transition words Good connections between paragraphs Unity within most paragraphs	15 11	Good and logical development of thesis Fairly substantive and detailed Almost no irrelevant information Somewhat interesting An adequate number of words for the amount of time given Clear relationship to the prompt
10 6	Appropriate layout with somewhat clear paragraph separation Some spelling errors in less frequent and more frequent vocabulary Several punctuation errors Several capitalization errors	10 6	Frequent errors in word order or attempts at complex structures Some errors that interfere with comprehension Frequent errors in morphology Minimal use of complex sentences Little sentence variety	10 6	Unsophisticated vocabulary Limited word choice with some errors obscuring meaning Repetitive choice of words Little resemblance to academic register	10 6	Some general coherent organization Minimal thesis statement or main idea Minimal introduction and conclusion Occasional use of transition words Some disjointed connections between paragraphs Some paragraphs may lack unity	10 6	Some development of thesis Not much substance or detail Some irrelevant information Somewhat uninteresting Limited number of words for the amount of time given Vague but discernable relationship to the prompt
5 0	No attempt to arrange essay into paragraphs Several spelling errors even in frequent vocabulary Many punctuation errors Many capitalization errors	5 0	Serious errors in word order or complex structures Frequent errors that interfere with comprehension Many errors in morphology Almost no attempt at complex sentences No sentence variety	5 0	Very simple vocabulary Severe errors in word choice that often obscure meaning No variety in word choice No resemblance to academic register	5 0	No coherent organization No thesis statement or main idea No introduction and conclusion No use of transition words Disjointed connections between paragraphs Paragraphs lack unity	5 0	No development of thesis No substance or details Substantial amount of irrelevant information Completely uninteresting Very few words for the amount of time given Relationship to the prompt not readily apparent

Figure 19. Reordered Rubric

APPENDIX E

Example Annotated Benchmark Essay



Benchmark A

The life is very defficult and complex. However, it is Intersiting. The people in nature always want more, they always don't satisfied about their lives. In general the life it's not about how the people do they want, but they try to get best life.

the better life required a lot of things, such as better job to get enough of money, a good hous or cozy hous, good friends and stay with family or communication with them. These things is very important to be the everybod in his life is happy and satisfied sometimes. but It isn't the all of things or it isn't enough to be sitefied. Sometimes, you can get the last things, but you cant not do what you want. That is meaning almost of people have goals in your life, these goals need to right surawding or right conditions to achiav it in addition to the things that I wrot it above. However, sometimes you can achaiv your goals. as result, I think the better job, more money, good house and friends important to help me or any one to feel satisfid in our lives.

Comment [1]: Vocabulary: Generally good choice of words that don't obscure meaning (complex, required, communication, surrounding, achieve)

Comment [2]: Language use: Errors in word order or complex structures. These errors sometimes interfere with comprehension.

Comment [3]: Content: Not much substance or detail. This is the central idea of the writer's argument, but there are no details or explanations.

Comment [4]: Organization: Good use of transition words (however, as a result, that is meaning, these goals), although this (as a result) is an example of an incorrect usage. Overall, the author makes a good attempt at internal cohesion.

Comment [5]: Mechanics: Spelling errors in frequent vocabulary (even in words used directly from the prompt; satisfied).

Content	Organization	Vocabulary	Language Use	Mechanics	Total
9	10/11	11	10/11	10	50-52

Figure 20. Example annotated benchmark essay

APPENDIX F

Example Critieria Importance Survey Excerpt

In general, what weight would you attach to each of the following criteria when evaluating writing samples? (Mark one per descriptor on the scale below).

	1 Unimportant	2	3	4 Somewhat important	5	6	7 Important	8	9	10 Very important
Clear connection between paragraphs	○	○	○	○	○	○	○	○	○	○
Appropriate punctuation	○	○	○	○	○	○	○	○	○	○
Appropriate use of academic register	○	○	○	○	○	○	○	○	○	○
Appropriate length	○	○	○	○	○	○	○	○	○	○
Sophisticated vocabulary	○	○	○	○	○	○	○	○	○	○
Detailed content	○	○	○	○	○	○	○	○	○	○
Proper spelling	○	○	○	○	○	○	○	○	○	○
Appropriate formatting	○	○	○	○	○	○	○	○	○	○

Figure 21. Example Critieria Importance Survey excerpt

APPENDIX G

Criteria Recall Task Sheet

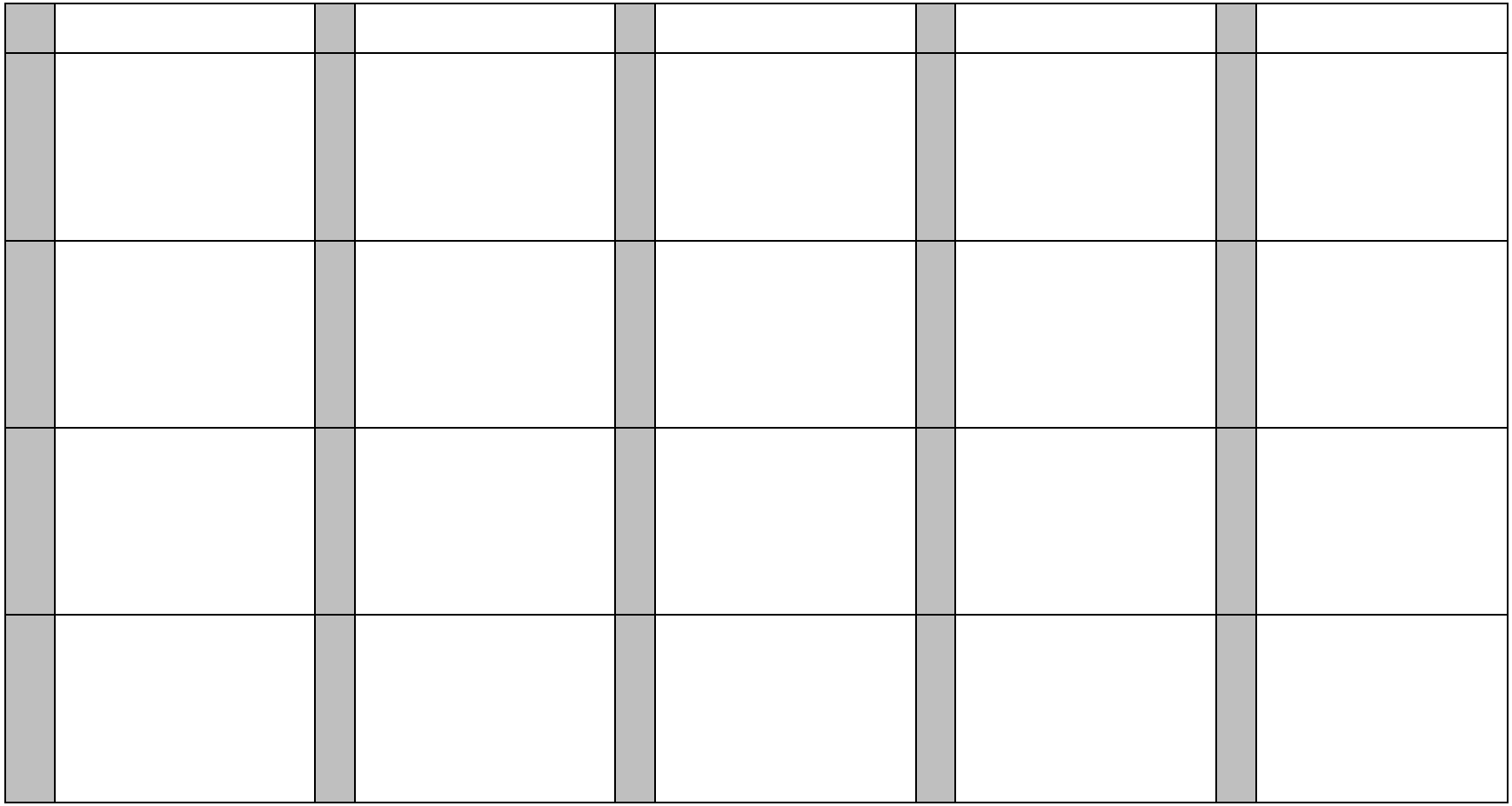


Figure 22. Criteria Recall Task sheet

APPENDIX H

Criteria Recall Task Coding Scheme

Overview

I want to know whether a participant was able to remember each criterion/descriptor/trait from the rubric. Each descriptor is part of a larger trait (e.g., academic register) that is represented in each vertical box within a category. The degree of that trait changes vertically, but it should be represented in each row (e.g., no resemblance to academic register, little resemblance to academic register, approaching academic register, academic register).

Trait Coding

On the reference rubric, each descriptor has a letter code (e.g., A-Z).
You will code each CRT rubric on the presence of each trait as follows:

0	The trait is not represented anywhere on the rubric
1	The trait is represented on the rubric, but it is in the wrong category
2	The trait is represented on the rubric, and it is in the correct category.

Please mark a 0/1/2 for each trait in the appropriate cell on the Excel sheet.

Title Coding

You will code each category title given as follows:

0	No title given for category
1	Alternative title given, but not exact title
2	Exact title given

Please mark a 0/1/2 for each trait in the appropriate cell on the Excel sheet.

Order of Appearance Coding

You will also code the order of appearance of each category, as outlined below:

1	Appears in column one
2	Appears in column two
3	Appears on column three
4	Appears in column four
5	Appears in column five
0	Does not appear on the rubric

Please mark 0-5 for each category in the appropriate cell in the Excel sheet.

APPENDIX I

Semi-structured-interview Questions

1. Tell me about your rating process that you outlined.
2. How do you typically arrive at a score?
3. What do you consider when thinking about a final score?
4. When is it easiest or most difficult for you to arrive at a score?
5. Describe how you typically use the rubric.
6. In your own words, how would you summarize or describe each **category** of the rubric?
7. In your own words, how would you describe/summarize essays in the 0-5, 6-10, 11-15, and 16-20 **bands**?
8. Describe whether and how your reliance on the rubric changed over the course of rating. (Tell me about changes in your rubric use over the course of rating).
9. Tell me about your experience rating this round as compared to last time.

APPENDIX J

Rater Background Questionnaire

1. Participant #:
2. Participant pseudonym:
3. Gender: Male Female
4. Age:

Language Background

5. First language (mother tongue):
6. Foreign language 1:
7. Please self-rate your proficiency in your (most dominant) second language:

Reading	1 (novice)	2	3	4	5	6 (near-native)
Listening	1 (novice)	2	3	4	5	6 (near-native)
Speaking	1 (novice)	2	3	4	5	6 (near-native)
Writing	1 (novice)	2	3	4	5	6 (near-native)

8. Foreign language 2 (if any):
9. Foreign language 2 (if any):
10. Are you bilingual (grew up speaking two languages)?
    Yes
    No
11. If so, what languages did you grow up speaking?
    Language 1:
    Language 2:

Educational Background

12. Please list your educational background information (degree/major/graduation year):
    Bachelor's:
    Master's:
    PhD:

Teaching Experience

13. Do you have any formal teaching experience?
    Yes
    No
14. How long have you been teaching?
    Years:
    Months:
15. What classes/subjects have your taught?
    Class/subject 1:
    Class/subject 2:
    Class/subject 3:
    Class/subject 4:
    Class/subject 5:

16. What is your main student population (e.g., high school students, college-level ESL students, etc.)?

Tutoring Experience

17. Do you have any experience as a tutor?
    Yes
    No
18. How long have you been tutoring?
    Years:
    Months:
19. What areas/subjects have your taught?
    Area/subject 1:
    Area/subject 2:
    Area/subject 3:
    Area/subject 4:
    Area/subject 5:
20. What is your main student population (e.g., high school students, college-level ESL students, etc.)?

Consulting Experience

21. Do you have any experience working as a writing consultant (e.g., in a writing lab)?
    Yes
    No
22. How long have you been a writing consultant?
    Years:
    Months:
23. What are your main areas of focus as a consultant?
    Area/subject 1:
    Area/subject 2:
    Area/subject 3:
    Area/subject 4:
    Area/subject 5:
24. What is your main student/client population (e.g., high school students, college-level ESL students, etc.)?

Rater Experience

25. Do you have any previous experience rating/scoring student writing?
    Yes
    No
26. How long have you been a rater?
    Years:
    Months:

27. In what context have you been rating students' writing (e.g., class assignments, departmental exams, etc.)?
    Context 1:

Context 2:
Context 3:
Context 4:

28. What is your main student/writer population (e.g. high school students, college-level ESL students, etc.)?

Rubric and Rater Proficiency

29. To what extent are you comfortable with using the essay rubric in this study?
1 (very uncomfortable) 2 3 4 5 6 (very comfortable)

30. How proficient/capable do you feel as an essay rater?
1 (novice rater) 2 3 4 5 6 (expert rater)

REFERENCES

REFERENCES

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

Anderson, N. H. (1965). Primacy effects in personality impression formation using a generalized order effect paradigm. *Journal of Personality and Social Psychology*, *2*, 1-9.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Level of processing theory and social facilitation theory perspectives. *Journal of Applied Psychology*, 72, 239–244.

Asch, S. E. (1946). Forming Impressions of Personality. *The Journal of Abnormal and Social Psychology*, *41*(3), 258–290. http://doi.org/10.1027/1864-9335/a000179

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? System, 29 (4), 371–383.

Baker, B. A. (2012). Individual Differences in Rater Decision-Making Style: An Exploratory Mixed-Methods Study. *Language Assessment Quarterly*, *9*(3), 225–248. http://doi.org/10.1080/15434303.2011.637262

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. http://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). Effects of scoring method and rater experience on ESL essay rating processes and outcomes (Unpublished doctoral thesis). University of Toronto, Canada

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(April 2015), 54–74. http://doi.org/10.1080/15434300903464418

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, *28*(1), 51–75. http://doi.org/10.1177/0265532210376379

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9. http://doi.org/10.1111/j.1745-3992.2012.00238.x

Bond, T. G., & Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. Psychology Press.

Brown, J. D. (2005). Testing in language programs: A comprehensive guide to English language assessment (2nd ed.). New York: McGraw-Hill.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. Language Learning, 34 (4), 21–42.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105–121. http://doi.org/10.1016/j.asw.2004.07.001

Burrows, C. (1994). Testing, testing, 1, 2, 3: An investigation of the reliability of the assessment guidleine for the Certificate of Spoken and Written English. *Making Connections*, 1994 ACTA-WATESOL National Conference, 11-17.

Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics*, 11 (2), 207–241.

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

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing : Multiple analyses of a common corpus. *Journal of Second Language Writing*, *26*, 1–9. http://doi.org/10.1016/j.jslw.2014.09.002

Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. *The Journal of Social Psychology*, *101*(1), 87–96.

Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing*, *26*, 1–4. http://doi.org/10.1016/j.asw.2015.08.002

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 31–51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision Making while Rating Tasks : A ESL / EFL Writing Framework Descriptive. *The Modern Language Journal*, *86*(1), 67–96.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*. http://doi.org/10.1177/0265532215582282

Eckes, T. (2008). *Rater types in writing performance assessments: A classification approach to rater variability*. *Language Testing* (Vol. 25). http://doi.org/10.1177/0265532207086780

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270–292. http://doi.org/10.1080/15434303.2011.649381

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37–64. http://doi.org/10.1177/0265532207071511

Field, A. (2009). *Discovering statistics using SPSS*. Sage publications.

Follman, J. & Anderson, J. (1967). An investigation of the reliability of five procedures for grading English themes. *Research in the Teaching of English*, 1, 190-200.

Forgas, J. P. (2011). Can negative affect eliminate the power of first impressions? Affective influences on primacy and recency effects in impression formation. *Journal of Experimental Social Psychology*, *47*(2), 425–429. http://doi.org/10.1016/j.jesp.2010.11.005

Godfroid, Al., & Spino, L. A. (2015). Reconceptializing Reactivity of Think-Alouds and Eye Tracking: Absence of Evidence Is Not Evidence of Absence. *Language Learning*.

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, 41 (3), 258–269.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education*, 27, 73–82.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In: L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, 29, 759–762.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41 (3), 337–373.

Henderson, J. M., Luke, S. G., Schmidt, J., & Richards, J. E. (2013). Co-registration of eye movements and event-related potentials in connected-text paragraph reading. *Frontiers in Systems Neuroscience*, *7*, 28. http://doi.org/10.3389/fnsys.2013.00028

Hendrick, C., & Costantini, A. F. (1970). Effects of varying trait inconsistency and response requirements on the primacy effect in impression formation. *Journal of Personality and Social Psychology*, *15*(2), 158–164. http://doi.org/10.1037/h0029203

Jacobs, H., Zingraf, S., Wormuth, D., Hartfiel, V. & Hughey, J. (1981). *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*(18), 1–16. http://doi.org/10.1016/j.asw.2015.07.002

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485–505. http://doi.org/10.1177/0265532209340186

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. http://doi.org/10.1016/j.edurev.2007.05.002

Kang, O. (2012). Impact of Rater Characteristics and Prosodic Features of Speaker Accentedness on Ratings of International Teaching Assistants' Oral Performance. *Language Assessment Quarterly*, *9*(3), 249–269. http://doi.org/10.1080/15434303.2011.642631

Kline, R. (2004). Beyond significance testing: Reforming data analysis methods in behavioral research. Washington, DC: American Psychological Association.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304. http://doi.org/10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81–96. http://doi.org/10.1016/j.asw.2011.02.003

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing*, *12*(1), 26–43. http://doi.org/10.1016/j.asw.2007.04.001

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3–31. http://doi.org/10.1191/0265532202lt218oa

Larson-Hall, J. (2010). *A Guide to Doing Statistics in Second Language Research Using SPSS* . New York : Routledge.

Li, H., & He, L. (2015). A Comparison of EFL Raters' Essay-Rating Processes Across Two Types of Rating Scales. *Language Assessment Quarterly*, *12*(2), 178–212. http://doi.org/10.1080/15434303.2015.1011738

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. http://doi.org/10.1177/0265532211406422

Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago, IL: MESA Press.

Luchins, A., & Luchins, E. (1970). The effects of order of presentation of information and explanatory models. *The Journal of Social Psychology*, *80*, 63–70. http://doi.org/10.1007/s13398-014-0173-7.2

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing*, *19*(3), 246–276. http://doi.org/10.1191/0265532202lt230oa

Lumley T (2005) *Assessing second language writing: The rater's perspective*. New York: Peter Lang.

Lumley, T., & McNamara, T. (1995). Rater Characteristics and Rater Bias: Implications for Training, 54–71.

Marzano, R. J. (2002). A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, *15*(3), 249–268. http://doi.org/10.1207/S15324818AME1503

McCray, G., & Brunfaut, T. (2016). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing*, 0265532216677105.

McDermott, W. L. (1986). The scalability of degrees of foreign accent. Ph.D. dissertation. Cornell University, Ithaca, NY.

McNamara, T. F. (1996). *Measuring Second Language Performance.* London and New York: Longman.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, 48(1), 73–97.

Myford, C. M. (2012). Rater cognition research: Some possible directions for the future. *Educational Measurement: Issues and Practice*, *31*(3), 48–49. http://doi.org/10.1111/j.1745-3992.2012.00243.x

Orr, M. (2002). The FCE Speaking test: using rater reports to help interpret test scores. *System*, *30*(2), 143–154. http://doi.org/10.1016/S0346-251X(02)00002-7

Polio, C. (2013). *Revising a writing rubric based on raters' comments.* Paper presented at the Midwestern Association of Language Testers (MwALT) conference, East Lansing, Michigan.

Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. Cognitive Psychology, 52, 1–56.

Rebitschek, F. G., Krems, J. F., & Jahn, G. (2015). Memory activation of multiple hypotheses in sequential diagnostic reasoning. *Journal of Cognitive Psychology*, *5911*(October), 1–17. http://doi.org/10.1080/20445911.2015.1026825

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and*

Organizational Psychology, *85*(2), 370–395. http://doi.org/10.1111/j.2044-8325.2011.02045.x

Sakyi, A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate ESL compositions. In: A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*(3), 303–325. http://doi.org/10.1191/026553201680188988

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). The Effect of Raters ' Background and Training on the Reliability of Direct Writing Tests. *Modern Language Journal*, *76*(1), 27–33.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In: G. Brindley (Ed.), *Studies in immigrant English language assessment* (pp. 159–189). Sydney, Australia: National Centre for English Language Teaching and Research, Macquarie University.

Smith, M. (2016). *Testing the shallow structure hypothesis in L2 Japanese* (Doctoral dissertation), Michigan State University.

Solano-Flores, G., & Li, M. (2006). The Use of Generalizability (G) Theory in the Testing of Linguistic Minorities. *Educational Measurement: Issues and Practice*, *25*(1), 13–22. http://doi.org/10.1111/j.1745-3992.2006.00048.x

Stuhlmann, J., Daniel, C., Dellinger, A., Kenton, R., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Reading Psychology*, *20*(2), 107-127.

Tulving, E. (2008). On the law of primacy. In *Memory and mind : a festschrift for Gordon H. Bower* (pp. 31–48).

Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In: A. Cumming & R. Berwick (Eds.), Validation in language testing (pp. 39–57). Clevedon: Multilingual Matters.

Underwood, G. (1975). Perceptual distinctiveness and proactive interference in the primacy effect. *The Quarterly Journal of Experimental Psychology*, *27*(September), 289–294. http://doi.org/10.1080/14640747508400487

Vaughan, C. (1991). Holistic assessment: What goes on in the raters'minds? In L. Hamp-Lyons (Ed.), *Assessing Second Language Writing in Academic Contexts* (pp. 111-126). Norwood, NJ: Ablex.

Weigle, S. (1994). Effect of training on taters of ESL compositions. *Language Testing*, *11*, 197–223.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287. http://doi.org/10.1191/026553298670883954

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Qualitative and qualitative approaches. *Assessing Writing*, 6, 145–178.

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Weigle, S. C., Boldt, H., & Valsecchi, M. I. (2003). Effects of Task and Rater Background on the Evaluation of ESL Student Writing : *TESOL Quarterly*, *37*(2), 345–354.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York, NY: Palgrave MacMillan.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. http://doi.org/10.1177/0265532212456968

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 37–53. http://doi.org/10.1016/j.asw.2015.05.002

Wiseman, C. (2005). A validation study comparing an analytic scoring rubric and a holistic scoring rubric in the assessment of L2 writing samples. Unpublished paper, Teachers College, Columbia University, NY. Cynthia

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150–173. http://doi.org/10.1016/j.asw.2011.12.001

Woehr, D. J. (1994). Understanding frame of reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology*, 79, 525–534.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment*, *10*(1).

Yorozuya, R., & Oller, J. W. (1980). Oral proficiency scales: Construct validity and the halo effect. *Language Learning*, *30*(1), 135-153.