Georgia State University

ScholarWorks @ Georgia State University

Applied Linguistics and English as a Second Language Dissertations

Department of Applied Linguistics and English as a Second Language

8-8-2017

Input-rich Writing Tasks and Student Writing on an English Language Proficiency Test

Megan J. Montee

Follow this and additional works at: https://scholarworks.gsu.edu/alesl_diss

Recommended Citation

Montee, Megan J., "Input-rich Writing Tasks and Student Writing on an English Language Proficiency Test." Dissertation, Georgia State University, 2017.

doi: https://doi.org/10.57709/10244112

This Dissertation is brought to you for free and open access by the Department of Applied Linguistics and English as a Second Language at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Applied Linguistics and English as a Second Language Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.

INPUT-RICH TASKS AND STUDENT WRITING ON AN ENGLISH LANGUAGE PROFICIENCY TEST

by

MEGAN JEAN MONTEE

Under the Direction of Sara Cushing, PhD

ABSTRACT

This project explores the assessment of academic writing for U.S. students learning English as a second language. Through the analysis of 1200 student responses to the writing component of a large-scale standardized test of academic English language development, the study explores how students in grades 3, 6, and 9 at four different score levels use language from task input in their responses. Drawing from research literature about integrated tasks and source-based writing (Shi, 2004; Weigle & Parker, 2012), the study adapts methodologies for analyzing student responses and applies these to a K-12 assessment context. Assessment tasks in the study are described as input-rich tasks and present students with text and graphic prompts in order to elicit responses that reflect academic language proficiency. Results suggest that while a large portion of language in student responses comes directly from the task input, extensive borrowing of longer strings of text is relatively rare across grade and score levels. Clear patterns of language use differentiate students by score level.

INDEX WORDS: Language assessment, Writing assessment, Writing tasks, Task input, Academic language, English as a second language

INPUT-RICH TASKS AND STUDENT WRITING ON AN ENGLISH LANGUAGE PROFICIENCY TEST

by

MEGAN JEAN MONTEE

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2017

Copyright by
Megan Montee
2017

INPUT-RICH TASKS AND STUDENT WRITING ON AN ENGLISH LANGUAGE PROFICIENCY TEST

by

MEGAN JEAN MONTEE

Committee Chair: Sara Cushing

Committee: Eric Friginal

YouJin Kim

Nadia Behizadeh

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2017

DEDICATION

For my family. Somewhere in the middle, the whole world fell apart. We survived.

ACKNOWLEDGEMENTS

I am grateful to the many strong women who have been a part of my personal and professional growth. Thanks, mom, for encouraging me to pursue work that I love and for sacrificing to make it possible. This has been one of your very best gifts to me.

Meg Malone, getting a job as your intern changed my life, and I am grateful to call you my mentor and dear friend. Thank you especially for introducing me to Sara Cushing. Sara, you have been an invaluable adviser, teacher, and role model. And very patient! As I look ahead to my post-graduate school career, I hope that I bring the same engagement, rigor, compassion, and balance to my life and work that I have seen modeled in Meg and Sara over the years.

A dissertation is a big endeavor, and I have had a lot of help over the years. I won the lottery with my PhD cohort. Jack Hardy and Audrey Roberson, I wouldn't have made it without your friendship. Thanks for being my people. I am also grateful for the wonderful students, faculty, and staff who are part of Georgia State's applied linguistics community. I wanted to attend GSU because I was impressed with both the caliber of scholarship and the kindness of the people I met. I haven't been disappointed. Thank you for three wonderful years as a full-time student, and for working with me to complete my dissertation while working. I am grateful to the faculty for their support and to my committee in particular.

Thank you also to my family, who cheer me on, and to dear friends in DC, Atlanta, Oklahoma, and beyond. You have cared for me and picked up my spirits on many occasions. My brother Alex taught be about resilience. Vicky Nier has been my coach and best cheerleader. Margo and Kyle Stedman cared for me so well along the way.

My colleagues at the WIDA Consortium have been incredibly helpful in providing access to data for this study. I am particularly grateful to Elizabeth Cranley for her assistance and

support. I am so glad I was able to design a study that bridged my academic and professional interests. Working on the WIDA project has been a wonderful part of my professional life, and it was a pleasure to extend this work as part of my dissertation.

Finally, I am grateful to my colleagues at the Center for Applied Linguistics (CAL). I couldn’t have a more supportive boss or better role model for work-life balance than Jennifer Norton. Anne Donovan and Jing Wei have been wonderful sounding boards, reviewers, and friends. And working with Dorry Kenyon has made me a better scholar, writer, and thinker. Working full time and finishing a dissertation isn’t easy, but working at CAL has always given purpose to my academic life. I am passionate about our mission, and am thankful for colleagues who made it possible for me to finish graduate school while doing work that I love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

In the U.S. K-12 context, English Language Learners (ELLs) are assessed annually in English language proficiency as part of annual accountability testing mandated by the federal government. These high-stakes tests are used to monitor evaluate schools and districts and to make decisions about language services for individual students. Scores are used as part of criteria for exiting out of English language support services. In order to ensure that students receive the language support that they need in order to access academic content, the assessments must assess academic English, or the language used in school for academic purposes (Bailey, 2006).

The purpose of this study is to examine how ELLs in grades 3, 6, and 9 at four different score levels use language from the task input when responding to writing test tasks on a test of academic English language proficiency. Through a linguistic analysis of 1200 student task responses, this exploratory study seeks to contribute to a better understanding of how students demonstrate their developing academic writing abilities on a large-scale English language proficiency test. The approach used in the study draws from research about integrated test tasks and source-based writing (Shi, 2004; Weigle & Parker, 2012) in order to analyze responses to input-rich writing tasks, or tasks that present students with extended graphic and text input.

1.1 Research motivation

One challenge of creating large-scale standardized test tasks of academic language is designing test tasks that assess language development rather than content knowledge. This means that the tasks must measure how well students can use English for academic purposes apart from their knowledge of specific academic content. This challenge is particularly revalent for performance-based tasks in the language domains of speaking and writing as students need to demonstrate the ability to speak or write about academic topics. One approach to this challenge has been to design tasks that provide students with rich graphic and linguistic input. These tasks, referred to in this study as input-rich tasks, provide all students with common information and content and thus mitigate construct-irrelevant variance due to topical knowledge or previous experience. The goal of input-rich tasks is to measure academic language proficiency as a distinct construct from academic content knowledge.

While input-rich tasks address the need to measure language development as separate from content knowledge, the task design means that students are necessarily dependent on much of the language provided in the input. Students may use this language in different ways. No published research to date has explored to what extent and in what ways students use language from task input in their responses and how this language relates to score levels. Understanding this is important to task design and to a clear conceptualization of the task construct.

1.2 Overview of research design

In input-rich tasks, linguistic input provides students with vocabulary and language structures that they can potentially use in their own writing. The purpose of this study is to explore how students use language from task input in their own responses. Data for the study comes from responses to operational WIDA ACCESS for ELL writing test tasks. WIDA is the name of a consortium of 39 U.S. State Education Agencies who agree to share a common assessment system. WIDA ACCESS for ELLs was a large-scale, paper-based test of academic English language proficiency used in the U.S. K-12 context until the 2015-16 school year, when the paper-based test was replaced with a computer-based assessment. Although the administration mode has changed, the task types used on the paper-based and computer-based versions are similar. The ACCESS assessment is used for purposes of federal reporting and accountability requirements and as part of determining when ELLs are ready to exit from English language support services.

Data for the study includes 400 responses from students at grade 3, grade 6, and grade 9 for a total of 1200 responses. Students at each grade level responded to a separate writing test task targeting language of math and science content areas, and the tasks were scored according to the Writing Rubric of the WIDA Consortium (n.d.) as part of operational testing.

Through a linguistic analysis of these responses, the study addresses research questions related to how test takers use language from the task input in their responses. The analysis includes three phases of coding. In phase 1, word-level coding, all words in responses that came from the task input were identified. In phase 2, phrase-level coding, I identified multiword strings of exactly copied and minimally revised text in the responses. This phase also includes an analysis of how frequently different words from the task input appear at each score level. The final phase of coding includes thematic coding of a subset of responses at each grade level.

1.3 Significance

This study explores how learners at different grade levels and stages of English language development copy and adapt material from writing test prompts to their own writing. Although there is a growing body of research related to this topic in postsecondary writing, particularly related to integrated reading and writing test tasks, this topic has not been addressed in the area of K-12 ESL assessment.

As an employee of the Center for Applied Linguistics (CAL), I work as a researcher and test developer for the WIDA ACCESS test. The study was conducted as a separate project from my professional responsibilities. All results and discussion reflect my own perspective and do not represent the views of CAL or of the WIDA Consortium. However, one goal of this project is to provide a research basis for some of the questions I have encountered in my work, and to produce results that will inform future test development work.

Thus, I expect that this study will have direct implications for the development of writing tasks. It is also relevant to language testing researchers interested in task design and the assessment of academic language development. While exploratory, the study also contributes to a positive, resource-based understanding of student language development by showing the rich and varied ways students leverage the linguistic resources available to them within a testing context

The results reveal patterns of task input use for different grade levels and proficiency levels that go beyond measures of how much input language students use in their responses and show the varied ways they interact with this language and adapt it for their own purposes. A better understanding of student writing illustrates how the ability utilize linguistic and graphic resources is an important aspect of language development. In turn, this may lead to improved test tasks that make better use of test input, and to clearer scoring guidelines with explicit directions for raters about how to address these issues.

1.4 Organization

This dissertation is organized into five chapters. After the introduction, Chapter 2 reviews relevant background issues and literature, including testing policy for K-12 ELLs and research literature about writing task design and test performance. Chapter 3 provides an overview of the research methods used in the study and includes a detailed explanation of coding procedures. Chapter 4 presents results for the study by research question, and Chapter 5 discusses the significance of the findings with a focus on implications for task design and scoring.

2 LITERATURE REVIEW

The purpose of this chapter is to provide background information related to the study. The chapter is divided into three sections. The first section discusses U.S. K-12 educational policy for ELLs with a focus on testing policies and practices for this population and describes the WIDA Consortium and the ACCESS for ELLs test. Section 2 reviews frameworks for understanding writing assessment and task features. The third section reviews research studies and methodological approaches related to how students use language from task input and sources in their writing.

2.1 K-12 English Language Learners in the U.S.

The context for the research study is K-12 ELLs and English language proficiency (ELP) assessment in the United States. In the 2012-13 school year, there were over 4.8 million ELLs enrolled in U.S. public schools, and these students accounted for 8.9% of total public school enrollment nationally (McHugh & Pompa, 2016). As established by the *Lau v. Nichols* (1974) Supreme Court case, ELLs are entitled to receive both language and content instruction as part of civil rights protections.

While learning English, students who are designated as Limited English Proficiency receive language and content instruction with the goal of exiting them into mainstream academic classes. It can take four to eight years for students to meet exit criteria (Hakuta, 2011), although data on this issue is limited. Research on the academic achievement of ELLs has shown that over time, these students often fall behind their native English-speaking peers in key academic areas (Abedi, Leon, & Mirocha, 2000/2005; Hakuta, 2011). Thus, research about the language and academic development of this population is a critical area to ensure that schools and language programs are adequately serving students and meeting legal requirements for equity. In other words, educational outcomes for ELLs are not only an education issue, but also a civil rights issue (Hakuta, 2011).

Educational and assessment policy for ELLs is established at the federal level by the *Elementary and Secondary Education Act* (ESEA). In 2015, the *Every Student Succeeds Act* (ESSA) replaced *No Child Left Behind* (NCLB, 2001) as the ESEA reauthorization bill. ESSA represents both important continuities as well as some major shifts for educational policy in general and for ELLs in particular. While the accountability and testing measures established by NCLB are still required, ESSA transfers a substantial amount of decision-making power from the federal government to the states (Klein, 2016; Wong, 2015). For example, beginning in the 2017-18 school year, states will be responsible for establishing their own accountability plans and will have more freedom to choose the indicators included in their plans, although these are subject to federal regulations and approval ("The Every Student Succeeds Act: Explained," 2015). This means that states will have more freedom to set their desired educational outcomes and to determine how to measure success.

For ELLs, ESSA represents several new developments. The law makes English language proficiency a core academic indicator, which was not the case under NCLB when English language development was reported separately from the results of content testing. This means that data for ELL students now counts in a more high-profile way than in the past. This has the potential to shine a spotlight on ELL issues since their progress in both English and academic subjects is a central component of accountability (McHugh & Pompa, 2016; "The Every Student Succeeds Act: Explained," 2015). However, there are also concerns about how well existing content tests are able to assess ELLs' content knowledge as test performance may be language-dependent (Abedi, 2002).

The future of ESSA regulations and implementation is uncertain (Ujifusa, 2017a, 2017b), as is the extent to which the new regulations will change the status quo when they go into effect during the 2017-18 school year. For now, federal guidance makes it clear that although some reporting requirements for ELLs are changing, “[each] State is still required to report […] the number and target number of English learners making progress and English learners attaining proficiency on the State’s annual English language proficiency assessment” (U.S. Department of Education, 2017, p. 30). For now, large-scale assessment will remain a key part of educational practice for ELLs.

Given the way test scores are used, ELD assessments must measure the how well students are able to use language in order to access academic content (Bailey & Wolf, 2012). The construct of ELD assessments is academic English, or the language of school. Academic language has been defined through a variety of theoretical perspectives and research approaches. Cummins (1980) first distinguished basic interpersonal communication skills (BICS) from cognitive/academic language proficiency (CALP) as a way to differentiate social and academic language. Definitions of academic language have since evolved and now academic language is generally viewed as a particular register of use. Importantly, this approach acknowledges that non-academic language is not less sophisticated or less demanding than academic registers (Bailey, 2006; Bunch, 2006; Schleppegrell, 2004). This study adopts Chamot and O’Malley’s (1994) definition of academic language as, “the language that is used by teachers and students for the purpose of acquiring new knowledge and skills […] imparting new information, describing abstract ideas, and developing students’ conceptual understanding” (p. 40). Bailey (2006) expands this definition with two additional features. First, she argues that academic language requires students to demonstrate knowledge both orally and in writing through conventional academic norms. Second, academic language is also characterized by fewer opportunities to negotiate meaning or to use contextual cues. These additions acknowledge the prominent role of knowledge demonstration through productive language as well as the communication constraints that often characterize academic settings.

For the purposes of large-scale assessment, the construct of academic English must be defined in a way that can be represented in and measured by test items. Standards, including both academic content and English Language development standards, help to define the language students need for school. Although not without controversy, content standards such as the Common Core State Standards for Mathematics and English Language Arts (National Governors Association Center for Best Practices, Council of Chief State School Officers, 2010) and the Next Generation Science Standards (NGSS Lead States, 2013) are widely used across the U.S. and provide a detailed description of academic content benchmarks by grade level. These standards both explicitly and implicitly specify certain language demands, including both receptive and productive language skills and abilities that students must develop in order to meet the standards (Council of Chief State School Officers, 2012). Thus, academic content standards are one important component of defining academic language for K-12 ELLs.

English language development standards also define this construct and form the basis for ELD assessments. As McKay (2000) notes, these standards must reflect the complex relationship between three different components:

1. Language development
2. Language demands of school at a particular level
3. Cognitive development learners from K-12 (p. 195)

Standards must reflect multiple developmental trajectories and the complex ways that these interact as students at different stages of cognitive development learn English and engage with academic content. These standards and the complex construct they represent are the basis for ELD assessment.

2.1.1 WIDA ACCESS for ELLs

ELD assessments align to both language and content standards as a way of operationalizing academic language and ensuring construct representation.

Under federal regulations, states are required to meet assessment and reporting requirements, but are not mandated to use a particular test. Some states develop their own testing programs. For example, in 2016-17, California, New York, and Texas all had state-level ELD assessment programs. In addition to statewide assessments, there are also two multi-state consortia for ELD assessments. In these assessment consortia, multiple states agree to work together to implement a shared assessment program. As of the 2016-17 school year, consortia for ELD assessments included WIDA and English Language Proficiency Assessment for the 21st Century (ELPA21).

According to the WIDA Website ("Mission & the WIDA Story, n.d.), the WIDA Consortium was originally founded in 2002. The name was chose to represent the lead states for the original grant: Wisconsin, Delaware, and Arkansas. However, because of the growth of the consortium beyond these states, the name is no longer used as an acronym. As of the 2016-17 school year, the WIDA Consortium consisted of 39 U.S. State Educational Agencies, making WIDA the largest ELD assessment consortium in the U.S.

In 2011, the WIDA Consortium was awarded a federal grant from the U.S. Department of Education to develop a computerized version of the WIDA ACCESS for ELLs assessment. ELPA21 was established as part of the same grant program. Thus, compared with WIDA, this consortium is relatively new. As of 2016-17, the ELPA21 consortium included seven U.S. states ("About ELPA21: FAQs." n.d.). Both WIDA and ELPA21 implement computer-based ELD assessments and test reading, writing, listening and speaking.

The computer-based WIDA test, known as WIDA ACCESS 2.0, became operational in 2015-16 and is currently used across the WIDA Consortium. WIDA ACCESS 2.0 does include paper-based versions in addition to the computer-administered test. WIDA ACCESS for ELLs, the older paper-based test, is no longer in use. In addition to the use of a shared annual ELD assessment, the WIDA consortium also uses shared English Language Development Standards (Gottlieb, 2004; Gottlieb et al., 2007; WIDA, 2012). The WIDA Standards describe language development in five areas:

- Social Instructional Language (SIL)
- Language of Language Arts (LoLA)
- Language of Math (LoMA)
- Language of Science (LoSS)
- Language of Social Studies (LoSS)

These standards and supporting publications form the basis of the annual assessment as well as instructional resources and professional learning opportunities for educators.

The WIDA approach focuses on a student-centered, "Can Do philosophy" which emphasizes the "assets, contributions, and potential of culturally and linguistically diverse children and youth" ("Mission & the WIDA Story, n.d.). This means that the assessment is designed to allow students at all levels of language to show what they can do in English. Collaboration is also a core value of the consortium. Educators are involved in many stages of test development and review. For example, teachers serve as item writers and panels of educators from the WIDA consortium review items and recommend revisions. The role of teachers in test development helps ensure that the assessment content represents authentic classroom contexts and that it is accessible and relevant to students. The Center for Applied Linguistics is responsible for test development for the WIDA consortium and along with WIDA staff coordinates the involvement of educators throughout the process. It typically takes one year or longer for test items to go from initial item writing to field testing, and performance-based tasks are often piloted with students before field testing to further refine the test content.

According to Behizadeh and Pang's (2016) framework for assessing content-based writing assessment, the ACCESS for ELLs writing test is classified as direct psychometric assessment, or an on-demand essay test. WIDA writing tasks are characterized by academically-based topics and rich linguistic and graphic input. For example, a WIDA writing task may present students with language and graphic input about a science experiment and ask students to write about the likely results. Although the tasks do place language processing demands on students, reading ability is not part of the writing construct measured by WIDA, and these tasks are not characterized as integrating multiple skills. Task input is designed to be clear and accessible to students at the targeted grade levels, and the graphic support is designed to both support linguistic input as well as minimize the need to include extensive text in order to present content. A summary of the test design describes the role of task graphics:

> Graphics are intended to reduce the potentially confounding influence of whatever linguistic channel is used to present the task context by opening a visual channel to frame that context. From another vantage point, the graphics also provide a non-linguistic means of supplying English language learners with necessary background knowledge to compensate for the advantage that students with academic preparation might otherwise have. The net effect of the use of theme graphics, then, is to increase the redundancy of task-specific contextual information. A concomitant effect is, typically, that the student test taker will have multiple pathways to finding or producing a correct or appropriate response. This notion ties importantly to our contention that ACCESS for ELLs® does not test individual skills or mechanical processing abilities, but tests language proficiency in a more comprehensive sense (Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007, p. 84).

This description highlights the overall goal of WIDA writing tasks to assess academic language while minimizing construct-irrelevant variance due to content knowledge, and the central role of task input in this goal.

Published research about the WIDA ACCESS test has been limited. Some of this has focused on evidence supporting the design of the test or research which helps establish guidelines for stakeholders (e.g., Cook, Boals, Wilmes, & Santos, 2008). Kenyon, MacGregor, Li, & Cook (2011) addressed measurement issues in creating a vertical scale for the assessment system. The vertical scaling procedure they describe allows multiple WIDA ACCESS test forms across grade levels to be placed on a single scale. This procedure supports the interpretation and use of test scores as stakeholders track student progress across multiple years.

In a separate study, Römhold, Kenyon, & MacGregor (2011) explored the role of general academic language development and domain- or content-specific knowledge on the test. Using latent factor analysis of WIDA ACCESS test data, they found evidence supporting the conceptual distinction between domain-general and domain-specific academic language. Domain-general knowledge was more prevalent during early stages of proficiency while domain-specific knowledge became more prevalent at high proficiency levels. They concluded that:

> [The results raise] an interesting question regarding the relationship between academic language proficiency and academic content knowledge. Test developers are generally quick to emphasize that tests of academic English proficiency are not tests of academic content knowledge. However, it is difficult to differentiate between the two forms of knowledge when language specific to a certain content domain is assessed. […] Given that domain-specific linguistic knowledge may play an increasingly more important role at mid- and high levels of proficiency, the distinction between academic language proficiency and academic content knowledge could also become increasingly blurred at these levels (p. 225).

Their results point to the practical challenge of developing test tasks that measure academic language proficiency as a distinct construct from content. At the highest levels of language development, these two seemed to be particularly intertwined. Bachman and Palmer (1996) speak to this issue when they describe three separate approaches to addressing the role of content knowledge (or "topical knowledge" in their terminology). Language ability and content knowledge can be defined as distinct constructs and assessment tasks can seek to measure language ability on their own. It is also possible to define the construct measured in an assessment as including both language ability and content knowledge, and creating assessment tasks which assess both together. And finally, it is also possible to define language and content knowledge as distinct constructs and measure these through distinct assessment tasks in order to make claims about both. The WIDA assessment takes the first approach in defining language ability as a distinct construct and seeking to measure this separately. The interpretation of test results makes no claims about student achievement in content areas. However, the results of the latent factor analysis research on the WIDA test suggest that this conceptual distinction may be difficult to sustain at higher proficiency levels as language ability and content knowledge become more intertwined.

2.2 Framework for writing assessment

The previous section focused on assessment policy and practice for ELLs. This section discusses frameworks for conceptualizing writing assessment, surveys approaches to categorizing tasks, and discusses considerations for input-rich tasks like the ones in this study. The purpose of this section is to provide a systematic way to understand the relationship between task features and language production and to review relevant approaches to describing writing tasks.

A writing performance on an assessment is a result of the interaction between a test taker and a task. Figure 1, adapted from Kenyon's (1992) model of performance-based assessment, illustrates this.

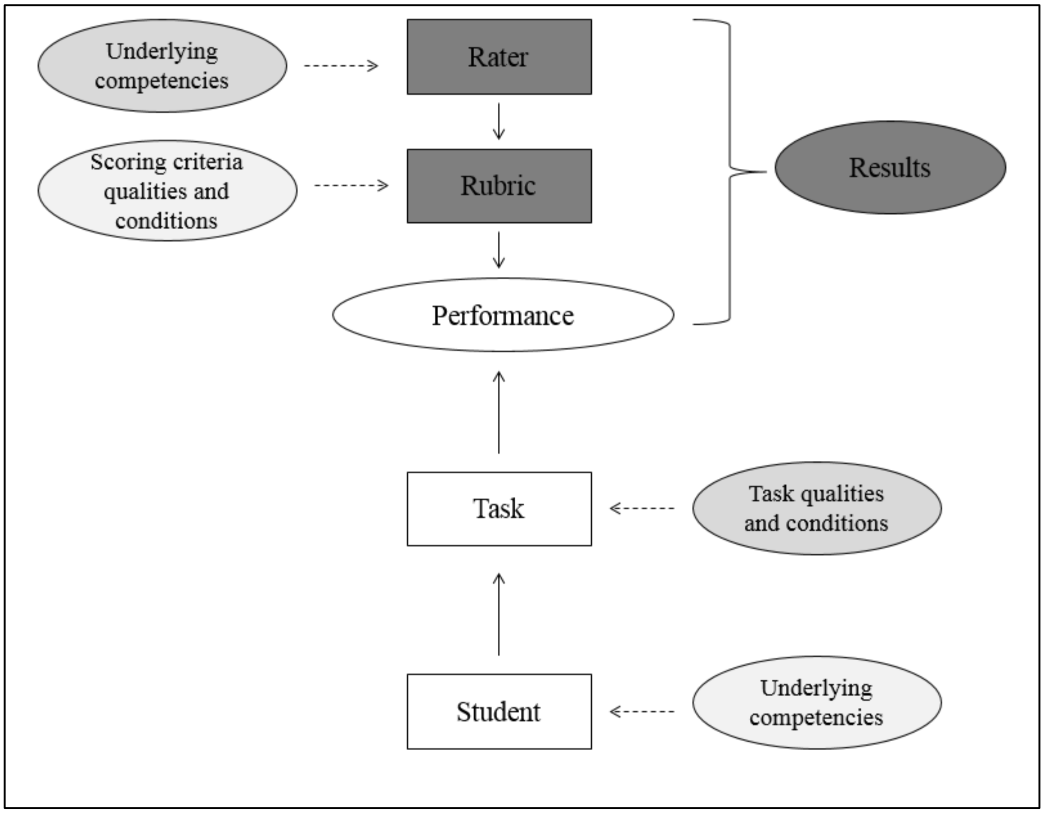


Figure 1. Model of writing assessment (adapted from Kenyon, 1992)

As Figure 1 shows, tasks mediate performances. Test takers bring underlying competencies, and these interact with task qualities and administration conditions to produce a performance. Kenyon's model also provides a framework for understanding how scores are produced. Raters apply rubrics to performances to produce test results. Thus, test scores are a result of a mediating process by which raters' expectations are mediated by rubrics. Rubric variables and scoring conditions inform the application of the rubric to performances.

2.2.1 Characterizing test tasks

In this review I take a construct-based approach to understanding performance assessment, as described by Bachman (2002). This approach views task performance as a reflection of underlying ability. This is contrasted with a task-based approach (e.g., Norris, Brown, Hudson, Yoshioka,1998) in which the task is the fundamental consideration in understanding performance, scoring, and interpretation. In taking a construct-based approach, I focus on conceptualizing task types and task characteristics from this perspective. This means linking task characteristics and performances to an understanding of underlying ability.

Writing tasks have been characterized in a number of ways. A distinction between independent and integrated tasks has been particularly prevalent in recent years, particularly related to research about the TOEFL iBT test. The TOEFL testing program defines independent tasks as writing tasks in in which test takers "formulate and express ideas on their own" (Enright, et al., 2008, p. 129) based on relatively minimal task prompts designed "to stimulate the examinee to generate his/her own ideas on the topic with supporting reasons and examples […]" (Cumming, Kantor, & Powers, 2001, p. 11). A second type of task, labeled "integrated task," is designed to "require comprehension of academic information and sustained written responses based on comprehension of that information" (Enright et al., 2008, p. 129). These tasks types are designed to reflect authentic language demands of university-level academic contexts (Weigle, 2002). Research has sought to establish the authenticity of integrated tasks (Plakans, 2008, 2009; Cumming, 2013), claiming that these tasks reflect the real-world integration of skills that is often required in academic contexts.

Integrated writing tasks typically ask test takers to read and/or listen to a passage and then respond in writing. Integrated tasks may be used to assess writing ability alone (e.g., Plakans & Gebril, 2012) or to assess multiple skills, such as reading comprehension and writing (e.g., Weigle, Yang, & Montee, 2013). The tasks may also vary according to how test takers are required to make use of passages in their responses. For example, test takers may be required to synthesize information in support of an opinion. In other cases, the item passage may provide context for the prompt. Thus, integrated tasks should be characterized by the mode of input (text, audio, video, or graphics), the language characteristics of this input, and by the cognitive task required of students when responding (e.g., synthesis, summary, recount).

While the distinction between independent and integrated tasks is useful, the WIDA writing assessment tasks in this study don't fit within these categories. Another way to categorize writing tasks comes from Kroll and Reid (1994), who describe three types of writing prompts based on the amount of task input. Bare prompts provide minimal input; framed prompts provide more extensive input in order to support the task; and text-based prompts provide extensive input in the form of reading texts. These categories are helpful in characterizing tasks with varying degrees of input. However, this framework is based only on the amount of task input and does not conceptualize a clear relationship between input and student responses or the underlying skills and abilities that relate to each task type. Other approaches to categorizing tasks focus on the rhetorical or functional task that students are asked to do. These categories often include narrative or argumentative tasks (Lim, 2010).

2.2.2 Task variables

The previous section discussed approaches to categorizing task types. This section focuses on the more fine-grained issue of task variables, or how to systematically account for different task features. Bachman (2002) describes the need to characterize task as a set of variables rather than as "holistic entities" (p. 469). This approach allows for research that systematically approaches the relationship between task features, test taker characteristics, and the interaction between them. Bachman and Palmer's (1996) widely-used framework of task characteristics catalogues a variety of task features. Sections of this framework are reproduced in Table 1 related to input characteristics, expected response, and the relationship between the input and the response.

Table 1. Key task characteristics from Bachman and Palmer (1996)

- Characteristics of the input
  - *Format*
    - Channel (aural, visual)
    - Form (language, non-language, both)
    - Language (native, target, both)
    - Length
    - Type (item, prompt)
    - Degree of speededness
    - Vehicle (‘live’, ‘reproduced’, both)
  - *Language of input*
    - Language characteristics
      - Organizational characteristics
        - Grammatical (vocabulary, syntax, phonology, graphology)
        - Textual (cohesion, rhetorical/conversational organization)
    - Pragmatic characteristics
      - Functional (ideational, manipulative, heuristic, imaginative)
      - Sociolinguistic (dialect/variety, register, naturalness, cultural reference and figurative language)
    - Topical characteristics

- Characteristics of the expected response
  - *Format*
    - Channel (aural, visual)
    - Form (language, non-language, both)
    - Language (native, target, both)
    - Length
    - Type (selected, limited production, extended production)
    - Degree of speededness
  - *Language of expected response*
    - Language characteristics
      - Organizational characteristics
        - Grammatical (vocabulary, syntax, phonology, graphology)
        - Textual (cohesion, rhetorical/conversational organization)
      - Pragmatic characteristics
        - Functional (ideational, manipulative, heuristic, imaginative)
        - Sociolinguistic (dialect/variety, register, naturalness, cultural reference and figurative language)
    - Topical characteristics

- Relationship between input and response
  - *Reactivity* (reciprocal, non-reciprocal, adaptive)
  - *Scope of relationship* (broad, narrow)
  - *Directness of relationship* (direct, indirect)

The features listed in Table 1 are helpful in surveying the ways in which task input can vary and provide a starting point for task categorization. In this study, task input comes in both language and graphic (non-language) form presented visually via printed test booklets. The language is highly topical and embedded within subjects related to math and science. The length and language features of the input are described in more detail in Chapter 3.

2.2.3 Summary of task features and variables

This section introduced a framework for understanding task responses and surveyed several approaches to classifying writing tasks and describing task variables. These approaches are applicable to the current study in that they provide a way to describe input-rich tasks, a task type that has not been explicitly described in previous research literature. Kroll and Reid's (1994) conception of "framed" prompts is helpful in identifying a task category between "bare" prompts and integrated tasks. However, it is also important to understand the purpose and characteristics of the input in these tasks. In this case, insights from integrated tasks are helpful in terms of describing variables related to the input mode and cognitive task demands, or what students need to do with the input in order to respond appropriately to the task. Finally, Bachman and Palmer's (1996) framework of task features is useful in cataloging task variables and in conceptualizing the relationship between the task and the response. The methods section returns to these concepts in describing the tasks used in this study.

2.3 Review of research related to textual appropriation

The WIDA writing tasks used in this study have not been widely studied, and there is no body of research literature directly related to the focus of the research. This section focuses on research from the field of language testing related to textual appropriation of task input in student responses, and focuses on research related to integrated writing tasks. After a brief survey of related studies, the section focuses on a description of three methodological approaches to studying how writers use task input in their writing responses.

The vast majority of published academic research on integrated tasks has focused on post-secondary assessment (e.g., Cumming et al., 2005; Plakans, 2008, 2009; Plakans & Gebril, 2012; Weigle & Parker, 2012). Research has explored the role of reading ability on integrated tasks (Esmaeili, 2002; Plakans, 2009; Plakans & Gebril, 2012). As Plakans & Gebril discuss, correlational studies show a weak relationship between reading and writing ability on integrated task while process-based approaches to research have shown that reading ability appears to factor into the writing process and performance.

The growing popularity of integrated assessment tasks, and reading-and-writing tasks in particular, has led to a focus on how test takers use language from sources in their responses (e.g., Weigle & Parker, 2012). This research relates to work in pedagogical contexts as a clear relationship exists between assessment tasks and ways source-based writing is used for academic purposes. Pedagogical research has focused the perception of transgressive practices (Flowerdew & Li, 2007; Hu & Lei, 2012; Pecorari, 2008; Pecorari & Petrić, 2014); disciplinary attitudes towards source-based writing (Shi, 2012; Davis & Morley, 2015; Pecorari & Shaw, 2012); genre-based perspectives on source use (Cheng, 2011); and features of effective source use (Petrić, 2012). Although there is extensive research in pedagogical contexts, Polio & Shi (2012) observe that assessment is a relatively new area for examining textual appropriation, or the borrowing of language from sources.

Some researchers approach textual appropriation through the lens of intertextuality. This is defined as a shared relationship between texts and describes how meaning is created in a text through discourse with other texts (Bakhtin, 1981, 1986; Fairclough, 1992, 2003; Kristeva, 1986;

Lemke, 1992). While intertextuality can be used in a variety of ways to capture how texts relate to each other, this framework often relates to notions of intentionality of the writer and purposeful connections between texts. For example, intertextuality can be a useful framework for understanding citation or plagiarism practices. Because of the task types used in this study, which are different from source-based writing tasks, I found it most useful to frame the project in terms of task characteristics and writing constructs rather than adopt intertextuality as a theoretical framework for the study. Thus, I do not review this literature in detail here. In this study, I focus on task features and adopt a framework for understanding writing based on assessment contexts.

2.3.1 Methodological approaches

This section reviews the coding methodology from three studies of textual appropriation in student writing. Shi (2004) developed a coding methodology for researching textual appropriation in a postsecondary pedagogical context. Shi analyzed source-based summary and opinion task responses from Chinese- and English-speaking university students in China and North America, respectively. Researchers identified strings of borrowed text in student writing. A string was defined as four or more consecutive words borrowed from the source text. Two consecutive content words from the source text and a string of three consecutive words that formed a syntactic constituent (e.g., a prepositional phrase) were also included in the coding. After borrowed strings were identified, these were then coded according to type.

The coding scheme used in this study is organized around whether or not the source of the borrowed string was acknowledged by the writer. Borrowed strings could be identified as having no reference, as having a reference, or as being directly quoted from the source. Text strings with and without reference are further identified according to the degree linguistic of modification. Strings were identified as copied, slightly modified, or syntactically reformulated. The coding scheme is summarized in Table 2.

Table 2. Shi's (2004) coding scheme for identifying textual appropriation

Coding categories
With no reference
-Copied
-Slightly modified
-Syntactically reformulated
With reference
-Copied
-Slightly modified
-Syntactically reformulated
With quotation

“Copied” strings were defined in this study as text taken directly from the source. “Slightly modified” strings were strings with minor modifications such as adding or deleting words or substituting a synonym for a word in a string. “Syntactically reformulated” strings were strings of closely paraphrased text in which the source had been reformulated or the wording had been modified. Total paraphrases were not identified in the coding scheme.

A statistical analysis of the amount of textual borrowing showed that both task type (summary or opinion) and first-language background affected the average amount of borrowed language in the responses. Overall, the summary task elicited a greater amount of borrowing than the opinion task, and Chinese students used more non-referenced language than students whose first language was English.

Weigle and Parker (2012) adapted Shi’s coding methodology in order to analyze 63 responses to a source-based writing task on an English proficiency test used for admission and placement purposes. The responses were stratified across three different score bands. This study used the same procedures for identifying strings of borrowed text and the same coding categories first developed by Shi.

The results of this study found that students borrowed an average of 2.73 strings per essay and that overall, most strings were short (3-4 words) and did not directly reference the text. Their data did include essays with longer non-referenced strings, and this was explained by either extensive borrowing in a small number of students and by the strategy of prompt rewording to begin responses. That is, some students would rephrase the task prompt as a strategy for responding. The researchers found no significant differences between score level for the percentage of borrowed words or for the rate of strings per text, and their data did not suggest a systematic relationship between score level and type of borrowing.

These two studies both used a manual coding scheme for identifying borrowed strings in student texts. Another set of studies (Keck, 2006, 2014) investigated similar issues using an automated approach to identifying borrowed and modified language. Keck created a custom computer program to identify paraphrase in 165 summary essays by L1 and L2 speakers of English in a U.S. university context. This study introduced a new construct of "attempted paraphrase," defined as sections of a summary which could be explicitly linked back to the source and which contained at least one word-level change to the source. Keck identified the following coding categories: Exact Copy, Near Copy, Minimal Revision, Moderate Revision, and Substantial Revision. This taxonomy of paraphrase types was created exclusively on linguistic criteria related to the number of "unique links" contained within a string. A unique link was defined as a word or string in the response that could be directly linked to part of the source and that did not occur elsewhere in the source. This automated approach to identifying paraphrase is systematic and replicable, but does require specialized computer-programming ability.

Keck (2006) found that paraphrase was an important strategy for writers responding to the summary task, and that attempted paraphrase (i.e., modified text) was more frequently used than exact copying by both L1 and L2 writers. A follow-up study (2014) looked at a total of 227 summary texts. This study went beyond the amount of paraphrasing to research how writers select and integrate paraphrase into their own writing. Results included the finding that writers tend to follow the order of ideas presented in the source text in their own writing. Keck also found that some sections of the source texts were most heavily paraphrased by writers, which is an indication of the centrality of these sections to the overall meaning of the source. Overall these studies indicate that the ability to identify and summarize key information in sources is a crucial aspect of summarization in academic writing.

The methodological approaches represented by Shi (2004), Weigle and Parker (2012), and Keck (2006, 2014) demonstrate different approaches to identifying borrowed language in student response writing, and highlight some of the differences between manual and automated approaches. Manual approaches to coding can be more time-intensive but offer researchers an in-depth understanding of data through the coding process. Automated approaches allow for the processing of a large number of texts but require specialized programming knowledge. Additionally, automated approaches consistently apply clear linguistic criteria to the coding process whereas human judgment, and error, may affect manual identification of borrowed language.

2.4 Summary and implications

This section began with an overview of testing policy and practice in the U.S. Because high-stakes assessment is both federally mandated and used to make important decisions about when students exit from language services, assessments must provide accurate data about the extent to which students are acquiring academic English, or the language they need for school. The WIDA ACCESS assessment is widely used for this purpose, and the writing section of the test uses an approach to task design intended to assess academic language as a distinct construct from topical content knowledge.

From an assessment perspective, it is important to understand task types and task characteristics as a way to ensure that assessments are representing the language construct and to support appropriate test score interpretation and use. Bachman and Palmer's (1996) framework of task characteristics is a widely used approach that provides a systematic inventory of ways in which tasks can vary. However, the existing literature in writing assessment does not provide an appropriate categorization for the types of writing tasks found on WIDA ACCESS. While insights from research about integrated tasks can be helpful, WIDA tasks use input for a different purpose. On the ACCESS writing test, rich task input is designed to elicit content-based, discipline-specific writing and to mitigate the role of background knowledge as construct-irrelevant variance. Integrated writing tasks, and reading-and-writing tasks in particular, are typically designed to assess the construct of source-based writing. These are clear differences in the purpose of task input and in the constructs being assessed by each type of task. For the purposes of this study, I adopt the term "input-rich tasks" to refer to the writing tasks used on the WIDA assessment. This term is unique to this study and fills a gap in existing task typologies. It refers to tasks with an extended stimulus which provide test takers with the background information they need to formulate an appropriate response.

In researching this new task type, methodological approaches from source-based writing are useful. Shi (2004) developed a coding scheme based on attribution and level of copying, and Weigle and Parker (2012) adapted this coding methodology for assessment research. Keck (2006; 2014) demonstrated the usefulness of an automated approach to identifying paraphrasing and copied strings of text, although this required the development of a specialized computer program.

The literature review has several implications for the study. First, it highlights the need to study assessment data from K-12 students and to better understand the nature of writing assessment for this population. Most research in language assessment has focused on university contexts, and a focus on younger learners promises useful insights related to how language ability and cognitive development work together. However, this also highlights the exploratory nature of any work in this area. This study is a first step in understanding how students respond to input-rich writing tasks, and how language from the task input relates to student response characteristics. Coding schemes from integrated tasks are a helpful starting point, but the conceptualization of input-rich tasks also necessitates some modifications to these approaches. The design of the research study seeks to provide a foundational understanding of how writers use task input in their writing responses.

3 METHODS

3.1 Methods overview

Given the need to better understand input-rich tasks and how task features relate to test takers responses, this study addresses three research questions:

1. For each grade level, to what extent do students use language from task input in their responses?
    a. Are there differences by score level?
2. For each grade level, what linguistic patterns emerge in terms of how students at different proficiency levels use language from the task input?
    a. To what extent do students at each score levels use different content words from the task input?
    b. To what extent do student at each score level appropriate strings of text from the task input?
    c. What patterns of input language use characterize each score level?
3. Are there different patterns of task input use by grade level?

The study addresses these questions through the analysis of 1200 student writing responses. These responses come from 400 students in each grade level: 3, 6, and 9. The first two research questions explore patterns by score level within each grade, and the final research question compares results across grade levels.

The study included three phases of coding to examine different aspects of how writers used language from the task input in their responses. Because each grade level group responded to a different test task, direct comparisons between grade levels are not possible. This chapter describes the methods used in the study, and includes a description of the test tasks, student responses, transcription and text analysis procedures, and the process for comparing results across groups.

As noted in the Introduction, my work on this project is motivated in part by my professional interest in the WIDA ACCESS for ELLs testing program. Through my position at the Center for Applied Linguistics, I was granted permission to access data for the project and designed the study with the hope that the results could be used to improve task design and scoring tools in addition to furthering the field's understanding of how task features relate to response characteristics more generally. In my position, I am inclined to view the test in a positive light and there necessarily limitations to my perspective. However, I felt that in spite of the challenges involved in an emic approach, my knowledge of the test design and task features would be an asset to the work. The research methodology I adopted was designed with an eye towards the potential usefulness and application of the results for my work while also attempting to link the applications to broader issues within the field of language testing.

3.2 Background

This section includes background information about the design of WIDA ACCESS for ELLs, henceforth WIDA ACCESS, a description of how operational tests were scored, and information about the test taker population.

3.2.1 Writing subtest

Data for the study comes from writing responses to the WIDA ACCESS test. As noted in Chapter 2, in 2015 the paper-based WIDA ACCESS test was replaced by ACCESS 2.0, an updated computer-based version. This study does not address WIDA ACCESS 2.0 and information only applies to the design of the paper-based version. However, because the task design remained relatively stable when the test was updated for the computer, the discussion chapter does review implications for WIDA ACCESS 2.0.

WIDA ACCESS included sections for four language domains: reading, listening, writing and speaking. Students completed a test form based on their grade-level cluster (1-2, 3-5, 6-8, or 9-12) and tier (A, B, or C). WIDA ACCESS tier placement was based on proficiency level and was determined by educators when ordering test forms. The writing subtest included three different writing tasks that were targeted to the student's level. Data for the current study comes from Tier C forms of the Grades 3-5, 6-8, and 9-12 forms of the test. Tier C was completed by students at the higher end of the proficiency scale, including students who may be eligible to exit from language services.

The Tier C test form included a 10-minute task, and 20-minute task, and a 30-minute task. In total, the recommended test administration time was 65 minutes, which included 5 minutes for directions. Table 3 shows the format of the WIDA ACCESS Tier C writing subtest, including information about the recommended administration time and the WIDA standards addressed by each task.

Table 3. Structure of the WIDA ACCESS for ELLs Tier C writing subtest

Task order	Recommended administration time	WIDA standards addressed
1	10 minutes	Social Instructional Language, Language of Language Arts, Language of Social Studies
2	20 minutes	Language of Math and Language of Science
3	30 minutes	Language of Language Arts

The current study uses data from the second of the three test tasks included on the Tier C test form. As noted in Table 3, this task has a recommended administration time of 20 minutes and addresses the Language of Math and Language of Science standards (abbreviated as MS).

Each task is based on one of five WIDA Standards, which means that these standards are reflected in task content and language elicited by task design. For example, MS tasks are based on language used in the content areas of math and science. The study includes student responses to three different MS test tasks from the Grades 3-5, 6-8 and 9-12 test forms. All test tasks were used as part of operational testing during 2010-2012 and have since been retired from use.

3.2.2 Operational scoring

The study uses operational task scores to categorize responses by level. During operational testing, responses to all three test tasks were scored using WIDA's writing rubric. This rubric was replaced by a new scoring scale as part of WIDA ACCESS 2.0 updates to the test design and scoring and is no longer used for operational scoring. The writing rubric is publicly available on the WIDA website ("Writing rubric of the WIDA Consortium," n.d.) as part of the 2007 edition of the WIDA Standards (Gottlieb, Cranley, & Oliver, 2007). Responses were scored by trained professional raters at MetriTech, Inc., the vendor responsible for scoring and score reporting of the WIDA ACCESS test. All responses were rated by a single rater, which was the score of record. A percentage of responses were double-scored to monitor inter-rater reliability, but this data was not used in this study.

The writing rubric was used as a holistic scoring scale with subscores by three rubric categories: Linguistic Complexity, Vocabulary Usage, and Language Control. Although subscores were assigned for each category, these were not conceptualized as analytic scores. Rather, a rater assigned a holistic score (e.g., 4) and then could assign a subscore at an adjacent level to indicate a strength or a weakness in a particular area. At least two of the three subscores must match the holistic level while one subscore may be at an adjacent level. For example, a score of 4-4-5 would present a holistic score of 4 with a strength in Vocabulary Usage. A score of 3-3-5 would not be possible, as the subscore must be at an adjacent level. The rubric descriptors for Linguistic Complexity at level 1 and level 2 contain references to language from the task input stating that, “varying amounts of text may be copied or adapted.” Language production at these two score levels is minimal. The expectation is that at level 1 students will produce, “single words, set phrases or chunks of language” and at level 2 students produce, “phrases and short sentences.” The quantity of language and discourse sophistication increases at each score level so that at level 5 the score descriptor states that students can produce, “a variety of sentence lengths of varying linguistic complexity in a single organized paragraph or in extended text.” The Vocabulary Usage descriptors characterize vocabulary in terms of “high frequency” words and “general language” at lower score levels and as “technical” and “precise” language at the higher score levels. Language Control descriptors focus on the overall comprehensibility of responses. This category also contains references to the task input for levels 1 and 2, noting that responses are level 1 responses are “generally comprehensible when text is adapted from […] source text [but that] comprehensibility may be significantly impeded in original text. Level 2 responses are described as “generally comprehensible when text is adapted from […] source text, or when original text is limited to simple text.” The use of adapted and original language is not addressed at other score levels.

After scoring, task scores were then weighted. Next raw score totals were converted to scale scores, and scale scores converted to writing proficiency levels for score reporting. Score reports include writing subscores and general proficiency levels, and writing subscores contributed to the composite proficiency levels for students across all language domains.

3.2.3 Test taker population

This section describes the population of test takers who completed WIDA ACCESS in 2010-2012. Data for the study comes from operational WIDA ACCESS testing during these years.

Information about the total population of test takers is cited from the publicly available Annual Technical Reports. The total population of test takers for 2010-11 was 824,590 (Yanosky et al., 2012, p. 40) and 975,142 in 2011-12 (Yanosky et al., 2013, p. 40).

Table 4 shows the total test taker population and gender distribution for the three grade levels included in this study.

Table 4. Student population and gender distribution by test year

	2010-2011				2011-2012			
	N	F	M	Missing	*N*	F	M	Missing
Grade 3	92,7224	46.8%	53.0%	0.2%	109,808	46.4%	53.2%	0.4%
Grade 6	46,066	44.8%	55.0%	0.2%	54,054	45.0%	54.6%	0.4%
Grade 9	47,417	44.8%	55.0%	0.2%	54,928	44.2%	55.2%	0.6%

As Table 4 shows, grade 3 had the largest number of test takers in each test year compared with grade 6 and grade 9. The total population of test takers in each grade increased from 2010-2011 to 2011-2012. The gender distribution of test takers was relatively equal between female and male students, although for each grade the percentage of male students was slightly higher.

Table 5 shows the ethnicity of the study population by test year. Available data provides only the number of Hispanic students. All other ethnicities are reported as “other.”

Table 5. Student ethnicity by test year

	2010-2011			2011-2012		
	Hispanic n (%)	Other n (%)	Missing n (%)	Hispanic n (%)	Other n (%)	Missing n (%)
3	61,854 (66.7%)	20,988 (22.6%)	9,880 (10.7%)	75,896 (69.1%)	30,354 (27.6%)	3,558 (3.2%)
6	28,045 (60.9%)	11,377 (24.7%)	6,644 (14.4%)	34,898 (64.6%)	16,722 (30.9%)	2,434 (4.5%)
9	27,214 (57.4%)	13,130 (27.7%)	7,073 (14.9%)	33,430 (60.9%)	18,408 (33.5%)	3,090 (5.6%)

As Table 5 shows, the majority of test takers in each grade level were Hispanic, and this reflects trends in the overall make-up of the U.S. ELL population. Data about test takers' home languages is not available, but it can be reasonably assumed that Spanish is the home language for most Hispanic students. According to the National Center for Educational Statistics, there were 4,635,185 K-12 English Language Learners enrolled in U.S. public schools in 2011-2012; 3,562,860 students, or 76.9% of the total ELL population, reported Spanish as their home language. Arabic, Chinese, and Vietnamese were the three most frequently reported home languages for ELLs other than English or Spanish (National Center for Education Statistics, 2016).

As described earlier, data for the study comes from the Grades 3-5, 6-8 and 9-12 Tier C forms of the writing test. Table 6 shows the ELD levels of students completing the Tier C form of the writing test for each grade-level. These proficiency levels are based on overall writing composite scores, with raw to scale score conversion and task weighting applied.

Table 6. Writing ELD levels by grade level and test year (Tier C forms only)

ELD Level:	1 n (%)	2 n (%)	3 n (%)	4 n (%)	5 n (%)	6 n (%)
Grade and Testing Year						
Grade 3 2010-11	96 (0.3%)	677 (1.9%)	5270 (14.7%)	24318 (67.7%)	5501 (15.3%)	61 (0.2%)
Grade 3 2011-12	111 (.02%)	629 (1.4%)	4136 (9.3%)	28207 (63.2%)	11297 (25.3%)	243 (0.5%)
Grade 6 2010-11	247 (1.2%)	1807 (8.6%)	12444 (59.1%)	6464 (30.7%)	95 (0.5%)	2 (0.0%)
Grade 6 2011-12	405 (1.7%)	1573 (6.6%)	15364 (64.2%)	6527 (27.3%)	46 (0.2%)	0 (0.0%)
Grade 9 2010-11	215 (1.3%)	280 (1.6%)	1779 (10.3%)	10738 (62.4%)	3972 (23.1%)	212 (1.2%)
Grade 9 2011-12	242 (1.1%)	269 (1.2%)	1224 (5.7%)	7705 (35.6%)	10506 (48.6%)	1676 (7.8%)

Table 6 provides a sense of the overall distribution of proficiency levels for the student population each year. The majority of students completing the Tier C writing test form have a writing proficiency level of 3 or 4. This is because, if tier placement rules are correctly applied, lower proficiency students should not be assigned the Tier C test form. There are fewer students at higher proficiency levels because these students would likely be exited from English language services and thus not included in the test taker population. It is important to note that the proficiency levels listed here are the result of scale score conversion and do not directly relate to the score levels used in the study, which are based on rubric scores for the MS writing task.

3.3 MS writing tasks

This section describes the writing tasks used in the study. Each of the three Tier C writing tasks was administered across two different testing years, 2010-11 and 2011-12, and then retired from operational use as part of the test refreshment cycle. I selected these tasks from a small pool of retired test tasks because of the WIDA Standards they address (Language of Math and Language of Science). I selected MS as the focal task type for this study because these tasks have rich input and, compared with the other two tasks on the WIDA writing test, are relatively constrained in the types of responses they elicit. For example, other types of task may rely more on background experience or opinions. My expectation was that responses to MS tasks would adhere closely to the task input and this would best be able to show patterns of how students use language from the task input in their responses.

Table 7 provides an overview of the three test tasks used in the study, including a description of the content, the grade-level, the WIDA English Language Development (ELD) levels targeted by the task, and the WIDA Standards that the tasks are aligned to.

Table 7. Summary of writing tasks included in the study

Task title	Grade-level cluster	Content	ELD levels	WIDA Standards
Electrical Circuits	3-5	Using lightbulbs and batteries to examine the flow of electricity through circuits	3-5	Language of Math and Language of Science (MS)
Using Scientific Instruments	6-8	Using scientific instruments to grow tomatoes for a science project	3-5	Language of Math and Language of Science (MS
Conservation of Energy	9-12	Applying the law of conservation of energy to an experiment using toy cars	3-5	Language of Math and Language of Science (MS)

As shown in Table 7, all three tasks targeted WIDA ELD levels three through five. Levels 3-5 are at the higher end of the six-point proficiency scale. At level five, student writing is expected to be comparable to fully English-proficient peers at the same grade level.

Each of the three tasks is based on two WIDA Standards: Language of Math and Language of Science, abbreviated as MS. MS tasks are designed to elicit grade-level appropriate language related to these content areas, as described in the WIDA Standards. Task input is aligned to these standards, and tasks are designed to reflect academic content similar to what students may encounter in math or science courses. The content of the tasks should be familiar to most students within the grade-level cluster. However, because the construct of the assessment is academic language proficiency and not content knowledge, the task input is intended to provide students with all the information needed to formulate a response.

Task input provided to students includes a task title, orientation text, a task graphic with text labels and captions, and a task prompt, which is defined here as the task question or directive to which the student responds. Table 8 describes key features of the task input for the three tasks included in the study including the total number of words and the exact text of the test prompt to which the student responds.

Table 8. Task characteristics and prompt wording

Task	Total input words	Prompt wording
Grade 3: Electrical Circuits	107	Explain how solving the problem with lightbulb B changes the flow of electricity in these circuits. Write a paragraph of 6 to 8 sentences explaining your answer.
Grade 6: Using Scientific Instruments	103	Look at the information about growing tomatoes. Write at least 8 sentences explaining how Alex will use the tools to help him grow healthy tomato plants.
Grade 9: Conservation of Energy	188	Describe what happens to the toy car’s energy and explain the steps Omri used to calculate the kinetic energy of the toy car at Point B. Write a paragraph of 8 to 10 sentences.

When calculating the number of words in the input, I included all text, titles, section headers, task-specific directions, and graphic labels in the calculation. Text related to general test information was not included. For example, page numbers and copyright information that appeared on each test page were not included in the word counts.

The task prompts explicitly state the length of responses students are expected to produce. In grades 3-5, students are instructed to write six to eight sentences. Grades 6-8 students are instructed to write "at least 8 sentences," and students in grades 9-12 are instructed to write, "a paragraph of 8 to 10 sentences." These instructions are designed to support students in understanding task expectations. However, students do not have to write a particular number of sentences in order to achieve a certain score, and raters did not calculate the total number of sentences students produced when scoring.

The test tasks are designed to allow students to demonstrate their English language ability rather than content knowledge. This means that all information students need to respond appropriately is provided in the task input. However, in each task, students do need to understand, synthesize and apply the input in order to formulate a response. For example, in response to the Grades 3-5 Electrical Circuits task, students read a short explanation of how electricity travels through wires and then must apply this to explain why a lightbulb depicted in the graphic is not working. In responding to the Grades 6-8 Using Scientific Instruments task, students must synthesize information about growing healthy tomatoes with a list of tools. In responding to the Grades 9-12 Conservation of Energy task, students are presented with a short description of the law of conservation of energy and with a formula for calculating energy use. Students must then apply this information in an explanation of how a toy car moves down a ramp. Thus, while each task provides students with information to respond, each task does include an element of application or critical thinking so that responses are not simply recounting task input.

Task graphics are a fundamental component of task input. Table 9 describes the type of graphic input provided for each task included in the study. All task graphics are full color and presented on a single page in the testing booklet.

Table 9. Summary of task graphics

Task	Graphic support and presentation
Grade 3: Electrical Circuits	The illustration shows a parallel circuit connected to a battery. Arrows indicate the flow of electricity through the circuit. One lightbulb (A) is connected to the circuit and is working. A second lightbulb (B), is not connected to the circuit because the path is broken. Text boxes and arrows provide students with an explanation of what is happening in the graphic.
Grade 6: Using Scientific Instruments	A text box labeled "How to Grow Healthy Tomatoes" shows a bulleted list of information about growing tomatoes with an illustration of tomatoes. Below this, a chart lists, "Tools for growing tomatoes" and includes a thermometer, rain gauge, and yardstick along with an illustration of each tool. The chart also indicates what each tool measures.
Grade 9: Conservation of Energy	The task graphic shows a toy car going down a ramp. Graphic text includes the equation for calculating total energy, and calculations for potential, kinetic, and total energy at three different points (A, B, and C) in the car's progress down the ramp. The graphic includes math calculations for determining the kinetic energy of the toy car at Point B.

As Table 9 indicates, task graphics also include text such as labels or short explanations of what is being shown. These graphics do not merely support the concepts presented in the task input. Rather, they are central to responding to the prompt appropriately, and students must understand what the graphic depicts and the text included in the graphic.

3.4 Student responses

This section describes characteristics of student responses in the study. A total of 1200 student responses to test tasks were analyzed. The dataset includes 400 responses at each of following grade levels: 3, 6, and 9.

3.4.1 Responses by score level

ACCESS test forms are designed to be administered to several adjacent grade levels. Each test task in this study appeared on one of three different test forms: Grades 3-5, Grades 6-8, and Grades 9-12. All task content is appropriate for the lowest grade-level within the band. For example, content on the Grades 3-5 form of the test is accessible to students in grade 3. The task responses analyzed in the study come from students in only three grade levels: grade 3, grade 6, and grade 9. These students completed the test task from the corresponding grade-level cluster.

Within a grade-level cluster, responses may vary due to developmental differences across grade levels. The study focuses on one grade-level within each grade-level cluster so that variations in grade-level do not affect the analysis of each task.

Responses are a stratified sample of all students at a grade level (3, 6 or 9) who took the Tier C form of WIDA ACCESS in 2010-2012. Within each grade level, responses were stratified across four different score levels: 2, 3, 4, and 5. For each of these score levels, the papers represented the following score profiles according to the writing rubric: 2-2-2, 3-3-3, 4-4-4, and 5-5-5. These scores reflect consistent subscores in each rubric category (Linguistic Complexity, Vocabulary Usage, and Language Control). The distribution of scores for level 5 responses deviated from the pattern of consistent scores for each category because of the limited availability of high-level samples. This is explained further below.

Table 10 lists the number of responses analyzed in the study by grade and score level. There are 100 responses per cell. The data request included 20% overage to allow for responses with illegible handwriting or scanning problems to be discarded.

Table 10. Number of responses by grade and score level

Grade	Task	Number of responses by score level				Total number of responses
		Level 2	Level 3	Level 4	Level 5	
Grade 3	Electrical Circuits	100	100	100	100	400
Grade 6	Using Scientific Instruments	100	100	100	100	400
Grade 9	Conservation of Energy	100	100	100	100	400

As Table 100 shows, the study included a total of 400 responses for each task. However, because there were a limited number of level 5 samples available, the data at this level the does not necessarily reflect a consistent score profile. If 120 papers with a score of 5-5-5 were not available, then the 120 highest scoring papers in the set of operational tests were sampled, beginning with score point 6-6-6 and moving downward through lower scores until the requisite number of papers were reached. Table 11 lists the score profiles for level 5 responses by grade.

Table 11. Score profiles of level 5 responses by grade.

	Grade 3	Grade 6	Grade 9
445	12		
454	26		
455	6		
544	14		
545	8		
554	15	42	
555	15	52	100
556	1		
565		1	
566	1		
655		2	
665	1	1	
666	1	2	
Grand Total	100	100	100

As this table shows, at grade 9, there were a total of 100 responses available at score point 5-5-5. Thus, all responses came from this category. For grade 3 and grade 6 responses, there were not sufficient responses at this level. Again, it is important to note that the ELD level distribution of the test taker population (described in Table 6) and the task scores are different. ELD levels are calculated based on composite writing scores to all three tasks with raw to scale-score conversion and task weighting. Thus, students with a composite proficiency level of 5 in writing may not have scored a 5-5-5 on the MS task. Scores at the highest end of the writing rubric are relatively rare.

3.4.2 Student background

This section describes student background data for the responses in the study. The dataset included information about test taker gender and the state where the test was administered. Table 12 describes the student gender distribution.

Table 12. Gender distribution of study data by grade level

Grade level	F n (%)	M n (%)	Missing n (%)
Grade 3	201 (50.3%)	199 (49.8%)	0 (0%)
Grade 6	213 (53.5%)	187 (46.8%)	0 (0%)
Grade 9	192 (48%)	205 (51.3%)	3 (0.8%)
Total	606 (50.5%)	591 (49.3%)	3 (0.8%)

As shown in Table 122, the gender distribution in all grades is fairly equal. Although there are a slightly higher percentage of male students in the overall test taker population for all grade levels, there is a slightly higher number of female students in the grade 3 and grade 6 data for this study. Table 133 shows information about the number of responses by state.

Table 13. State distribution of writing responses by grade level

	Grade 3	Grade 6	Grade 9	Total
State	n	n	n	
AK	1	3	3	7
AL	6	7	3	16
DC	3	3	6	12
DE	5	4	4	13
GA	46	38	32	116
HI	17	13	7	37
IL	67	82	70	219
KY	11	5	17	33
MD	12	9	2	23
ME	5	6	5	16
MN	10	16	13	39
MO	15	10	6	31
MS	4	2	2	8
NC	10	14	16	40
ND	3	2	5	10
NH	11	2	4	17
NJ	33	14	14	61
NM	15	38	39	92
OK	11	14	10	35
PA	28	27	37	92
RI	3	4	2	9
SD	1	1	1	3
VA	54	44	56	154
VT	0	0	3	3
WI	28	39	41	108
WY	1	3	2	6

A total of 26 different states are represented in the data. The largest number of responses came from Illinois, Virginia, Georgia, and Wisconsin. This distribution reflects the overall distribution of students in the WIDA consortium at the time (Yanosky et al., 2012, 2013).

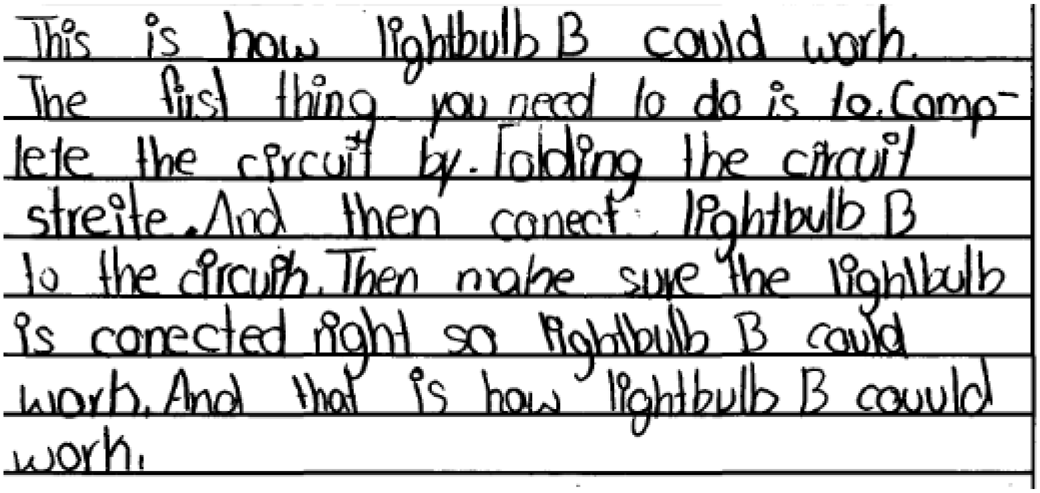
3.5 Data transfer

With authorization from WIDA, data for the study were delivered to me electronically from MetriTech, Inc., the organization responsible for operational scoring of all WIDA ACCESS writing test responses during operational testing in 2010-2012. MetriTech was also responsible for maintaining electronic records of all test responses. All data were transferred using a secure file share system. Data included .tiff image files of student responses and a spreadsheet linking anonymous response ID numbers to background information. This information included each test taker's gender, state, and task score. No additional information was provided about the response files or the test takers.

3.6 Transcription

Image files were transcribed by trained transcriptionists as .txt files using Microsoft word and saved using the same ID number as the image file. See Appendix A for transcription procedures. During the transcription process, spelling errors were typically corrected. These corrections were made to allow for analysis using computer-based tools. Figure 2 shows a sample response image file and the corresponding transcription.

Image File (Grade 3, 4-4-4, ID 10153630)



This is how lightbulb B could work.
The first thing you need to do is to.Comp-
lete the circuit by.Folding the circuit
streite.And then conect lightbulb B
to the circuth.Then make sure the lightbulb
is conected right so lightbulb B could
work.And that is how lightbulb B couuld
work.

Transcribed Text

This is how lightbulb B could work. The first thing you need to do is to complete the circuit by folding the circuit straight. And then connect lightbulb B to the circuit. Then make sure the lightbulb is connected right so lightbulb B could work. And that is how lightbulb B could work.

Figure 2. Sample transcription

As shown in Figure 2, spelling was standardized for “straight.” This practice of correcting student spelling was generally followed during transcription. However, invented words or non-standard words were not corrected. For example, if students wrote “gonna” for “going to,” this was maintained in the transcription. Because sentence-final punctuation was necessary in the analysis software used in the study, punctuation was added or modified in student texts as needed in order to form sentences.

I personally transcribed grade 3 and grade 9 responses (*n* = 800). Grade 6 responses (*n* = 400) had been previously transcribed by a group of three trained researchers as part of a separate research study. The same transcription conventions were followed for all texts, and I reviewed all grade 6 texts to ensure that transcription conventions were consistent.

3.7 Coding and analysis

To address the research question, I analyzed the texts in three distinct phases. Table 144 summarizes the three different analysis phases, the number of responses included in the phase, and the research questions addressed by the results of each phase.

Table 14. Summary of phases of analysis

Analysis phase	Description	Number of responses per grade-level	Research question addressed
I Word-level analysis	Identification of all words copied verbatim or modified from the task input; analysis of frequency of use of different content words	400	RQ1, RQ2
II Phrase-level analysis	Identification of strings of more than four words exactly copied or minimally revised from the task input	400	RQ2, RQ3
III Qualitative coding	Coding and qualitative description of rephrasing patterns and text-level features	100	RQ2, RQ3

As Table 14 shows, 400 texts in each grade level were included in the word and phrase-level analysis, and 100 texts from each grade level were coded using qualitative coding. The following sections describe the procedures for each analysis phase in detail.

3.7.1 Word-level analysis

All words in student responses that were exactly copied or modified from the task input were identified. Within each category (copied from the input or modified), words were further identified by content and function categories and by grammatical function. Content words, or lexical words, are words that carry information in a text (Biber, Conrad, & Leech, 2002, p.15).

This class of word is an open class, meaning new words can be added to it, and includes noun, lexical verbs, adjectives, and adverbs. Function words are words that indicate relationships and help interpret units (Biber, Conrad, & Leech, 2002, p. 16). Function words are closed class of words, and include prepositions, coordinators, auxiliary verbs, and pronouns.

To develop codes for each grade level, I first categorized each word from the task input as content or function. Next, I identified the grammatical category of each word. For content words, I then listed the possible modified forms that might occur in the responses. For example, this included the singular form of plural nouns, different verb forms, and adjectival forms derived from nouns. The modified forms also included the use of words in different grammatical categories than what was represented in the input. For example, “water” occurs as a noun in the grade 6 Using Scientific Instruments task input but occurs as both a noun and as a verb in student responses. Thus, I identified “water” used a as a noun as a copied input content word and “water” used as a verb as a modified input content word. Table 155 lists the word-level codes used in the study. The codes include words taken directly from the task input (input words) and words modified from the task input (modified words).

Table 15. Summary of word-level codes

Code	Meaning	Description
[INC]	Input Content Words	Content words (e.g., nouns, verbs) copied from the task input.
[INF]	Input Function Words	Function words (e.g., auxiliary verbs, prepositions) copied from the task input.
[MS]	Math Symbols (grade 9 only)	Math symbols (e.g., =, +) copied from the task input.
[MOC]	Modified Content Words	Content words (e.g., nouns, verbs) in a modified form from how they occur in the task input.
[MOF]	Modified Function Words	Function words in a modified form from how they occur in the task input.

I treated mathematical symbols, which appeared only in the grade 9 task input, as a special category of input words. These are addressed in more detail later in this section. During coding, I noted the grammatical category of each word. For example, I coded nouns copied directly from the task input as [INC_N]. A complete list of codes, including grammatical categorization, is included in Appendix B. Because of differences in the task input by grade level, there are separate codes for each task, although many codes are shared.

One challenge that presented itself when identifying a word's grammatical category was coding words that can belong to different grammatical classes depending on use. In these cases, I differentiated content words based on grammatical category (as described above). I assigned function words a general code and did not analyze these further by grammatical category. Table 16 describes how different types of function words were handled in coding.

Table 16. Coding decisions for function words

Issue	Coding decision
Preposition words that can also appear as adverbial participles as part of either phrasal or prepositional verbs	All instances were coded as prepositions
"To" can occur as a preposition or as part of infinitive marker	Did not code for grammatical class; identified only as an input function word
Wh-words can occur as several grammatical classes, including determiners, pronouns or adverbs.	Created wh-words coding category and did not further differentiate grammatical function
It's	Coded as a single unit function word with code [INF_ITS]

The decision not to analyze these function words for grammatical transformations in use was a practical consideration. These function words occur frequently in the texts. Coding specific grammatical uses and modified uses of function words would have been time-consuming and would not yield results directly related to the research questions. The grammatical categorizations of both content and function words are not reported in the study results. These were used primarily to identify modified forms and to facilitate future research. For the purposes of the research questions, the important distinctions are between content and function words.

Other special cases in the data included the use of numbers. The grade 6 Using Scientific Instruments task and the grade 9 Conservation of Energy task both included numerals in the task input. Numerals are a special class of function words that are most commonly used like “determiners or heads in noun phrases” (Biber, Conrand & Leech, 2002, p. 34). Task responses used numeric representations as well as the lexical forms of numbers that appeared in the input. Instances where students wrote out the numerals that appeared in the input were coded as modified function words ([MOF_NUM]). Thus, while function words are a closed class of words and cannot be added to or modified the way that content words can, there were some modified function words included in the coding.

In addition to numbers, the task input for Conservation of Energy also included mathematical symbols, or symbolic representations of mathematical concepts (i.e., “+” “-,” and “=”). These were frequently incorporated into student responses. For example, students discussed mathematical operations in sentences and wrote the equations. These were coded as math symbols ([MS]. Instances where students wrote out these words (i.e., “equals,” “plus”) were treated as original language and not coded. Because math symbols could not be processed by the analysis software, math symbols were replaced within the texts as letters (“pls” for “+”, “mns” for “-“ and “eqs” for “=”) so that the math symbols could be coded and processed for analysis. Numbers and math symbols were counted as part of the total word count, with each number or symbol counting as a single word.

After I identified and checked all word-level codes for a task, I created a code book, then identified and coded all instances of the words within the response texts. The complete procedures for coding texts are described in Appendix B along with a list of how all special cases were handled (e.g., math symbols, wh-words). In order to assign codes, I copied all responses from text files into a single Microsoft Word document. Next, I used the search and replace function in Microsoft Word to identify and code each word. This process allowed the reliable identification and efficient coding of all words in the code book. After initial coding was complete, I conducted two data checks in order to confirm that all instances of a word had been captured and correctly coded. The data checking process is included in the procedures document.

I formatted the responses and codes according to requirements for the analysis software. The software required that each sentence start on a new line and begin with a code for the writer (in this case, “S” for student). Figure 3 shows a formatted grade 3 response with word-level codes applied.

S This[INF_AUX] is[INF_AUX] how[INF_WH] lightbulb[INC_N] B[INC_ALS] could work.

S The[INF_ART] first thing you need to[INF_TO] do is[INF_AUX] to[INF_TO] complete[INC_ADJ] the[INF_ART] circuit[INC_N] by folding the[INF_ART] circuit[INC_N] straight.

S And[INF_COOR] then connect[MOC_VB] lightbulb[INC_N] B[INC_ALS] to[INF_TO] the[INF_ART] circuit[INC_N].

S Then make sure the[INF_ART] lightbulb[INC_N] is[INF_AUX] connected[INC_PP] right so lightbulb[INC_N] B[INC_ALS] could work.

S And[INF_COOR] that[INF_PREP] is[INF_AUX] how[INF_WH] lightbulb[INC_N] B[INC_ALS] could work.

Figure 3. Example of grade 3 response with word-level coding

After I completed and verified word-level coding within Microsoft Word, I entered responses into Systematic Analysis of Language Transcripts (SALT) software (Miller & Iglesias, 2012). This software is a tool for tabulating codes. Although created for analyzing oral language, SALT was effective in this study as a way to analyze basic feature of texts, including number of words and counts of codes. Texts were run through SALT and results were downloaded by grade level. Appendix C lists the procedures for analyzing transcripts using SALT.

Data from word-level coding is reported as descriptive statistics. Data includes the average percentages of copied and modified words from the task input used at each grade and score level.

I also conducted Kruskal-Wallis H test to determine if there were statistically significant difference in percentages of input language use by score level. I chose this statistical test because the response data for all-three grade levels violated key assumptions of ANOVAs. The Kruskal-

Wallis H test is a rank-based non-parametric test that can be used as an alternative to one-way ANOVA. All analyses were conducted using SPSS version 22. This data addresses the first research question, which focuses on the amount of input language students use in their responses and differences by score level

Word-level coding data is also used to address the second research question, which focuses on the relationship between task input use and task scores. To analyze how frequently texts at each score level used different content words from the task input, I used AntConc, a concordance software (Anthony, 2014). To analyze the frequency of input content words, I conducted a search of all input content words. When conducting a search for each word, I included all forms of the word coded as both directly copied from the task input ([INC]) and all modified forms of the word ([MOC]). I then identified all words which appeared in at least 20% of texts for at least one score level. Data for word frequency is presented as normalized frequencies per 100 words and as the percentage of texts at each level in which the word appears. This data is used to address research question 2a.

3.7.2 Phrase-level coding

The second phase of coding looked at how writers appropriate multi-word strings of text from the task input. During this phase, I identified strings of more than four words that were copied or minimally revised from the task input. During piloting, I used the coding schemes from Shi (2004) and Weigle and Parker (2012) as a starting point, and developed a simplified set of two codes best suited to the response data. I found that a four-word string was as the shortest for which I could reliably identify minimal revisions. Additionally, because word-level coding captured the overall amount of input language use, this level of coding could focus on if and how writers used longer strings of task input in their writing. Table 177 describes the two codes used in this phase along with an example of each.

Table 17. Phrase-level codes

Code	Definition	Example
Exact Copy [EC]	String of four or more words exactly copied from the task input	Task input: An object's potential and kinetic energy will always add up to the same total amount of energy. Response text: Omri toy car energy increases which makes the toy move because when **an object's potential and kinetic energy will always add up to the same total amount of energy**.
Minimal Revision [MINR]	String of four or more words minimally revised from the task input. Minimal revisions are defined as approximately one change per every four words.	Task input: An object's potential and kinetic energy will always add up to the same total amount of energy. Response text: A toy car is **an object that has potential and kinetic energy and that energy will always add up to the same amount of energy**

During this phase of coding, I identified each string of input text by type as well as by the length of the string. Text strings were identified in Microsoft Word using the search function. Appendix D describes procedures for phrase-level coding and data checking.

When analyzing grade 9 responses, I added an additional level of coding to identify math equations that were either copied or minimally revised from the task input. The task input included several equations and formulas, and students frequently incorporated these into their responses. Information about math equations is thus reported separately as patterns of how writers incorporate these into their responses. It is relevant to the research question and somewhat distinct from other patterns of use.

Results from phrase-level coding are reported as descriptive statistics by grade and score level. Results include the rate of strings per 100 words, the percentage of texts that contain one or more string, and the average length of borrowed string. To look at differences by score level, I conducted one-way ANOVAs using results from the phrase-level coding to compare results by score level. I conducted separate ANOVAs for each grade level using score level as the independent variable and rate of strings per 100 words as the dependent variable. These results address the second research question, which focuses on how writers at different grades and score levels take up and use task input.

3.7.3 Qualitative coding

The third phase of text coding involved a qualitative review of responses. The purpose of this phase of coding was to provide a holistic sense of how writers at each grade-level and score point understand and use language from the task input. Because this phase of coding was more intensive than previous phases, I analyzed a subset of 20% of the total response data. I reviewed 20 responses at each score level for a total of 100 texts per grade level. These were selected using a random number generator (Haahr, 2006). To verify that 20 responses per score point were sufficient to develop a sense of overall trends, I created a coding scheme and description of grade 9, score level 2 using 10 responses. Next, I reviewed these codes and the level description using an additional 10 texts. No new codes were added with the addition of 10 more texts. Thus, I determined that 20 responses per level would yield a robust understanding of key features of the level.

Appendix E lists complete procedures for qualitative coding. For this phase I adopted a structural coding approach (Saldaña, 2013). This approach focuses on the categorization of data segments using descriptive, functional codes. Before coding began, I identified a list of general issues to pay attention to. This included: understanding of task content, extent to which the response addresses the task prompt, and patterns of rephrasing of task input. This list of issues was informed by the literature review and my own professional experience with the tasks. Within each grade level, I began coding with level 2 responses and worked upward through the score levels. As texts grew longer and more complex, I added further codes as needed. I then developed codes as I read through texts repeatedly, made notes, and developed a set of features that captured key aspects of the texts relative to the task input. There are some codes that appear across all grade levels and some that apply specifically to the features of one task. Coding was done using printed copies of texts and handwritten notes. I entered data for each response into Microsoft Excel for tabulation. I also compiled typed notes about the responses for review and wrote a brief memo capturing my overall sense of the score level. Although some code quantification is reported in this section, the results are primarily presented through a qualitative description of sets of texts by grade level and score point.

3.8 Comparisons between tasks and grade levels

The third research question addresses differences in task input use that may be related to task or grade-level differences. Because this study relies on operational test data, students from different grade-levels responded to different tasks. This limits direct comparisons between sets of texts. It is difficult to determine if patterns of difference are due to cognitive differences between students or because of other task characteristics. Thus, a comparison across tasks and grade levels is limited to qualitative data and a discussion of patterns by grade level. This component of the study is exploratory and may suggest future directions for research.

3.9 Summary of methods

This chapter has described word-level, phrase-level and qualitative coding approaches used to address the three research questions. The research questions are exploratory in nature and focus on how students at different grade and score levels use language from the task input in their responses.

4 RESULTS

This chapter presents results for the study organized by research question.

4.1 Results: Research question 1

The first research question is:

1. To what extent do students at each grade level use language from the task input?
   a. Are there differences by score level?

This question was addressed through word-level coding in which I identified words in response texts that were either exactly copied or modified from the task input and then analyzed using SALT software. Table 188 lists the codes used throughout this section. More detailed descriptions of these codes and examples of each are included in Chapter 2.

Table 18. Word-level codes

Code	Meaning	Description
[INC]	Input Content Words	Content words (e.g., nouns, verbs) copied from the task input.
[INF]	Input Function Words	Function words (e.g., auxiliary verbs, prepositions) copied from the task input.
[MS]	Math Symbols (grade 9 only)	Math symbols (e.g., =, +) copied from the task input.
INPUT	All input words	All of the content and function words copied directly from the task input.
[MOC]	Modified Content Words	Content words (e.g., nouns, verbs) in a modified form from how they occur in the task input.
[MOF]	Modified Function Words	Function Words in a modified form from how they occur in the task input.
MODIFIED	All modified words	All content and function words that appear in a modified form from how they occur in the task input.

Table 19 shows information about response length by grade and score level. Results include descriptive statistics for texts at each grade-level and score point.

Table 19. Total words by grade and score level

	Score level:	Level 2	Level 3	Level 4	Level 5
Grade 3	**Total Words**				
	Mean (SD)	45.69 (24.30)	53.31 (21.51)	75.46 (19.01)	85.15 (23.32)
	Min	7	17	37	45
	Max	115	117	140	175
Grade 6	**Total Words**				
	Mean (SD)	67.03 (22.10)	87.48 (24.85)	111.04 (27.39)	132.05 (28.61)
	Min	23	30	43	71
	Max	140	165	211	210
Grade 9	**Total Words**				
	Mean (SD)	54.57 (37.19)	83.70 (33.06)	113.57 (28.12)	139.15 (29.38)
	Min	4	19	52	64
	Max	253	178	190	223

As shown in **Error! Reference source not found.**, within each grade level the average number of total words increased at each score level. For example, grade 9 level 2 texts had an average length of 54.6 words. The average text length at level 5 was more than double this length at 139.2 words. Grade 3 texts were the shortest at each score level when compared with texts from grade 6 and grade 9. For example, the average length of grade 3 level 5 texts was 85.15 words, compared with a mean of 132.05 words at the same score level in grade 6. However, because each grade responded to a different test task, it is unclear if differences in length were due to grade-level differences or task design. The prompts to which students responded specified the target response length. As described in Chapter 3, the grade 3 prompt instructed students to, “write a paragraph of 6 to 8 sentences.” The grade 6 prompt specified, “at least 8 sentences,” and the grade 9 prompt, “ a paragraph of 8 to 10 sentences.” Thus the difference in response length may have reflected the task directions rather than fluency differences between grade levels, although the task prompts were developed to reflect reasonable expectations for the quantity of writing students should have been able produce in each grade-level cluster (3-5, 6-8, and 9-12).

Table 20 shows results of word-level coding by grade level. It lists the average percentage of input words ("INPUT") and modified words ("MODIFIED") for each grade level as well as the average for both input and modified input words. This data provides a broad overview of how much input language students at each grade level produced in response to prompts.

Table 20. Percentage of input and modified words by grade level

	Grade 3 (*n* = 400) Task: Electrical Circuits	Grade 6 (*n* = 400) Task: Using Scientific Instruments	Grade 9 (*n* = 400) Task: Conservation of Energy
Total Words			
Mean (SD)	64.90 (27.31)	99.40 (35.59)	97.75 (45.16)
Percent INPUT			
Mean (SD)	61.80 (15.10)	54.93 (11.25)	58.16 (13.18)
Min	22.99	16.92	20.00
Max	100.00	84.91	94.44
Percent MODIFIED	4.19 (3.72)	4.11 (2.71)	1.94 (3.12)
Mean (SD)	0.00	0.00	0.00
Min	22.58	14.74	16.68
Max			
Percent INPUT & MODIFIED			
Mean (SD)	65.98 (14.44)	59.04 (11.70)	60.10 (13.21)
Min	26.44	16.92	20.00
Max	100.00	91.30	94.44

As Table 20 shows, at grade 3 an average of 61.80% of words in student responses were taken from the task input. The average was 54.93% for grade 6 and 58.16% for grade 9. The range is quite large. For example, in grade 9, percentage of language from the task input ranged from 20.00% to 94.44%. Next, Table 20 presents mean percentages for modified words. These are words that represent either a word that is used in a different grammatical category than it appears in the input (e.g., “water” appears as a noun in the task input and used as a verb in responses) or a modification to an input word (e.g., the plural form of a singular noun that appeared in the task input). The mean percentages for modified words were 4.19% at grade 3, 4.11% at grade 6, and 1.94% at grade 9. The final rows in Table 20 show the total mean percentages for both modified and input words. This was 65.98% at grade 3, 59.04% at grade 6, and 60.10% at grade 9. This means that at each grade level approximately 35-40% of the words in responses consisted of original language and the rest came from the task input. Given the relative brevity of the responses, results indicate that relatively few words in responses may have been language beyond that presented in the task. These results show a general picture of the prevalence of input language in test taker responses.

4.1.1 Results by grade and score level

This section presents result by grade and score level. Table 21 presents results for each coding category for grade 3. Results are presented using the coding categories listed in Table 18, including average percentages of both input content ([INC]) and input function ([INF]) words and modified content ([MOC]) and modified function ([MOF]) categories. The “INPUT” and “MODFIED” categories represent both the content and function words for that type. As noted in Chapter 3, only a small number of words were identified as modified function words, and the results for this category are minimal. For grade 3, this category included the use of “one” and “other” as pronouns in response papers. These words appeared as a determiner and semi-determiner, respectively, in the task input.

Table 21. Grade 3 word-level coding results by score level

Score level:	2 (*n* = 100)	3 (*n* = 100)	4 (*n* = 100)	5 (*n* = 100)
Total Words				
Mean (SD)	45.69 (24.30)	53.31 (21.51)	75.46 (19.01)	85.15 (23.32)
Percent [INC]				
Mean (SD)	16.72 (12.30)	22.43 (9.78)	27.04 (7.20)	25.88 (8.33)
Min	0.00	4.44	5.36	8.11
Max	53.85	57.89	43.84	50.82
Percent [INF]				
Mean (SD)	37.83 (10.22)	40.13 (9.10)	39.01 (7.12)	38.15 (6.80)
Min	15.00	10.71	21.95	25.89
Max	64.58	59.79	56.43	58.82
Percent INPUT				
Mean (SD)	54.5 (19.06)	62.56 (14.64)	66.06 (10.80)	64.03 (11.89)
Min	22.99	33.96	36.61	36.94
Max	100.00	100.00	88.06	85.11
Percent [MOC]				
Mean (SD)	3.50 (4.68)	4.75 (3.85)	4.02 (2.90)	4.00 (2.96)
Min	0.00	0.00	0.00	0.00
Max	22.58	20.63	12.24	13.21
Percent [MOF]				
Mean (SD)	0.25 (1.08)	0.07 (0.35)	0.08 (0.38)	0.07 (0.35)
Min	0.00	0.00	0.00	0.00
Max	9.09	2.33	2.82	2.56
Percent				
MODIFIED	3.75 (4.70)	4.82 (3.90)	4.10 (2.95)	4.07 (2.97)
Mean (SD)	0.00	0.00	0.00	0.00
Min	22.58	20.63	12.24	13.21
Max				
Percent INPUT & MODIFIED				
Mean (SD)	58.30 (18.49)	67.38 (13.31)	70.16 (10.30)	68.10 (11.23)
Min	26.44	39.58	42.86	38.74
Max	100.00	100.00	88.64	87.23

Table 21 shows clear patterns of difference by score in grade 3 for some coding categories. Grade 3 level 2 texts had an average input content word use ([INC]) of 16.72% (*SD* = 12.30) while this increased to 22.43% (*SD* = 9.78) at level 3. Levels 4 and 5 were similar to level 3. The standard deviation was also higher at level 2, indicating greater variation between texts than at other levels. Because level 2 texts tended to be short, the percentage of input content words resulted in approximately seven words in a response deriving from the input content words. There are no clear patterns of difference between the average use of input function words ([INF]) by score level, with responses at each level consisting of approximately 40% of function words from the task input. As input function words for the grade 3 task in this study (as well as grade 6 and grade 9) included some of the most frequent function words in English (determiners, prepositions, auxiliary verbs), this finding is not surprising. However, this study focuses on the use of content words only.

The average use of modified content words was relatively low at all score levels, with the lowest average of 3.50% (*SD* = 4.68) at level 2 and the highest average of 4.75% (*SD* = 3.85) at level 3. At level 3 this means that students used an average of two to three modified content words in each response. A review of student responses showed that the content words “connected” and “lightbulb” were among the most frequently modified content words in responses.

The final row of the table tabulates the totals of all codes, input and modified, content and function words. A Kruskal-Wallis H test was run to determine if there were differences in percentage of input language use between the four different score levels. Distributions of percentages were similar for all groups, as assessed by visual inspection of a boxplot. Median percentages of input language use were statistically significantly different between groups, *H*(3)

= 35.543, $p < .0005$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed statistically significant differences between level 2 (*Mdn* = 56.87) and level 3 (*Mdn* = 68.13), level 4 (*Mdn* = 70.89), and level 5 (*Mdn* = 70.29) responses. There were no significant differences between other score levels.

Table 22 presents grade 6 results by coding category and score level.

Table 22. Grade 6 word-level coding results by score level

Score level:	Level 2	Level 3	Level 4	Level 5
Total words				
Mean (SD)	67.03 (22.10)	87.48 (24.85)	111.04 (27.39)	132.05 (28.61)
Percent [INC]				
Mean (SD)	20.07 (9.87)	23.48 (7.86)	26.66 (6.59)	27.53 (6.57)
Min	2.74	7.05	10.93	10.26
Max	54.55	54.17	44.32	43.14
Percent [INF]				
Mean (SD)	27.41 (7.31)	31.90 (6.11)	32.29 (4.60)	30.38 (4.95)
Min	10.26	16.67	19.40	20.51
Max	46.15	44.71	42.86	41.67
Percent INPUT				
Mean (SD)	47.48 (13.72)	55.38 (9.85)	58.95 (7.94)	57.91 (8.77)
Min	16.92	35.54	37.21	35.26
Max	80.43	84.91	78.41	79.71
Percent [MOC]				
Mean (SD)	2.99 (2.57)	3.41 (2.23)	3.47 (2.32)	3.84 (1.99)
Min	0.00	0.00	0.00	0.57
Max	11.11	12.22	9.57	10.53
Percent [MOF]				
Mean (SD)	0.54 (1.70)	0.58 (1.61)	0.65 (1.35)	0.96 (1.51)
Min	0.00	0.00	0.00	0.00
Max	8.70	8.51	5.41	5.26
Percent MODIFIED	3.53 (3.00)	3.99 (2.72)	4.12 (2.48)	4.80 (2.42)
Mean (SD)	0.00	0.00	0.00	0.57
Min	11.11	12.50	10.81	14.74
Max				
Percent INPUT & MODIFIED				
Mean (SD)	51.01 (14.34)	59.38 (10.20)	63.07 (7.92)	62.71 (8.98)
Min	16.92	36.73	43.88	38.46
Max	91.30	85.42	81.69	83.70

Results for grade 6 followed a similar trend to grade 3, with level 2 responses showing the lowest average percentages in all coding categories. Level 2 responses used an average of 20.07% (*SD* = 9.87) input content words ([INC]). Average percentages were 23.45% (*SD* = 7.86) at level 3, 26.66% (*SD* = 6.59) at level 4, and 27.53% (*SD* = 6.57) at level 5, showing a gradual increase by score level. At all score levels there were relatively high standard deviations and large ranges, indicating variation in the amount of input content language used by responses within a level. It is important to note that some responses were quite short. For example, the lowest word count for grade 6 level 2 was 23 words and 30 words at level 3.

There are no clear patterns of difference between score levels for input function words. As with grade 3 responses, this coding category captured many common function words in English, including articles and auxiliary verbs. The number of items coded within the category of modified function words ([MOF]) was, as expected, quite minimal. For grade 6 responses, this primarily included numbers that appeared in the input as numerals and as text in responses. All score levels had approximately 3% modified content words with relatively high standard deviations, indicating that use of these words varied considerably within a given score level. The highest maximum for modified content words was at level 3 (12.22%). Given the average word count for level 3 (87.48), this translates to approximately ten modified content words in a given text. A review of response texts shows that “rain” and “water,” which both appeared in the input as nouns, were among the most frequently modified content words in student responses.

The final row of the Table 22 tabulates the totals of all codes, input and modified, content and function words. This provides an overall picture of the extent to which grade 6 responses at each score level used words from the task input. Level 2 responses had the lowest average percentage at 51.01% (*SD* = 14.34), and this reflects the trends within each coding category. A

Kruskal-Wallis H test was run to determine if there were differences in percentage of input language use between the four different score levels. Distributions of percentages were similar for all groups, as assessed by visual inspection of a boxplot. Median percentages of input language use were statistically significantly different between groups, $H(3) = 55.753$, $p < .0005$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed statistically significant differences between level 2 (*Mdn* = 52.00) and level 3 (*Mdn* = 60.79), level 4 (*Mdn* = 63.24), and level 5 (*Mdn* = 61.89) responses. There were no significant differences between other score levels.

Table 23 presents grade 9 results by coding category and score level. The grade 9 task, Law of Conservation, included math equations and math symbols. As described in the methods section, math symbols, including “+,” “-,” and “=” were counted in the total word count for grade 9 response texts and coded using a category for math symbols ([MS]). I did not code any modifications for these math symbols, such as writing out “equals” instead of using the symbol. However, practices related to modifying and writing out math equations were captured in the phrase-level coding analysis.

Table 23. Grade 9 word-level coding results by score level

Score level:	2	3	4	5
Total Words				
Mean (SD)	54.57 (37.19)	83.70 (33.06)	113.57 (28.12)	139.15 (29.38)
Percent [INC]				
Mean (SD)	16.45 (12.27)	19.54 (10.01)	28.98 (7.44)	30.50 (6.72)
Min	0.00	1.33	12.33	14.78
Max	50.00	42.86	48.98	47.73
Percent [INF]				
Mean (SD)	32.13 (7.77)	33.09 (6.29)	32.66 (5.16)	31.99 (4.99)
Min	13.75	12.28	20.97	21.95
Max	55.26	50.00	45.90	45.38
Percent [MS]				
Mean (SD)	0.41 (1.63)	2.92 (5.54)	2.50 (3.95)	2.25 (2.41)
Min	0.00	0.00	0.00	0.00
Max	9.30	36.00	28.21	10.45
Percent INPUT				
Mean (SD)	49.05 (14.06)	54.98 (13.01)	63.87 (9.76)	64.74 (7.94)
Min	20.00	26.32	42.37	42.15
Max	94.44	89.87	86.96	84.38
Percent [MOC]				
Mean (SD)	0.92 (2.41)	1.42 (2.66)	1.33 (2.20)	1.21 (1.72)
Min	0.00	0.00	0.00	0.00
Max	16.67	15.84	14.52	6.99
Percent [MOF]				
Mean (SD)	0.30 (1.70)	0.30 (1.41)	1.11 (2.44)	1.16 (2.18)
Min	0.00	0.00	0.00	0.00
Max	12.96	11.59	14.52	8.49
Percent				
MODIFIED	1.22 (3.10)	1.73 (3.07)	2.44 (3.20)	2.37 (2.94)
Mean (SD)	0.00	0.00	0.00	0.00
Min	16.67	15.94	15.38	13.40
Max				
Percent INPUT & MODIFIED				
Mean (SD)	50.27 (13.87)	56.71 (12.59)	66.31 (9.60)	67.11 (7.70)
Min	20.00	28.30	43.17	42.15
Max	94.44	89.87	86.96	84.85

For grade 9, there is a pattern of lower input content word use ([INC]) at level 2 (*M* = 16.45%, *SD* = 12.27) and level 3 (*M* = 19.54%, *SD* = 10.01) compared with level 4 (*M* = 28.98, *SD* = 7.44%) and level 5 (*M* = 30.50%, *SD* = 6.72). Standard deviations at higher score levels were also smaller, indicating greater consistency in the use of input content words at these levels. The use of input function words was relatively consistent at approximately 32-33% at all score levels. As at other grade levels, the grade 9 task input included high frequency function words and use of these words in student responses does not necessarily reflect intentional use of input language.

The average use of math symbols ([MS]) is quite low at level 2 (*M* = 0.41, *SD* = 1.63), indicating that these were not used as frequently as compared with other score levels. The averages ranged from 2.25% (*SD* = 2.41) at level 5 to 2.92% (*SD* = 5.54) at level 3. The ranges were also quite large for math symbols, particularly at level 3 and level 4. The maximums for these levels were 36.00% and 28.21%, respectively. In some cases, responses at these score levels consisted primarily of math equations copied or adapted from the task input which would account for the high percentage of math symbols in the response texts.

The use of modified content words ([MOC]) was minimal at all score levels, with a range of 0.92% at level 2 and 1.42% at level 3. A review of responses shows that the verbs “use” and “move” were among the most frequently modified content words from the task input. The use of modified function words was also minimal ([MOF]), which is as expected given the small number of words in this coding category. For grade 9, this category consisted primarily of numerals from the task input written out as text. There were slightly higher percentages of use at score levels 4 and 5 than at levels 2 and 3.

The final row of Table 23 tabulates the totals of all codes, input and modified, content and function words. This provides an overall picture of the extent to which grade 9 responses at each score level used words from the task input. The same patterns seen at grade 3 and grade 6 also held true at grade 9, with level 2 responses having the lowest average percentage of overall input and modified language use (*M* = 50.27, *SD* = 13.87). The average increased at each score level with the highest average at level 5 (*M* = 67.11, *SD* = 7.70). Given the differences by coding category, the categories of input content words and math symbols seem to account for most of the total differences between score levels.

A Kruskal-Wallis H test was run to determine if there were differences in percentage of input language use between the four different score levels. Distributions of percentages were similar for all groups, as assessed by visual inspection of a boxplot. Median percentages of input language use were statistically significantly different between groups, $H(3) = 113.299$, $p < .0005$. Pairwise comparisons were performed using Dunn's (1964) procedure with a Bonferroni correction for multiple comparisons. This post hoc analysis revealed statistically significant differences between level 2 (*Mdn* = 50.00) and level 3 (Mdn = 56.36), level 4 (Mdn = 67.83), and level 5 (Mdn = 67.71) responses. Input language use at level 3 was also significantly lower from levels 4 and 5.

4.1.2 Summary of results for research question 1

Research question 1 focused on the extent to which responses at each grade and score use language from the task input. In student responses, words from the task were coded as either taken directly from the task input or modified in some way, and within each category (input or modified) coded as either content or function words. Grade 9 included an additional coding category of math symbols to capture how responses used math equations included in the task graphic.

Across all grade levels, the mean percentages of task language use for all coding categories ranged from approximately 50 to 70 percent. Standard deviations and ranges were relatively large, indicating a large amount of variation within each score level. About one quarter of responses at levels 3, 4 and 5 consisted of input content language. The average percentage of input content language in responses was lower at level 2 for all grade levels. Additionally, for grade 9 data, level 2 responses had a lower average percentage of math symbol use, and these were infrequently used at this score level compared with other levels. The use of function words did not show any clear patterns by score level, and as noted throughout the results, the task input at all grade levels included high-frequency function words such as determiners and auxiliary verbs. Use of these words in responses texts does not necessarily indicate the intentional use of language from the task input by test takers. The percentages of modified function words were also quite small, which is expected given the relatively few words included in this category. For grades 6 and 9, this consisted primarily of numbers presented as text.

These results for research question 1 provide a general overview of the extent to which responses in grades 3, 6, and 9 used language from the task input, and they suggest some patterns of difference by score level. The remaining research questions provide a picture of patterns of language use by grade and score level and explore qualitative differences between score levels.

4.2 Results: Research question 2

This section presents results for research question 2, which asks:

2. For each grade level, what linguistic patterns emerge in terms of how students at different proficiency levels use language from the task input?

a. To what extent do students at each score levels use different content words from the task input?

b. To what extent do student at each score level appropriate strings of text from the task input?

c. What patterns of input language use characterize each score level?

The focus of this research question is patterns of difference between score levels in terms of how responses use language from the task input. The research question includes three sub-questions, and these were addressed through separate analysis procedures. Question 2a, referring to the extent to which students at each score level use different words from the task input, is addressed using results from an analysis of responses in AntConc, a concordance program (Anthony, 2014). This data includes content words from the task input that occur in at least 20% of texts for at least one score level at a given grade level. When calculating the frequency of words in the response data, all modified forms of the word were included.

Words are organized in two categories: high frequency words and less frequent words. High frequency words include content words other than verbs that appear in the top 1000 most frequent words in English according to word frequency data from COCA, the Corpus of Contemporary American English (Davies, 2008). The less frequent words category includes words other than verbs that do not appear in the top 1000 COCA words. Data about content word use includes the COCA ranking, if available, the normalized frequency in response texts per 100 words, and the percentage of responses at each score level which contain at least one instance of the word. I chose to list words by COCA frequency in order to explore whether or not there were any patterns of use based on word frequency information. After reviewing available word lists, I chose COCA word frequency data because I was able to obtain data on many of the words from the tasks.

Question 2b is addressed through data from coding strings of text appropriated from the task input. Results are presented by grade level, and within each grade level the different components of the research question are addressed. I coded strings of four or more words from the task input using the coding categorized in Table 24.

Table 24. Summary of phrase-level codes

Code	Meaning	Description
[EC]	Exact Copy	String of four or more words exactly copied from the task input
[MINR]	Minimal Revision	String of four or more words from the task input with minimal revisions (approximately one change for every four words)
[MEEC]	Math Equation Exact Copy (grade 9 only)	A math equation of four or more distinct units exactly copied from math equations in the task input
[MEMINR]	Math Equation Minimal Revision (grade 9 only)	String of four or more distinct units from from math equations in the task input with minimal revisions (approximately one change for every four words)

The codes for math equations apply only to grade 9 responses. Codes for exact copies and minimally revised strings apply to all grade levels.

Question 2c is addressed with results from qualitative coding. For the qualitative analysis I analyzed a subset of 20% of texts within each grade level. Results are presented by grade level and within each grade level organized by sub-question. The grade-level results also includes sample responses for each score level, and I selected these responses to represent typical features of each score level in terms of total number of words, percentage of language from the task input, and use of borrowed strings.

4.2.1 Grade 3 results

This section presents grade 3 results for the three different components of research question 2. The results include the frequency of different content words, phrase-level coding, and qualitative analysis for each score level.

4.2.1.1 Frequency of input content words

This section presents the frequency of different input content words in grade 3 texts by score level. Table 25 includes content words that appear in at least 20% of grade 3 texts at one score level. Data includes the normalized frequency at each score level per 100 words, the percentage of texts which included at least one instance of the word, and the COCA frequency ranking for each words. The word "lightbulb" did not appear in COCA and often appears as two separate words in other contexts. However, this word was preserved as it appears on the test. "Light" is among the 500 most frequent words in English, and the frequency data for "lightbulb" includes both "light" and "bulb" as modified forms of the word.

Table 25. Frequency of content word use by score level for grade 3

Content Word	COCA Frequency	Level 2	Level 3	Level 4	Level 5
		High frequency words			
problem	171				
Per 100 words		0.26	0.15	0.46	0.48
% of texts		6.00	7.00	24.00	28.00
part	178				
Per 100 words		0.44	0.21	0.34	0.27
% of texts		15.00	8.00	20.00	19.00
change	307				
Per 100 words		0.33	0.02	0.29	0.53
% of texts		10.00	1.00	18.00	28.00
		Less frequent words			
complete	1211				
Per 100 words		0.37	0.32	1.13	0.93
% of texts		14.00	13.00	51.00	42.00
path	1343				
Per 100 words		0.44	0.21	1.25	0.80
% of texts		19.00	10.00	45.00	42.00
connect	1645				
Per 100 words		0.39	2.53	2.20	2.14
% of texts		16.00	64.00	75.00	75.00
solves	1839				
Per 100 words		0.09	0.13	0.36	0.32
% of texts		3.00	5.00	16.00	22.00
flow	1997				
Per 100 words		0.44	0.26	1.76	1.56
% of texts		18.00	13.00	69.00	59.00
wire	2692				
Per 100 words		0.20	2.78	1.54	1.94
% of texts		7.00	63.00	53.00	55.00
broken	2717				
Per 100 words		1.07	0.96	1.15	0.80
% of texts		35.00	33.00	57.00	45.00
electricity	2743				
Per 100 words		2.01	1.46	2.21	2.58
% of texts		45.00	38.00	79.00	77.00
battery	3221				
Per 100 words		2.87	2.06	1.36	1.43
% of texts		62.00	63.00	58.00	64.00
circuit	4063				
Per 100 words		0.85	0.38	2.82	2.56
% of texts		18.00	8.00	72.00	73.00
incomplete	n/a				
Per 100 words		0.35	0.19	0.61	0.48
% of texts		15.00	9.00	40.00	33.00
lightbulb	n/a				
Per 100 words		3.33	5.70	5.72	5.60
% of texts		57.00	84.00	95.00	95.00

The word-level results in Table 25 show that some words clearly distinguish between score levels, but the ordering of words by COCA frequency does not reveal any clear patterns. Figure 4 provides a visualization of word-use data. In this figure, words are ordered by the percentage of level 5 responses with at least one instance of use. The numbers in each cell show the percentage of responses that use this word, and these are shaded according to five different bands.

Content Word	Level 2	Level 3	Level 4	Level 5
part	15	8	20	19
solves	3	5	16	22
change	10	1	18	28
problem	6	7	24	28
incomplete	15	9	40	33
complete	14	13	51	42
path	19	10	45	42
broken	35	33	57	45
wire	7	63	53	55
flow	18	13	69	59
battery	62	63	58	64
circuit	18	8	72	73
connect	16	64	75	75
electricity	45	38	79	77
lightbulb	57	84	95	95

0-20%	21-40%	41-60%	61-80%	81-100%

Figure 4. Percentage of grade 3 responses using content words by score level

When presented in this format, word frequency data suggests which content words are most relevant to the task. The words “circuit,” “connect,” and “electricity” were present in at least 70% of responses at levels 4 and 5, and almost all level 4 and 5 responses use “lightbulb.” The percentage of responses that used these words at levels 2 and 3 is lower. This indicates that task essentialness, or the importance of a word to completing the writing task, best explains the patterns of use by score level. The data also suggest that there may be an interaction between word difficulty and the task essentialness of a word, as indicated by word frequency data. For example, “circuit” was widely used by higher-level texts and was highly relevant to task completion. However, it is also a relatively infrequent word in English, which may in part account for the small number of level 2 responses (18.00%) that used this word. A word like “lightbulb,” and particularly the modified form “light” is presumably more familiar to students and thus while there are patterns of difference across score levels, these are not as distinct.

4.2.1.2 Phrase-level codes

Table 26 and Table 27 present the results of phrase-level coding for grade 3 by score level and the total for all grade 3 texts. Table 26 lists the number of instances of each coding category, Exact Copy ([EC]) and Minimal Revision ([MINR]) identified at each score level, the mean number of occurrences by per text, and the percentage of texts which had at least one instance of the feature identified. Table 27 presents the mean string length by score level.

Table 26. Number of borrowed strings for grade 3 responses

Category	Level 2 *n* = 100	Level 3 *n* = 100	Level 4 *n* = 100	Level 5 *n* = 100	Total *N* = 400
[EC]					
Total N	13	8	38	47	106
Per 100 words	0.28 (0.85)	0.14 (0.51)	0.50 (0.88)	0.55 (0.78)	0.37 (0.79)
Percentage of texts	10.00	7.00	28.00	37.00	20.50
[MINR]					
Total N	19	18	71	58	166
Per 100 words	0.56 (1.54)	0.38 (1.36)	0.96 (1.25)	0.72 (1.10)	0.66 (1.34)
Percentage of texts	14.00	15.00	49.00	41.00	29.75
TOTAL					
Total N	32	26	109	105	272
Per 100 words	0.84 (1.87)	0.52 (1.47)	1.46 (1.68)	1.27 (1.44)	1.02 (1.67)
Percentage of texts	20.00	19.00	57.00	58.00	38.50

Table 27. Mean string length in grade 3 responses by score level

	Level 2	Level 3	Level 4	Level 5	Total
[EC]					
Mean (SD)	12.69 (4.41)	12.38 (6.10)	9.03 (4.80)	7.83 (4.39)	9.20 (5.01)
Min	5	5	4	4	4
Max	22	25	22	25	25
[MINR]					
Mean (SD)	11.21 (3.40)	8.78 (4.05)	7.85 (3.14)	8.38 (3.23)	8.52 (3.46)
Min	5	4	4	4	4
Max	15	20	16	18	20
Total					
Mean (SD)	11.81 (3.91)	9.88 (5.06)	8.26 (3.85)	8.13 (3.80)	8.78 (4.15)
Min	5	5	4	4	4
Max	22	25	22	25	25

As Table 26 shows, a small number of phrases (both [EC] and [MINR]) were identified at levels 2 and 3. At level 2, 32 strings of text were identified in 20 different responses. At level 3, 26 strings were coded in 19 different texts. The number of strings identified at higher score levels more than tripled with 109 instances in 57 texts and level 4 and 105 instances in 58 texts at level 5. This means that at higher score levels approximately half of the responses included at least one instances of a string of text taken or revised from the task input.

Table 27 presents the average length of string by score level. The shortest average for both coding categories combined is at level 5 (*M* = 8.13, *SD* = 3.80) with average string length increasing at each lower score point for a mean of 8.26 (*SD* = 3.85) at level 4, 9.88 (*SD* = 5.06) at level 3, and 11.81 (*SD* = 3.91) at level 2. These data indicate that responses at lower score points tended to appropriate longer strings of text. However, the overall frequency of these strings was quite low at these levels. This means that lower proficiency responses did not frequently use copied or minimally revised strings of text, but when they did these strings tended to be long, and this was particularly true for strings of text exactly copied from the task input.

4.2.1.3 Typical grade 3 responses by score level

Figure 5 presents typical grade 3 responses for each score level. These responses were selected to reflect the average word counts, percentage of task language use for each score level and to exemplify how writers used strings of text from the input. Words directly from the task input are in bold, words modified from the task input are in bold italics, strings of copied words are underlined with a solid line, and minimally revised strings of words are underlined with a dotted line.

Grade 3, Level 2
3_055
55 words, 60.00% language from task input

Because when you put **the battery** on **the** lamp **it** will turn on **the *light***. Many people need ***light*** so they **can** do **the** homework work do ***writing***, reading, **and** math. **When the battery is** died you put another **battery**. **When all parts are connected it is *completed*. Electricity can not flow through an *uncompleted*.**

Grade 3, Level 3
3_134
59 words, 67.80% language from task input

Do you know **what happened to *light* B**. **It** got ***broked***. **The electricity** goes **to *light* A**. You know wh**y because the electricity can not flow** over **an incomplete circuit. It** runs **the wires in *light* A.** If **it is complete then electricity is** free. If you hook up ***light* B all the electricity** will go **to light A.**

Grade 3, Level 4
3_257
77 words, 66.23% language from task input
Electricity does not flow through or **to lightbulb B because the wires are not connected to it** so the **electricity** will only **flow through lightbulb A. Because the wires are connected and it has its** own cycle or **path.** But if **the wires** were c**onnected to lightbulb B** there would be **a complete** cycle or **path**. Actually **it** would be **a complete circuit** for everything. Also **the battery** helps. Without **the battery** you couldn’t make **a circuit.**

Grade 3, Level 5
3_390
97 words, 73.20% language from task input
Lightbulb B changes flow of electricity in these circuits because lightbulb B is not connect to the *wire* so **the circuit is not complete.** But if you ***connect* the *wire* the circuit** will be **complete. It** would ***change* the electricity flow** so **it is** more wider **and a** lot more **electricity. Electricity travels through the *wire* from one end of the batter to the other end of the battery. Battery** give **the electricity** power **to flow through.** For example **with battery the electricity** could easily pass through. After you read this you might know more **about electricity.**

Figure 5. Typical grade 3 responses by score level

These examples typify responses at each score level and show how responses integrated borrowed strings of text from the input into original language constructions. Note that the level 2 response did not use “circuit,” which is one of the words that distinguished higher proficiency texts. This word was present in the sample responses for each of the higher score levels.

These examples show that as the response length and sophistication increased at each score level. The figure also visually represents the nominal amount of original language included in each response. At score level 2, this language was related to the task content but not directly relevant to the task prompt (e.g., “lamp,” “homework”) while at higher score levels original language was integrated with language from the task input to provide responses more directly related to the prompt. Simple transition words like “also,” “for example,” and “but” used throughout level 4 and 5 responses contribute to an overall sense of cohesion at these score levels. The next phase of analysis explores these text-level differences in more detail.

4.2.1.4 Qualitative coding

This section describes the results of grade 3 qualitative coding by score level. The prompt for the grade 3 task asked students to “explain how solving the problem with lightbulb B will change the flow of electricity in these circuits.” Table 28 lists the coding categories that emerged from analyzing the grade 3 responses. A total of 20% of the 400 grade 3 responses were included in the qualitative coding, or 20 responses per score level.

Table 28. Grade 3 qualitative coding categories

Coding category	Description
Addresses prompt	Responses were coded for the extent to which they addressed the prompt: does not address prompt, partially addresses prompt, or fully addresses prompt. Entirely off-topic responses were also identified. These responses focused on a related topic but were not directly relevant to the prompt.
Misunderstands input	Evidence that the student misunderstood some part of the task input or the concepts presented in the task. Some responses were coded as partial understanding, meaning that they reflected an accurate understanding of part of the input but did not meaningfully engage with all task input.
Prompt rephrase	Responses coded for whether or not they included a rephrasing or reformulation of the task prompt
Input links	Responses coded for links to the task input (either through exact copying, minimal revisions, or more extensive rephrasing). Each instance was linked back to a particular part of the task input.
Original language	Notations about original language used in responses. This was coded in instances where the language was related to but not directly relevant to the prompt and represented background knowledge or language.

The description of responses at each level addresses these coding categories along with any distinctive features of the level noted during the coding process. Response excerpts throughout this section include a notation of the response identification number.

At level 2, only one of the twenty responses was coded as fully addressing the prompt. Most responses partially addressed the prompt and were topically related to the lightbulb and electricity but did not describe circuits. Rather than describing the flow of electricity, some level 2 responses provided an explanation of why the lightbulb didn’t work and how it could be fixed. In many cases this was based on students’ background knowledge or experience rather than information provided in the task input. For example, one response explained that the light was left on too long, and another response focused on the need to pay for electricity: “If you do not but your electricity your light are going to get off and you need to stay with the light off" (3_039). At this score level, the problem of a lightbulb being off seemed to be an accessible problem for students to write about, and thus responses focused on this issue. A total of eight of the 20 responses used background knowledge not included in the task input to explain why the lightbulb was broken or how to fix it. Three responses focused on uses for electricity (e.g., television, play station).

Responses typically did not bring in language or information from the input about electrical circuits. Only one response of the 20 coded at level 2 discussed circuits: "The electricity can incomplete the path broken. It couldn't be broken. It goes circuit round and round" (3_073). While this response took up some language from the input, including "path," "broken," and "circuit," the response is not particular coherent, and it is not clear that the student can correctly use the input vocabulary to formulate original ideas. This response is an example of a level 2 response that is more related to the task input than responses that relied on background knowledge or language. A total of five of the 20 level 2 responses were identified as picture description responses, meaning that the writing focused on describing the item graphic. For example: "Lightbulb B was broken and it not flowing. And [indistinguishable] A is not flowing and the battery is on the bottom" (3_026). This response described what is happening in the task graphic using both input vocabulary (lightbulb, broken, flowing, battery) along with some original language ("on the bottom").

Three of the 20 level 2 responses were coded as rephrasing the task prompt. In each case, the writer took up and rephrased a short chunk of language from the prompt. The language linked to the prompt is in bold:

1. “I think **the problem with lightbulb B** is because…” ()
2. “It is battery now **solving the problem**.” (3_022)
3. “**It changes it because** it only goes half way…” (3_027)

Overall level 2 responses showed minimal links to the task prompt or to other task input, either through exact copying, minimally revised strings, or through rephrasing. Students at this score level tended to use vocabulary from the task input but did not use extensive rephrasing of task input. This is consistent with the finding that level 2 responses often brought in background language and ideas rather than directly addressing the task prompt.

Typical level 3 responses did not fully address the prompt, but responses at this level did tend to engage more with the task input than did responses at level 2. For example, as with level 2, responses focused on explaining why the lightbulb does not work rather than explaining how the flow of electricity will change if the problem with lightbulb B is fixed. However, level 3 responses tended to base these explanations on the task input rather than background knowledge, even if their understanding of the task input was not complete. The response excerpts below exemplify this:

4. I think lightbulb B doesn’t work because the string is not on it and it has to get a string like lightbulb A. (3_109)
5. Because if the lightbulb B is not plugged to the battery it can’t work. And A is lighting up because it is plugged to the battery. (3_125)
6. I think if we connect more wire it will work and the electricity will go to the wire. (3_140)

In each of these excerpts, the response reflects an understanding of the task graphic and the problem in depicts. The writers also used their own language to describe what was happening rather than using terms from the task input. Terms like “string”, “wire,” and “plugged to the battery” are used instead of “incomplete circuit” or “path.” Similar original terms were used throughout level 3 responses to describe the problem with the lightbulb.

In terms of the extent to which writers at level 3 understood the task input, there were few misunderstandings reflected in the responses. Fourteen of the 20 responses were coded as reflecting a partial understanding of the input, meaning that they engaged with some of the input and reflected this accurately in their responses, but did not engage with other parts of the task input. In most cases, this meant that responses did not discuss circuits or the flow of energy but were able to accurately describe why the lightbulb was broken.

No level 3 responses included prompt rephrasing. Four responses did include other links to task input, and in each case these were longer chunks of copied or minimally revised input rather than rephrasing. Entire sentences or long phrases were included in the responses with little integration with original language at the sentence level. For example, this excerpt shows how the writer copied the graphic label verbatim and used this as a conclusion to the response: “The electricity helps the lightbulb to have more light. **When all the parts in a circuit are connected it is a complete circuit”** (3_182). All four responses that were coded as having links to the task input copied one or more of the graphic labels. This result points to the important role graphics and graphic texts have in shaping student responses.

Level 4 responses were distinct from lower proficiency responses in that they typically reflected a complete understanding of the task input. Only one response in this set was coded as having a partial misunderstanding of the task input. Seven out of 20 level 4 responses were coded as fully addressing the task prompt, and 13 coded as partially addressing the prompt. As with previous levels, responses focused on the problem and how to fix it, but often did not explain the flow of electricity, as directed by the prompt.

Unlike responses at levels 2 and 3, at level 4 writers less frequently brought in background or original language unless it was directly related to the task. Three instances of off-task background language were coded in the responses, and these digressions were relatively brief. Examples of task-relevant original language include the frequent use of “wires” for circuit. Other terms for the circuit at this level included “string” and “square.” When describing the task graphic, one student wrote that, “Lightbulb B will not work unless the circuit gets back together. Because as you see lightbulb A it is working perfect” (3_225). In this response excerpt, the student used the term “circuit” along with original phrases “gets back together” and “working perfect” to describe the task graphic.

One instance of prompt rephrasing was coded at level 4. This appeared as a concluding sentence to a response: “So that’s how I think **I can solve this problem**.” In this case the writer adapted the phrase “solving the problem” from the task prompt. Task rephrasing was not frequent for level 4 responses. However, in nine of the 20 responses there were instances of other rephrasing links to the task input. As with level 3, these tended to be longer strings of copied or minimally revised input, although there are also examples of more extensive rephrasing. Writers at level 4 tended to integrate longer strings into their own sentences rather than including copied sentences verbatim. This is shown in the following excerpts:

7. But for right **now the electricity can not flow into an incomplete circuit**. (3_246)
8. The **electricity can’t flow through** lightbulb B because it is **an incomplete circuit**. (3_291)

In these excerpts the writers were able to adapt the same graphic label ("Electricity can not flow through an incomplete circuit […]") and integrate this into their own sentences. In excerpt h, the writer also demonstrates an ability to apply the general principle described by the label about the flow of electricity to the specific problem described in the task prompt. In general, writers at level 4 tended to copy or adapt longer strings of input from the task graphic and integrate this into original sentences. In one instance, a response directly referenced the task input when using long strings of minimally revised task input: "The caption says that **electricity can not flow through an incomplete circuit because its path is broken**. The other caption says that **when all the parts of a circuit are connected it is a complete circuit**." (3_218). In the level 4 data included in the qualitative analysis, this was the only instance of a response referencing the task input directly.

One limitation of level 5 texts is the score distribution includes texts with a similar score profile to level 4. As noted in the methods section, there were a limited number of high-level texts in grade 3 and so the level 5 texts represent the highest scoring texts from the testing program, but do not represent a consistent score profile. With these limitations noted, there were clear patterns that made level 5 texts distinct during the qualitative analysis.

As with other score levels, level 5 responses did not necessarily fully address the task prompt. Seven of 20 responses were coded as fully responding the prompt, as these responses focused on how fixing the problem would affect the flow of electricity. The fact that fewer than half of level 5 responses completely addressed the prompt suggest a few possibilities. First, it may be that the concepts presented in the task were complex for the grade level. Grade 3 is the lowest grade assigned this task on the grade 3 test form, and it would be interesting to compare grade 3 responses to higher grade levels. Older students with more content background may be more familiar with the task content and thus better able to address the prompt. The findings also suggest that multi-part tasks may be challenging for younger students. They may not attend to multiple parts of the prompt or realize that they need to address different issues. Again, it would be useful to compare grade 3 responses with older students.

For the most part, responses at level 5 reflected a full understanding of the task input and often used a range of vocabulary from the task correctly. Responses incorporated original language and background knowledge, but this was typically done as a brief aside rather than an extended digression. For example, only one response in this set was extensively off-topic and described the battery catching fire. Responses more typically followed the pattern in the following excerpt: “So energy flows through the lightbulb to make light. The battery is a cylinder filled with energy. If you connect B it will light up” (3_324). In this example, a brief aside about batteries using original language is integrated into a relevant response.

Four of the 20 level 5 responses were coded as rephrasing the task prompt.

9. If we **solve the problem with lightbulb B it will change the flow of electricity**. (3_316)
10. That’s how **the flow of electricity changes**. (3_336)
11. **Solving the problem with lightbulb B changes the flow of electricity in these circuits**. (3_345)
12. I think **solving the problem with lightbulb B changes the flow of electricity**. (3_384)

Examples 9 and 10 appeared as concluding sentences in responses. Examples 11 and 12 appeared as introductory sentences. In 9, 10, and 12 the writers integrated longer chunks of copied or minimally revised input language into their own sentences. In example 11, the entire sentence was copied directly from the prompt. Although the use of prompt rephrasing was limited at level 5, it did occur most frequently here and writers adapted longer strings of input than at level 4.

Nine texts were coded for links to other input language. Most links were to the graphic labels, and as at level 4 there was one instance of a response directly referencing the input: “Also it will not be able to have a path because the label says it’s broken” (3_372). In this excerpt, the writer directly referenced the graphic caption and summarized its meaning.

Other links to input language show that at level 5 writers still adapted longer strings of language but integrated this with their own language or rephrased it. Additionally, responses show that writers apply the general principles described in the task graphic to the specific issue presented in the prompt. This is demonstrated in the following excerpts:

13. Lightbulb A has **a complete circuit because all the parts are connected**. (3_367)
14. I think lightbulb B does not work because **the path has been broken**. (3_389)

In each example, language was adapted and applied to the specific context of the prompt and input strings were integrated with original language. This was typical of level 5 responses that used copied strings or rephrasing.

4.2.1.5 Summary of grade 3 qualitative coding

The qualitative results showed clear differences in how responses at each score level understood the task input, responded to the prompt, and used language from the input in their responses. Table 29 summarizes the key characteristics of grade 3 responses by score level in terms of understanding of task input, extent to which responses addressed the task prompt, and response characteristics.

Table 29. Summary of grade 3 responses by score level

Score level	Summary of key characteristics	
	Understanding of task input and extent to which responses address the prompt	Response characteristics and use of language from the task input
Level 2	Level 2 responses typically did not fully address the task prompt. Responses tended to describe the problem with the lightbulb or describe how it can be fixed, but relied on background language and knowledge to do so rather than information from the task input. Responses often reflected a partial misunderstanding of task input.	Level 2 responses were brief. Original language in the response was likely related to the topic but not directly relevant to the prompt and may have been included in digressions and asides. Responses tended to use vocabulary from the task input (e.g., lightbulb, electricity) but did not typically include longer strings or rephrased task input.
Level 3	In general, level 3 partially addressed the task prompt and would typically describe the problem with the lightbulb using concepts from the task input.	They may have used their own language and terms to describe the task graphic (e.g., “string” instead of “circuits”). Some responses may have included some longer strings of copied or minimally revised text and this was not typically reformulated or integrated with original language. In particular, students tended to include sentences from the graphic labels into their responses.
Level 4	Responses may have fully or partially addressed the task prompt. Off-task digressions were relatively brief.	Responses relied on content language from the task input and may also have used original language and terms, particularly when describing the task graphic. Responses may have included longer strings of text from the task input, and this was typically integrated into original sentences. Prompt rephrasing was relatively infrequent.
Level 5	Responses may have fully or partially addressed the task prompt and typically reflected a good understanding of the task input, including vocabulary terms.	Responses tended to use language from the task input, and may also have included longer strings of text adapted from the task input and integrated with original language. Students were able to adapt general principles from the task input (e.g., how electricity flows through circuits) and apply this to the situation set forth in the prompt.

Across all grade 3 responses, the task graphic seemed to be the primary force shaping student responses. Students tended to ignore part of the prompt and input about the flow of electricity and describe the problem with lightbulb B. Their understanding of the problem (or their ability to explain it) varied by score level, with students at level 2 being most likely to have brought in original language and digressions not directly relevant to the prompt.

At level 4, there was a clear shift in the way students integrated strings of text into their responses. While at level 3 there were a small number of longer strings in responses, these were not typically integrated into original language within sentences. At level 4 students were able to integrate longer strings into their own sentences. This pattern emerged even further at level 5, with responses more frequently adapting general principles from the task input to the context of the prompt.

4.2.1.6 Summary of all grade 3 results for research question 2

This section presented results for grade 3 from word frequency data, phrase-level analysis, and qualitative coding. Taken together, these results show distinct patterns of input language use by score level and also demonstrate a progression in terms of how students understood task input and the task prompt. Students at lower proficiency levels sometimes used original language that was not directly related to the task prompt through digressions. This may have been a strategy for responding to the prompt when they did not fully understand what to write, or did not have the language ability to respond appropriately. This finding corresponds with grade 3 results for research question 1, which showed that grade 3 responses have a lower average percentage of input content language. As the score level increases, responses were more fully engaged with both the language and content of the task.

The analysis of borrowed strings also showed patterns by score level. At levels 2 and 3, responses did not typically include copied or minimally revised strings. When they did, these strings tended to be longer than at higher score levels with an average of about 12 words per length of string compared with an average of about 7 or 8 words at levels 4 and 5. At each score level, there were responses which borrowed long strings of input text, but this was atypical.

4.2.2 Grade 6 results

This section presents grade 6 results for the three different components of research question 2. The results include the frequency of different content words, phrase-level analysis, and qualitative analysis for each score level.

4.2.2.1 Frequency of input content words

This section presents the frequency of different input content words in grade 6 responses by score level.

Table 30 includes content words that appear in at least 20% of grade 6 responses at one score level. Data includes the normalized frequency at each score level per 100 words, the percentage of texts which include at least one instance of the word, and the COCA frequency ranking for each word.

Table 30. Frequency of content word use by score level for grade 6

Content Word	COCA Frequency	Level 2	Level 3	Level 4	Level 5
		High frequency words			
want	83				
Per 100 words		0.16	0.16	0.12	0.29
% of texts		11.00	14.00	11.00	30.00
use	92				
Per 100 words		0.33	1.71	2.10	1.98
% of texts		14.00	65.00	78.00	82.00
high	141				
Per 100 words		0.42	0.32	0.51	0.50
% of texts		28.00	25.00	55.00	60.00
help	167				
Per 100 words		0.25	0.85	0.39	0.73
% of texts		12.00	31.00	28.00	47.00
week	188				
Per 100 words		0.70	0.55	0.66	0.70
% of texts		42.00	37.00	63.00	71.00
water	227				
Per 100 words		1.83	1.28	1.09	1.13
% of texts		83.00	69.00	80.00	83.00
air	371				
Per 100 words		0.33	0.66	0.50	0.52
% of texts		20.00	39.00	53.00	57.00
full	504				
Per 100 words		0.39	0.26	0.32	0.24
% of texts		24.00	20.00	35.00	28.00
plant (N)	624				
Per 100 words		1.67	1.42	1.61	1.69
% of texts		65.00	55.00	74.00	80.00
		Less frequent words			
garden	1047				
Per 100 words		0.28	0.22	0.22	0.26
% of texts		14.00	14.00	18.00	23.00
rich	1079				
Per 100 words		0.61	0.48	0.50	1.82
% of texts		39.00	36.00	51.00	98.00
sun	1239				
Per 100 words		1.28	0.79	0.61	0.47
% of texts		63.00	53.00	57.00	50.00
distance	1241				
Per 100 words		0.16	0.25	0.40	0.39
% of texts		11.00	20.00	42.00	45.00
tool	1298				
Per 100 words		0.13	0.94	0.80	1.08
% of texts		8.00	44.00	49.00	60.00
measure	1384				
Per 100 words		0.21	0.95	1.21	1.51
% of texts		10.00	47.00	63.00	83.00
healthy	1476				
Per 100 words		1.07	1.03	0.77	1.10
% of texts		42.00	49.00	55.00	72.00

rain	1559				
Per 100 words		0.42	1.23	2.05	1.82
% of texts		23.00	55.00	97.00	98.00
temperature	1631				
Per 100 words		0.92	1.10	1.07	1.01
% of texts		50.00	69.00	82.00	86.00
deep	1719				
Per 100 words		0.16	0.19	0.30	0.29
% of texts		11.00	17.00	33.00	32.00
soil	1805				
Per 100 words		0.95	0.63	0.67	0.51
% of texts		48.00	43.00	59.00	50.00
seed	1933				
Per 100 words		1.43	0.98	1.31	1.11
% of texts		62.00	49.00	84.00	84.00
apart	1984				
Per 100 words		0.54	0.42	0.61	0.51
% of texts		34.00	35.00	64.00	63.00
tomato	2422				
Per 100 words		3.09	2.79	2.33	2.88
% of texts		80.00	87.00	90.00	97.00
Alex	n/a				
Per 100 words		1.10	1.94	1.93	2.01
% of texts		56.00	65.00	81.00	78.00
centimeter	n/a				
Per 100 words		0.42	0.47	0.65	0.51
% of texts		28.00	39.00	64.00	62.00
Fahrenheit	n/a				
Per 100 words		0.43	0.11	0.48	0.44
% of texts		28.00	9.00	50.00	51.00
gauge	n/a				
Per 100 words		0.04	0.27	1.17	1.13
% of texts		3.00	21.00	96.00	97.00
inch	n/a				
Per 100 words		0.58	0.41	0.76	0.59
% of texts		34.00	33.00	70.00	70.00
thermometer	n/a				
Per 100 words		0.04	0.65	1.07	1.04
% of texts		3.00	51.00	92.00	96.00
yardstick	n/a				
Per 100 words		0.03	0.69	1.15	1.13
% of texts		2.00	51.00	90.00	94.00

Figure 6 provides a visualization of word-use data. In this figure, words are ordered by the percentage of level 5 responses with at least one instance of use. The numbers in each cell show the percentage of responses which use the word, and these are shaded according to five different frequency bands.

Content Word	Level 2	Level 3	Level 4	Level 5
garden	14	14	18	23
full	24	20	35	28
want	11	14	11	30
deep	11	17	33	32
distance	11	20	42	45
help	12	31	28	47
sun	63	53	57	50
soil	48	43	59	50
Fahrenheit	28	9	50	51
air	20	39	53	57
high	28	25	55	60
tool	8	44	49	60
centimeter	28	39	64	62
apart	34	35	64	63
inch	34	33	70	70
week	42	37	63	71
healthy	42	49	55	72
Alex	56	65	81	78
plant (N)	65	55	74	80
use	4	65	78	82
measure	10	47	63	83
water	83	69	80	83
grow	65	79	77	84
seed	62	49	84	84
temperature	50	69	82	86
yardstick	2	51	90	94
thermometer	3	51	92	96
tomato	80	87	90	97
gauge	3	21	96	97
rich	39	36	51	98
rain	23	55	97	98

0-20%	21-40%	41-60%	61-80%	81-100%

Figure 6. Percentage of grade 6 responses using content words by score level

As with grade 3 word-level data, the visual representation of data shows which content words in the input were most relevant for successful task completion. There were a number of content words in the grade 6 input which were used by most of the higher-scoring responses. Words that indicate tools, including “yardstick,” “thermometer,” and “gauge” were used by over 90% of level 4 and 5 responses and show a clear pattern of less frequent use at levels 2 and 3. The task prompt asked students to describe how Alex, a character in the task, will use tools to grow tomatoes. Thus the use of these words not only distinguished between score levels, but also indicates task completion at the higher levels. This issue is explored further in the qualitative coding.

While task essentialness seemed to be the most relevant feature driving the use of particular words (rather than general word difficulty features), there does seem to be some interaction between word frequency data and word use by score level. For example, “gauge” is a relatively infrequent word in English. It was used by fewer level 3 responses than a more high frequency word like “thermometer,” suggesting that this word may not be as accessible to students below a certain score level.

In general, the grade 6 task included a large number of content words and these words appear frequently across score levels. The task input included two separate tables with information and labels, and the task prompt was framed in such a way that students should ideally have used and integrated the language in the tables in order to respond. It may also be that the topic of gardens and plants was generally familiar to students, and thus they had access to the task vocabulary.

4.2.2.2 Phrase-level codes

Table 301 and Table 32 present the results of phrase-level coding for grade 6 by score level and the total for all grade 6 responses. Table 31 lists the number of instances of each coding category, Exact Copy ([EC]) and Minimal Revision ([MINR]) identified at each score level, the mean number of occurrences by per text, and the percentage of texts which had at least one instance of the feature identified. Table 32 presents the mean string length by score level.

Table 31. Number of borrowed strings for grade 6 responses

Category	Level 2 *n* = 100	Level 3 *n* = 100	Level 4 *n* = 100	Level 5 *n* = 100	Total *N* = 400
[EC]					
Total N	38	28	61	69	196
Per 100 words	0.67 (1.22)	0.33 (0.66)	0.57 (0.80)	0.53 (0.75)	0.52 (0.89)
Percentage of texts	29.00	23.00	42.00	42.00	34.00
[MINR]					
Total N	67	76	159	194	496
Per 100 words	1.05 (1.50)	0.96 (1.41)	1.50 (1.20)	1.51 (1.11)	1.26 (1.34)
Percentage of texts	43.00	47.00	77.00	86.00	63.25
TOTAL					
Total N	105	104	220	263	692
Per 100 words	1.72 (2.00)	1.29 (1.73)	2.07 (1.51)	2.01 (1.32)	1.78 (1.69)
Percentage of texts	58.00	53.00	84.00	91.00	71.50

Table 32. Mean string length in grade 6 responses by score level

	Level 2	**Level 3**	**Level 4**	**Level 5**	**Total**
[EC]					
Mean (SD)	5.68 (3.03)	4.93 (1.56)	5.05 (2.05)	5.25 (2.45)	5.22 (2.37)
Min	4	4	4	4	4
Max	17	12	12	13	17
[MINR]					
Mean (SD)	7.36 (3.84)	7.49 (3.32)	7.16 (2.48)	7.09 (2.75)	7.21 (2.94)
Min	4	4	4	4	4
Max	25	21	17	20	25
Total					
Mean (SD)	6.75 (3.65)	6.80 (3.16)	6.57 (2.55)	6.60 (2.80)	6.65 (2.93)
Min	4	4	4	4	4
Max	25	21	17	20	25

The results of phrase-level coding show that about half of all level 2 and level 3 responses include at least one string of text appropriated from the task input and that most higher proficiency responses included at least one borrowed string. Copied or minimally revised strings were identified in 84% of level 4 responses and 91% of level 5 responses. The average length of string is about five words for exactly copied strings at each score level and about seven words for minimally revised strings at each score level. For the grade 6 task, there were a number of short chunks of language in the task input that could easily have been borrowed and integrated into responses. For example, details about conditions for growing tomatoes were presented in a list format and included phrases such as “50 degrees Fahrenheit or higher,” “ water 15-20 centimeters every week.” These short phrases were frequently used in responses, as will be illustrated in the sample responses presented in the next section. The presence of these phrases accounts for the prevalence of appropriated strings of text in higher-level responses as well as the average length of borrowed string. Compared with the grade 3 task, the grade 6 Using Scientific Instrument task included fewer complete sentences in the input which were directly relevant to crafting a response. Thus, while there are certainly instances of longer strings of borrowed text at grade 6, the design of the task seemed to drive a particular type of borrowing which is the integration of short chunks into original sentences.

4.2.2.3 Typical grade 6 responses by score level

Figure 7 presents typical grade 6 responses for each score level. These responses were selected to reflect the average word counts, percentage of task language use for each score level and to exemplify how writers use strings of text from the input. Words directly from the task input are in bold, words modified from the task input are in bold italics, strings of copied words are underlined with a solid line, and minimally revised strings of words are underlined with a dotted line.

Grade 6, Level 2 6_006 66 words, 48.48% language from task input **He *followed* the** directions correctly in order. If you put **seeds** in **the** dirt that are **24 apart**. **It** needs **full *sunlight*. The tomatoes** need **rich health** good **soil for it to plant** in. **The** [indistinguishable] needs 50 degrees **Fahrenheit or higher. Tomatoes** need **water 15-20 centimeters every week. Tomatoes** need care. **And it** really needs **to** be **healthy** so you can eat **it.**
Grade 6, Level 3 6_198 92 words, 59.78% language from task input **Alex** is going **to** need **24 seeds** so **the tomatoes** can **grow. He** is also going **to** need **sun.** Because **the tomatoes** need ***sunlight* to grow. He** is going **to** need **soil to plant the seeds in the soil. He** is going **to** need that **the temperature be at 50 degrees or higher. He** is going **to** have **to** put **15-20 *centimeter*** of **water every** day. Because **the tomatoes** need **water to grow. He** is going **to** need **a** ruler **to measure. How deep the** hole in **the soil** is.
Grade 6, Level 4 6_231 111 words, 63.96% language from task input First, **Alex will** set **the seeds *twenty-four* inches** across **from** each other. **Alex** needs **a thermometer, to** see **the temperature** outside, because if there is bad weather **the tomatoes will** not **grow.** You need **a yardstick to** *measure* **how** many **inches the seeds** need **to** be **apart. The** needs **to** be ***fifty*** degrees **Fahrenheit or higher** so **the tomatoes will grow** perfectly. **The tomatoes** need **a** lot of **sun** so they can **grow.** Finally **Alex** needs **to** have **rich soil** so **he** can **plant the tomatoes** in **the garden, for his science project.** Then **Alex** needs **to use the rain gauge to** put ***fifteen* to *twenty* centimeters** of **water every week.**
Grade 6, Level 5 6_335 132 words, 67.42% Language from task input **Alex will use tools to help him grow health tomato plants. He will use thermometer to** see **the air temperature. It** is important because **tomatoes** are fruits **and** they have **to** be exact **temperature. For** example: **Tomatoes** has **to** be **50 Fahrenheit or higher.** If **it’s** lower than **50 F**, **tomatoes** won’t be **healthy to** eat. **Rain gauge** is **the tool** that **Alex** should **use it. It will measure rainfall and** that tells **him** that **how** much **water he** needs **and *measure* it for him.** There’s another **tool** that is really important **to grow tomatoes healthy. It’s yardstick. Yardstick will measure distance and/or depth. Alex** should **measure distance and/or depth. Alex** should **measure seeds** with **yardstick.** Because **seeds** has **to** be exact **and it** has **to** be **24 inches apart.**

Figure 7. Typical grade 6 responses by score level

These typical grade 6 responses demonstrate how writers at each level use language from the task input in their responses. As noted in the phrase-level analysis, grade 6 responses tended to borrow strings of about five to seven words from the task input. This can be seen in the sample responses as each one used chunks of language from the input related the temperature, the amount of water, and the distance of the seeds. These chunks of language were integrated with original language in ways that reveal different levels of proficiency. For example, the level 2 response and the level 4 response used the same short phrase from the task input about the amount of water needed to grow tomatoes:

> Level 2: Tomatoes need **water 15-20 centimeters every week**.
>
> Level 4: Then Alex needs to use the rain gauge to put **fifteen to twenty centimeters of water every week**.

At level 2 the phrase was inserted somewhat awkwardly into the sentence. At level 4, the integration with the writer's own words was much more seamless. Additionally, this sentence integrated information from two different sections of the task input to apply the need to use a rain gauge to the measure the total amount of water needed each week. Even as responses used similar phrases and vocabulary, the task allowed writers to demonstrate different levels of ability related to how they used this language.

4.2.2.4 Qualitative coding

This section describes the results of grade 6 qualitative coding by score level. The prompt for the grade 6 task asked students to explain "how Alex will use the tools to help him grow healthy tomato plants." Table 33 lists the coding categories that were applied to grade 6 responses. A total of 20% of all responses were included in the qualitative coding (or 20 responses per score level).

Table 33. Grade 6 qualitative coding categories

Coding category	Description
Addresses prompt	Responses were coded for the extent to which they addressed the prompt: does not address prompt, partially addresses prompt, or fully addresses prompt. Entirely off-topic responses were also identified. These responses focused on a related topic but were not directly relevant to the prompt.
Misunderstands input	Evidence that the student misunderstood some part of the task input or the concepts presented in the task.
Prompt rephrase	Responses coded for whether or not they included a rephrasing or reformulation of the task prompt
Input links	Responses coded for links to the task input (either through exact copying, minimal revisions, or more extensive rephrasing). Each instance was linked back to a particular part of the task input.
Original language	Notations about original language used in responses.

The description of responses at each level addresses these coding categories along with any distinctive features of the level noted during the coding process. Response excerpts throughout this section include a notation of the response identification number.

All grade 6 level 2 responses were coded as partially addressing the prompt. The responses tended to focus on how to grow tomatoes or items that are needed in list format, but did not address how Alex would use the tools to grow tomatoes, as specified in the task prompt. The following excerpt illustrates this type of response:

1. The first thing you do to grow healthy tomato plants is the seeds has to be 24 inches apart. But the land has to have bright full sun, rich soil used to make. (6_047)

This example used words and phrases from the task input to provide a relevant response, but did not fully address the prompt. None of the level 2 responses showed a misunderstanding of the task input, indicating that the topic and input were fairly accessible to students.

Few responses were off-topic, but level 2 responses did frequently include brief digressions related to gardening or tomatoes. These were typically a few sentences in length and incorporated original language. For example:

2. Finally you need to make a hole so you could put the seeds there. You need to put something around the seeds so no animals could not eat the tomatoes. (6_036)
3. And he is going to need seeds to make them big and healthy. There are a lot of different kinds of tomatoes. He has to pick the right seeds for it to grow. (6_097)

In these examples, the responses included short digressions that showed some level of background knowledge or experience with the task topic. Original vocabulary at level 2 tended to be used in these type of digressions.

One of the level 2 responses was identified as rephrasing the task prompt. In this response, the student wrote that, “Alex will use the tools to help him grow healthy to use them” (6_055). The responses did not elaborate further how Alex would use specific tools. Fourteen responses were coded as having links to other parts of the input. The task input included bulleted lists of information in short chunks and at level 2, responses integrated these short phrases of four to five copied or minimally revised words into their own writing. For example:

4. […] and the most important the tomatoes need to have good **water 15-20 every week**. (6_009)
5. He also needs to **water them 15-20 centimeters**. (6_072)
6. Put **fifteen to twenty centimeters of water every week**. (6_090)

In each of these three excerpts, students adapted information that appeared in the input as part of a bulleted list into their own sentences. The input states: “Water: 15-20 centimeters every week” as part of a chart labeled “How to Grow Healthy Tomatoes.” As these examples show, level 2 responses may have grammatical errors or awkward phrasing when students tried to integrate phrases from the input into their writing. This was characteristic of input language use at level 2.

Level 3 responses were distinct from level 2 in that 18 responses were coded as fully addressing the prompt. That is, at level 3 most responses addressed how Alex would use the tools to plant tomatoes rather than generally describing how to plant tomatoes, as tended to be the case at level 2. This is illustrated in the following excerpts:

7. I think Alex should use the thermometer first to see how the weather is. (6_127)
8. Alex will use his tools to grow healthy stuff by measuring temperature, rainfall, and distances. (6_134).

As these examples illustrate, level 3 responses engage more directly with the task prompt. This distinction is supported by data from

Figure 6, which lists the frequency of content words by score level. For example, “thermometer” was used by 3% of level 2 texts and 51% of level 3 texts. This supports the qualitative finding that level 3 responses focused more on how tools were used.

Level 3 responses at times integrated original language in the form of short digressions related to growing tomatoes or gardening. These asides often reflect background experience with the topic. However, in general the responses were more focused and relevant to the prompt than at level 2. Three of the level 3 responses were coded for prompt rephrasing.

9. That **how Alex grow healthy tomatoes**. (6_116)
10. **Alex tools will help him grow big red, healthy tomato plants. These tools will help Alex** because […] (6_132)
11. **Using the tools Alex has I think he will grow healthy tomato plants**. (6_193)

Example 9 occurred as a response conclusion. Examples 10 and 11 implemented prompt rephrasing as part of the introduction. In each example, the students rephrase the prompt and integrate language into original sentences that serve an organization purpose within their responses. Overall, rephrasing the task prompt was infrequent in level 3 responses.

A total of 10 level 3 responses were coded for the presence of other links to the task input. As at level 2, these links were typically short phrases from a chart in the task input. For example:

12. **Water it every week** with **15-20 centimeters of water**. (6_141)
13. Last, Alex will use the rain gauge to **water the plants 15-20 centimeters every week**. (6_122)

These examples show greater grammatical accuracy and better integration into sentences than the examples at level 2 using the same phrase. Example *g* also illustrates a response that focused on the use of tools and that integrated information from two charts ("How to Grow Healthy Tomatoes" and a chart showing the tools). While the integration of information from across parts of the task input was not typical at level 3, it did occur as writers adapted language from the task input to address the prompt.

Level 4 responses to this task tended to be clear, straightforward responses that fully addressed the task. Nineteen responses at this level were coded as fully addressing the prompt with limited digressions or asides. Level 4 responses were characterized by the effective use of input vocabulary, although responses did incorporate some relevant background knowledge and experience with the topic.

As with level 3, responses at level 4 focused on describing how tools are used, but often provided more elaboration than responses at lower score points. Responses were often more organized and cohesive at the discourse level rather than providing list-like descriptions of how Alex would use tools. Responses also demonstrated an ability to adapt information from the task chart and build on it using original language. For example:

14. Third, he will need to make sure the temperature is **50 degrees Fahrenheit or higher** and he will use the thermometer to measure the temperature. (6_238)
15. First thing to do when you are growing a healthy tomato garden is use the yardstick to put the **seeds 24 inches apart** from each other. (6_256)

In these two examples, the responses integrated language from the two different charts into clear, coherent descriptions of how to use tools to grow tomatoes. The short chunks of text from the first chart, in bold, were integrated seamlessly into original sentences.

Prompt rephrasing occurred in three of the level 4 texts. The instances are listed below:

16. **How Alex will use the tools** by first using the yardstick. (6_233)
17. In conclusion, that is **how you grow healthy tomato plants with tools**. (6_238)
18. These are **the tools Alex would need to grow the tomatoes**, and to use them. (6_259)

Example 16 used prompt rephrasing in the introduction and examples 17 and 18 used it as a concluding sentence. In each case the prompt rephrasing serves an organizational purpose within the response, although this feature was not particularly common at level 4. Fourteen of the 20 level 4 responses included at least one link to the task input language. Many of the links were to language in the task chart, and at level 4 responses often integrated information from the two charts to address the prompt, thus creating more sophisticated responses.

Responses at level 5 were clear and reflected a full understanding of the task input. All level 5 responses were coded as fully addressing the prompt. While level 4 responses are notable because they effectively used and adapted language from the task input, level 5 responses were often characterized by the use of original language that was directly related to the task prompt. Often a few words or phrases in each response stood out as particularly apt or precise, and these words tended to distinguish level 5 responses from lower proficiency responses. The following excerpts illustrate this type of language:

19. You can also use a different strategy to water your tomato seeds. (3_306)
20. The in [sic] a spacious area, dig a hole, and place the rain gauge in the hole. (6_320)
21. If Alex uses these instruments properly, his tomatoes will be healthy. (6_347)

In each of these examples, a word or short phrase stands out: “strategy”, “spacious area”, “uses these instrument properly.” While not all level 5 responses included notable original vocabulary, it was relatively frequent at this score level and most responses included at least one or two instances of original language used in a sophisticated way.

Six level 5 responses included a rephrasing of the task prompt. Two typical examples are included below:

22. With the tools show, **Alex can use each tool to help him grow healthy tomatoes**. (6_333)
23. **These tools will help Alex** by knowing when it’s ready to water them, harvest them and eat them.

Nineteen level 5 texts included at least one link to task input language. At this level, responses integrated language from the two charts, and also from the task introduction, which introduced the scenario of a science experiment. Compared with other levels, at level 5 prompt rephrasing tended to be more creative as writers put ideas into their own words and integrate both input language and original language.

4.2.2.5 Summary of grade 6 qualitative results

Table 34 summarizes the key characteristics of grade 3 responses by score level in terms of understanding of task input, extent to which responses addressed the task prompt, and response characteristics.

Table 34. Summary of grade 6 responses by score level

Level	Summary of key characteristics	
	Understanding of task input and extent to which responses address the prompt	Response characteristics and use of language from the task input
Level 2	Level 2 responses were often brief and partially rather than fully addressed the task prompt. They most often described how to grow tomatoes generally and may not have mentioned the use of specific tools.	Short chunks of copied or minimally revised language from the task input may have been integrated into original sentences, although it was often characterized by grammatical errors or awkward wording.
Level 3	In general, level 3 responses fully addressed the task prompt by explaining how tools are used to grow tomatoes. The response may have been somewhat list-like and may have followed the organization structure presented in the task graphic.	Level 3 responses often included specific vocabulary from the task input related to tools (e.g., thermometer, yardstick). Short phrases from the task input were integrated into original language, although at times the wording may have been awkward.
Level 4	Level 4 responses fully addressed the task prompt using organized, cohesive discourse.	Responses typically integrated language from different sections of the task input to address the prompt. Words and phrases from the task input were integrated seamlessly into original sentences.
Level 5	Level 5 responses fully addressed the task prompt using sophisticated and organized discourse.	Responses typically reflected notable original words and phrases that were relevant to the prompt and reflected precision, creativity, or sophistication. Language from different parts of the task input was integrated and woven seamlessly into original structures and sentences.

The qualitative analysis shows that in the grade 6 responses, students at all levels took up and used short chunks of language from the task input and integrated this into their writing. The frequency of these short chunks of language is not surprising given the design of the grade 6 task, which included bulleted lists and short phrases as part of the task input. As responses increase in proficiency level, the sophistication of how writers integrated this language, and the grammatical accuracy of the constructions, also increases. At levels 4 and 5, the integration of language from the task input was fairly seamless, and flowed naturally with the writer's original language and structures. Prompt rephrasing did not characterize these grade 6 responses, although this feature was most frequent at level 5.

4.2.2.6 Summary of grade 6 results for research question 2

This section presented results for grade 6 using word frequency data, phrase-level analysis, and qualitative coding. Taken together, these results show distinct patterns of input language use by score level and also demonstrate a progression in terms of how students understand task input and the task prompt. Word use shows that there were a number of words from the task input that distinguished between score levels. When cross-referenced with the results of the qualitative analysis, it seems clear that the words used less frequently in level 2 responses reflect the fact that, in general, these responses did not fully address the task prompt. It seems that students at lower proficiency levels used language and concepts from the task input as well as digressions based on their background knowledge to provide a topically-relevant response.

The analysis of borrowed strings of input text showed that most high-level responses in grade 6 used at least one copied or minimally revised string and that these strings were typically between five and seven words. In this case, the design of the task likely shaped how students use this input as the task graphic included charts with phrases and short chunks of language that students could integrate into their responses. The qualitative analysis showed that the level of sophistication in terms of how students implement this could differentiate proficiency levels.

4.2.3 Grade 9 results

This section presents grade 9 results for the three different components of research question 2. The results include the frequency of different content words, phrase-level coding, and qualitative analysis for each score level.

4.2.3.1 Frequency of input content words

This section presents the frequency of input content words in grade 9 responses by score level. Table 35 includes content words that appear in at least 20% of grade 9 responses at one score level. Data includes the normalized frequency at each score level per 100 words, the percentage of texts which include at least one instance of the word, and the COCA frequency ranking for each word.

Table 35. Frequency of content word use by text level for grade 9

Content Word	COCA Frequency	Level 2	Level 3	Level 4	Level 5
			High frequency words		
now	72				
Per 100 words		0.02	0.14	0.28	0.18
% of texts		1.00	8.00	22.00	20.00
use	92				
Per 100 words		0.24	0.39	0.53	1.32
% of texts		10.00	24.00	32.00	77.00
same	161				
Per 100 words		0.31	0.43	0.22	0.18
% of texts		12.00	20.00	17.00	21.00
point	212				
Per 100 words		1.37	2.22	2.25	2.58
% of texts		35.00	61.00	83.00	89.00
car	290				
Per 100 words		4.58	3.72	2.87	2.52
% of texts		85.00	91.00	96.00	97.00
add	341				
Per 100 words		0.13	0.37	0.38	0.43
% of texts		7.00	19.00	31.00	43.00
explain	481				
Per 100 words		0.00	0.24	0.04	0.63
% of texts		0.00	17.00	3.00	49.00
energy	616				
Per 100 words		3.15	4.52	8.22	8.47
% of texts		54.00	82.00	100.00	100.00
rest	673				
Per 100 words		0.02	0.13	0.23	0.33
% of texts		1.00	9.00	21.00	30.00
amount	782				
Per 100 words		0.11	0.14	0.35	0.43
% of texts		6.00	10.00	21.00	34.00
			Less frequent words		
total	1042				
Per 100 words		0.31	0.93	1.66	1.68
% of texts		11.00	41.00	78.00	81.00
object	1156				
Per 100 words		0.07	0.04	0.32	0.42
% of texts		2.00	3.00	19.00	31.00
potential	1266				
Per 100 words		0.73	1.31	3.08	3.28
% of texts		23.00	39.00	89.00	100.00
transfer	2335				
Per 100 words		0.02	0.06	0.22	0.27
% of texts		1.00	2.00	16.00	27.00
toy	2441				
Per 100 words		1.19	1.05	1.06	1.19
% of texts		40.00	48.00	57.00	72.00
calculate	3064				
Per 100 words		0.13	0.19	0.23	0.30
% of texts		5.00	14.00	22.00	32.00

kinetic	n/a				
Per 100 words		0.82	1.19	3.51	3.72
% of texts		25.00	38.00	97.00	100.00
Omri	n/a				
Per 100 words		0.26	0.79	1.01	1.03
% of texts		8.00	35.00	60.00	78.00

Figure 8 shows a visualization of word-use data. In this figure, words are ordered by the percentage of level 5 responses with at least one instance of use. The numbers in each cell show the percentage of texts, and these are shaded according to five different bands.

Content Word	Level 2	Level 3	Level 4	Level 5
now	1	8	22	20
same	12	20	17	21
transfer	1	2	16	27
rest	1	9	21	30
object	2	3	19	31
calculate	5	14	22	32
amount	6	10	21	34
add	7	19	31	43
explain	0	17	3	49
move	28	25	51	60
toy	40	48	57	72
use	10	24	32	77
Omri	8	35	60	78
total	11	41	78	81
point	35	61	83	89
car	85	91	96	97
energy	54	82	100	100
potential	23	39	89	100
kinetic	25	38	97	100

0-20%	21-40%	41-60%	61-80%	81-100%

Figure 8. Percentage of grade 9 responses using content words by score level

As in other grade levels, the word-use data shows which content words are essential to the writing task. For this task, which asked students to describe how a toy car’s energy changes as it goes down a ramp, the words “potential,” “kinetic,” “energy,” and “car” were used by most high-proficiency responses. For example, “energy” was used by 100% of level 4 and level 5 responses. And while a word like “car” was essential to responding to the task, this word does not differentiate score levels particularly well, as most level 2 responses used “car.” However, only about one quarter of level 2 responses used “kinetic” and “potential.” In this case the word difficulty may interact with the importance of a particular word to completing the task. Both “potential” and “kinetic” are relatively infrequent words in English, and so while they were important to the task they may be less accessible for lower proficiency students.

Another interesting pattern relates to the use of the proper noun “Omri.” This character was presented as part of contextualizing the task input. Omri was conducting a science experiment. While this aspect of the task was relatively minimal and it was not necessary to mention the background context in a successful response, 78% of level 5 responses used “Omri” while only 8% of level 2 and 35% of level 3 responses used this word. This may indicate that students at higher levels were more engaged with the task scenario in their responses. In addition, the prompt explicitly asked students to “explain the steps Omri used” to calculate the car’s energy. The word-level coding showed that at higher levels students used more math symbols in their responses, and thus presumably addressed the math calculation aspect of the task prompt. The fact that higher-level students also used “Omri” more in their responses, which more fully addressed the task, accords with that finding.

4.2.3.2 Phrase-level codes

Table 36 lists the number of occurrences for each phrase-level code by score point. As noted earlier, the grade 9 task input included math equations. These were treated separately and coded as either exact copies ([MEEC]) or as minimally revised ([MEMINR]). Table 36 also lists the percentage of texts at each score point that contained at least one instance of a copied or revised string. For example, at level 2, there were nine occurrences of exactly copied strings identified, and these occurred in 7% of the texts.

Table 36. Number of borrowed strings for grade 9 responses

Category	Level 2 *n* = 100	Level 3 *n* = 100	Level 4 *n* = 100	Level 5 *n* = 100	Total *n* = 400
[EC]					
Total N	9	25	65	92	191
Per 100 words	0.21 (1.02)	0.25 (0.61)	0.61 (0.83)	0.64 (0.76)	0.42 (0.84)
Percentage of texts	7.00	19.00	46.00	52.00	31.00
[MINR]					
Total N	5	27	84	110	226
Per 100 words	0.15 (0.96)	0.34 (0.92)	0.75 (0.92)	0.82 (0.87)	0.52 (0.96)
Percentage of texts	3.00	18.00	52.00	64.00	34.25
[MEEC]					
Total N	2	17	26	45	90
Per 100 words	0.03 (0.25)	0.20 (0.52)	0.22 (0.51)	0.31 (0.59)	0.19 (0.50)
Percentage of texts	2.00	15.00	20.00	29.00	16.50
[MEMINR]					
Total N	11	7	37	58	113
Per 100 words	0.19 (1.01)	0.06 (0.24)	0.32 (0.57)	0.42 (0.58)	0.25 (0.68)
Percentage of texts	4.00	6.00	31.00	42.00	20.75
TOTAL					
Total N	27	76	212	305	620
Per 100 words	0.58 (1.78)	0.85 (1.29)	1.90 (1.52)	2.18 (1.50)	1.38 (1.68)
Percentage of texts	13.00	43.00	86.00	92.00	58.50

As Table 36 shows, the number of each type of string increased in frequency with each score level. At level 5, 52% of texts used an exactly copied string and 64% used a minimally revised string. At level 2, only 7% of texts used an exactly copied string and 3% used a minimally revised string. The use of copied and minimally revised math equations also showed a clear pattern by proficiency level. The use of math equations with minimal revisions increased from 11 instances in 4% of texts at level 2 to 58 instances in 42% of texts at level 5. These data suggest clear distinctions between each score level in terms of the use of input text strings. In particular, there appears to be a pronounced increase between score levels 3 and 4. For exactly copied and minimally revised strings and for minimally revised math equations, the percentage of texts more than doubled between these two score points.

Table 37 lists the mean number of words in each string by score level.

Table 37. Mean string length in grade 9 responses by score level

	Level 2	Level 3	Level 4	Level 5	Total
[EC]					
Mean (SD)	6.89 (4.46)	4.76 (1.88)	5.32 (2.85)	5.89 (3.17)	5.60 (3.03)
Min	4	4	4	4	4
Max	17	12	17	18	18
[MINR]					
Mean (SD)	5.40 (1.02)	6.96 (2.77)	8.25 (3.45)	8.76 (4.57)	8.28 (4.00)
Min	4	4	4	4	4
Max	7	17	18	21	21
[MEEC]					
Mean (SD)	8.00 (0.00)	7.18 (3.50)	5.65 (1.21)	5.93 (2.00)	6.13 (2.26)
Min	8	5	5	5	5
Max	8	18	8	13	18
[MEMINR]					
Mean (SD)	6.00 (3.16)	8.43 (3.99)	7.11 (2.67)	7.24 (2.04)	7.15 (2.58)
Min	5	5	4	4	4
Max	16	17	17	12	17
ALL CODES					
Mean (SD)	6.33 (3.38)	6.42 (3.11)	6.83 (3.21)	7.19 (3.86)	6.94 (3.46)
Min	4	4	4	4	4
Max	17	18	18	21	21

For exactly copied strings, there are no clear patterns by score level. Score level 2 shows the highest average string length (*M* = 6.89, *SD* = 4.46). The range for each score level shows that some writers copied relatively long strings from the task input. The use of minimally revised strings does show an increase in average length by score point, with a noticeable increase between score levels 2 and 3. It is important to note that the number of instances of each type of string at score level 2 was relatively small, and this makes it difficult to draw clear comparisons. For example, at score level 2 there were only two instances of copied math equations while there were 45 instances at score level 5. Taken together with the data in Table 37, these data suggest that students at score levels 4 and 5 and more frequently incorporated strings of all types into their responses, but that there are no clear patterns in the average string length by score level.

4.2.3.3 Typical grade 9 results by score level

Figure 9 presents typical grade 9 responses for each score level. These responses were selected to reflect the average word counts, percentage of task language use for each score level, and to exemplify how writers use strings of text from the input. Words directly from the task input are in bold, words modified from the task input are in bold italics, strings of copied words are underlined with a solid line, and minimally revised strings of words are underlined with a dotted line.

Grade 9, Level 2
9_027
49 words, 48.98% language from task input

What happens to the toy car energy is that first **it** goes faster. Then **it** goes more slowly because **the** velocity **is** decreasing. **The *step*** that he do were, that he **add and** then multiply **the** numbers **of the** answer that he got for **potential energy and kinetic energy**.

Grade 9, Level 3
9_181
92 words, 52.17% language from task input

When the toy car goes down **the** hill its **energy** goes **up**. **The energy** goes **up** because **it's** going down **the same** thing. So gravity pushes down on **the car** making **it** go faster. **When it** gets **to the** bottom **it** starts **to** slow down. **It** slows down because **it's** not going down **the** hill anymore. **The energy** goes down till **it** gets **to a** stop. In ***step*** 1 **the energy is 20.** And then **the energy** goes down **10** then **it** goes **to 0**. That's **what happens to the toy car.**

Grade 9, Level 4
9_235
119 words, 64.71% language from task input

The toy car's energy will have **the same amount of energy**, if we **add up the potential and kinetic energy**. We know that ***adding* the potential energy and the kinetic energy**, we'll get **the total energy**. So **the steps** that **Omri *uses* is**, since he knows that **potential energy is 10**. He then multiply **it** by "**x**" because we don't know **the kinetic energy**. **Omri** knows that **the total amount is 20**. So he ***uses* the equation 10 + x = 20**. He then simplify **it to** get "**x**" on one side. After solving, he got **x = 10**. So he **add the potential energ**y (10) by **kinetic energy (10) and** get **20** for **the total amount of energy**.

Grade 9, Level 5
9_322
143 words, 68.53% language from task input

A **toy car is an object** that **has potential and kinetic energy and** that **energy will always add up to the same amount of energy**. **At the** beginning **of the** process in which **the car** starts **to** run **the potential energy of the car** equals **20 and the kinetic energy** equals **10**. ***Adding*** both ***energies*** would give you **the total energy of 20**. **At point B the potential energy** equals **10 and the kinetic energy is** unknown which **wil**l be represented with **the** variable **x**. You then place **the** number **and** variable to form **a** mathematic problem **to** find **out what the total energy** would be. **When** doing this **at the** end **the** results would be **10**. **At the** last **point, point C, the potential energy is 0 and the kinetic energy** equals **20** so it ***adds* up to a *total* of 20**.

Figure 9. Typical grade 9 responses by score level

While each score level contains a variety of responses and response types, Figure 9 provides examples of typical features at each level. Response length increased at each score level as did the percentage of language from the task input. Responses at higher score levels more frequently incorporated strings of copied or revised text into their responses, integrating them with original words and structures.

4.2.3.4 Qualitative coding

This section describes the results of grade 9 qualitative coding by score level. The prompt for the grade 9 task asked students to “describe what happens to the toy car’s energy and explain the steps Omri used to calculate the kinetic energy of the toy car at Point B.” Table 38 lists the coding categories that were applied to grade 9 responses. A total of 20% of all responses were included in the qualitative coding (or 20 responses per score level).

Table 38. Grade 9 qualitative coding categories

Coding category	Description
Addresses prompt	Responses were coded for the extent to which they addressed the prompt: does not address prompt, partially addresses prompt, or fully addresses prompt. Entirely off-topic responses were also identified. These responses focused on a related topic but were not directly relevant to the prompt.
Misunderstands input	Evidence that the student misunderstood some part of the task input or the concepts presented in the task. Some responses were coded as partial understanding, meaning that they reflected an accurate understanding of part of the input but did not meaningfully engage with all task input.
Prompt rephrase	Responses coded for whether or not they included a rephrasing or reformulation of the task prompt.
Input links	Responses coded for links to the task input (either through exact copying, minimal revisions, or more extensive rephrasing). Each instance was linked back to a particular part of the task input.
Use of background knowledge or language	Instances when responses incorporated relevant or irrelevant details or language beyond what was provided in the task prompt.
Graphic description	Identified responses which responded to the prompt by describing the task graphic.
Use of math equations	Categorized each response based on whether or not it incorporated information from the math equations included in the task input.

The description of responses at each level addresses these coding categories along with any distinctive features of the level noted during the coding process. Response excerpts throughout this section include a notation of the response identification number.

Level 2 responses were typically brief and exhibited a misunderstanding or partial misunderstanding of the task input. Twelve grade 9 responses were coded as off-topic and eight as partially addressing the prompt. Of the 20 texts at this score level, 15 showed at least a partial misunderstanding of the task input and three texts were coded as not engaged with the task input, which means that they did not address the topic sufficiently to evaluate the writer's understanding of what was presented in the task. The following response exemplifies an incomplete understanding of the task input:

1. The calculation Omri used was that at point A it has less energy than point B because the car is not moving. Therefore point B has more energy because it is going downhill. (9_050)

This response did not reflect the concepts presented in the task input and instead viewed the changes as the car moved as a loss of energy rather than energy transfer. While partial misunderstanding of task input typified responses at score level 2, it is difficult to ascertain whether writers lacked the language to comprehend the task input or if they lacked the language to express their understanding in writing. Writers at this level often provided a response that only partially addressed the task prompt or was off-topic, meaning the response was related to the task input but did not directly address the prompt. For example, in in the following excerpt the writer describes the car's movement:

2. At point A the toy car is standing still. At the point B the car is going down. (9_015)

I coded this response as both off-topic and as a graphic description response because the writer described the task graphic rather than addressing the prompt. Ten level 2 students incorporated similar graphic description or narration into their response and this was often done in lieu of responding to the task prompt. For low proficiency writers, this may have been a strategy for providing a relevant response when they did not yet have the writing proficiency to sufficiently understand and respond to the prompt.

In other cases, writers at level 2 incorporated background knowledge and language that, while related to the topic, was not directly relevant to the prompt. Fourteen responses were coded for instances of background language. The topics they wrote about were frequently related to the speed of the car, which was not referenced in the task input. For example, this response describes the car's speed and force:

3. As the car is at the top of the toy ramp he goes down very slow at first and the energy gets higher and the car goes faster than what it went before. And all that force adds up to make the car go faster. (9_098)

This is an example of a response that engaged with the task input and brought in background language but did not directly address the prompt.

Six level 2 responses incorporated prompt rephrasing. The two examples below are typical of level 2:

4. **The toy car's energy** increases because it's going down the hill. (9_031)
5. **What happened to the toy car** that the car got faster by going down the little ramp […] (9_081)

In these examples, students incorporated brief chunks of language from the task prompt into the introductions to their responses. Other examples at level 2 are similar in that they used short strings of language from the prompt. At level 2 the links to language in the task input were limited, and these responses did not include extensive rephrasing.

At level 3, the understanding of and interaction with the task prompt is noticeable when compared with level 2 responses. Three responses were coded as off-topic, fourteen as partially addressing the task prompt, and three as fully addressing the prompt. In general, responses were relevant to the task but not complete. Responses often recounted math calculations from the task graphic. While math equations and numerical data were rarely incorporated into level 2 responses, they are a noticeable feature at level 3. This information was often presented using short sentences that were strongly related to the task graphic. That is, responses incorporated minimal original language to transform math equations and graphic labels into sentences. For example:

6. Point A the potential energy is 20. The kinetic energy is 0 so the total energy is 20. (9_165).
7. The potential of the car is 10. And the kinetic is 10. And the total equals to 20. (9_172)
8. When the car was at point A the car had no kinetic energy and the potential energy and the total energy was 20. And when the car went to point B it was rolling down a hill. And the potential energy changed to 10. (9_180)

In these three response excerpts, language from the graphic labels, which listed the amount of kinetic, potential, and total energy at each point, was incorporated into original sentences. The response information and order of presentation seemed to drive the structure of the responses.

Six level 3 responses were coded as including prompt rephrasing. The following examples are typical of level 3:

9. **The steps that Omri used** was 10+x=20. (9_170)
10. **What happens to the energy of the car** is that the first it has his own energy. (9_197)

The prompt included two distinct parts (what happens to energy and steps in the calculation) and all prompt rephrasing at level 3 focused on one but not both components of the prompt. This reflects the coding of the majority of level 3 responses as partially addressing the prompt. In addition to prompt rephrasing, seven responses were coded as having links to other input language, typically, to graphic labels and math equations rather than to the input text at the beginning of the task that describes the law of the conservation of energy. At level 3, responses tended to be concrete and describe what happened to the car or the energy, but did not address the underlying reasons for these changes, such as energy transfer, which was discussed in the input text. The responses relied heavily on the task graphic.

Responses at level 4 typically displayed a strong understanding of the task input. Three responses were coded as partially understanding input and the rest indicated no difficulty with comprehension of task input. Thirteen responses were coded as fully addressing the task prompt, meaning that they described changes in energy and recounted the steps to calculate the car's kinetic energy. Seven responses were coded as partially addressing the task.

At level 4, responses often incorporated language from the task graphic, but this language was more varied and included more original language and rephrasing than at level 3. For example:

11. At the point B, the car is going down of the hill, so it is starting to transfer the potential energy into kinetic energy. The picture shows that the car has 10 potential energy and 10 kinetic energy but the total energy is still the same.

In this excerpt, the response is closely linked to information in the task graphic, but also incorporated original language ("down the hill," "still the same") and integrated concepts about energy transfer from the task introduction text. Responses at level 4 tended to synthesize information from the input at a higher rate than level 3 responses. There were some awkward phrasings and grammatical infelicities in the integration of input language (e.g., "the car has 10 potential energy"), but the meaning is clear. Level 4 responses used words and phrases from the task input extensively, although the use may have not flowed naturally if the vocabulary was unfamiliar to the students.

A total of 18 level 4 responses had links to task input language, which included the extensive use of math equations and numerical information from the task graphic as well as more extensive rephrasing of task input. Eight texts used language from the “Law of Conservation of Energy” portion of the input, which was rarely used in level 3 responses. Eight level 4 responses included prompt rephrasing. These two examples are typical of level 4:

12. That's how he **calculated the kinetic energy of the toy car at point B** (9_283)
13. **The calculation Omri used** to solve point B […] (9_298)

At level 4, students tended to borrow longer phrases from the task prompt. Rephrasing the task prompt seems to be a more consistent part of responses in grade 9 than in other grade levels. As at other score levels, prompt rephrasing was used in both response introductions and conclusions, and students tended to rephrase only one of the two components of the prompt.

At level 5, all responses demonstrated a complete understanding of the task input and 19 of the 20 were coded as completely addressing the task prompt. Eighteen of the responses included links to the task input, and responses frequently were able to integrate information from different components of the task input into their response. The following example is typical of level 5:

14. At point A the toy car had no kinetic energy and its potential energy was 20. When the toy car was pushed down from the top of the hill, all the stored energy was turning into kinetic energy, When the toy car reaches point B, its potential energy was 20. (9_347)

This response is closely linked to the task graphic and integrated input language into original constructions in a way that sounds natural. While level 4 constructions at times sounded awkward, level 5 responses were able to seamlessly integrate input language. In addition, the original vocabulary was noticeably more sophisticated than at lower proficiency levels. Typically a few words or phrases set a level 5 response apart. Examples of original language use at level 5 include:

15. The last thing he did was simplify and he got x=10 so that the kinetic energy at point B was 10.
16. First he placed the toy car on an elevated area of the hill and then he calculated that the potential energy of the toy car is 10. (9_386)

In these examples, language use like “simplify” and “elevated area” are notable uses of original language, and it is these types of words and phrases that tended to distinguish level 5 texts. The specific original language used varied markedly between texts, but in general most level 5 texts did have this type of original language as a defining feature.

Half of the level 5 responses incorporated prompt rephrasing. The following examples are typical of level 5:

17. That's how **Omri calculated the kinetic energy of the toy car at point B**. (9_363)
18. **Omri calculated the kinetic energy of the toy car at point B** by steps below. (9_376)

As at other levels, prompt rephrasing was used in response introductions and conclusions. At level 5 all instances of prompt rephrasing were linked to the part of the prompt about calculating the car’s kinetic energy.

4.2.3.5 Summary of grade 9 qualitative coding

The results of qualitative coding showed clear patterns by score level. Table 39 summarizes the key characteristics of grade 9 responses by score level in terms of understanding of task input, extent to which responses address the task prompt, and response characteristics.

Table 39. Summary of grade 9 responses by score level

Level	Summary of Key characteristics	
	Understanding of task input and extent to which responses address the prompt	Response characteristics and use of language from the task input
Level 2	Short responses that often reflected a partial misunderstanding of the task input. Responses did not typically respond completely to the task prompt or may have been off-topic.	Responses may have relied on a description of the graphic or may have incorporated topic-related background language that was not directly relevant to the prompt. Level 2 responses did not typically incorporate math equations or a description of mathematical procedures.
Level 3	Level 3 responses typically partially addressed the task prompt and may have reflected a partial understanding of the task input, or a misunderstanding of key information.	Responses frequently incorporated math equations or numerical data from the task graphic as part of addressing the prompt. In many cases, the responses followed the graphic closely and adapted graphic labels into sentences using repetitive sentence structures with minimal original language.
Level 4	Responses typically fully addressed both parts of the prompt, although this was not always the case.	Responses typically included math equations or numerical information from the task graphic, and this was often synthesized or integrated into original language structures. The use of original words and structures was varied rather than repetitive and included details and elaboration.
Level 5	Level 5 responses typically reflected a complete understanding of task input and fully addressed both components of the task prompt.	Responses seamlessly integrated language from the input with original language and structures. Most level 5 responses included a few instances of noticeably sophisticated or high-level vocabulary used to provide relevant detail and elaboration when responding to the prompt.

One distinction in the grade 9 responses was between level 2 and all higher-level responses. At level 2, responses rarely incorporated math equations or numerical information from the graphic labels, but this was a frequent feature at all other levels. The level of integration with input and original language was also a noticeable difference. Level 3 responses tended to use simple and repetitive original language. Level 4 responses were often characterized by some degree of awkwardness, while level 5 responses were typically seamless and natural when incorporating input language with original language. Prompt rephrasing occurred more frequently in grade 9 responses than in grade 3 or grade 6. This may be a test-taking strategy that students learn in higher grade levels.

4.2.3.6 Summary of grade 9 results for research question 2

This section presented results for grade 9 using word frequency data, phrase-level analysis, and qualitative coding. Taken together, these results show distinct patterns of input language use by score level and also demonstrate a progression in terms of how students understand task input and the task prompt. As with other grade levels, word use data shows patterns by score level. Phrase-level results show that at higher proficiency levels, most students incorporated short phrases of input language into their responses but that extensive copying of task input was not widespread. The grade 9 task was distinct in that it included math equations in the task input. These were used more frequently by students at higher scoring levels, and a qualitative analysis of responses by score level shows that use of math equations was related to how completely responses fulfilled task demands.

4.3 Results: Research question 3

The third research question asks:

3. Are there different patterns of task input use by grade level?

Because students in each grade-level completed a different task, data to this question is limited to a comparative discussion of descriptive statistics and qualitative differences which emerged between grade levels.

4.3.1 Differences in response features

This section presents data used to address research questions 1 and 2 (which appears elsewhere in this report) in a way that allows for comparisons between grade levels. As noted in the results for research question 1, there was a clear difference in response length from grade 3 to grade 9, with grade 3 texts having the shortest overall mean (*M* = 64.90, *SD* = 27.31). The mean text length for grade 6 (*M* = 99.40, *SD* = 35.59) and grade 9 (*M* = 97.75, *SD* = 45.16) were similar. However, because the task directions for each grade level specified different response lengths, these differences suggest but cannot be convincingly argued to support fluency differences between grade level.

Word-level coding did not reveal any clear differences between grade levels. The relative percentage of language use from the task input was fairly consistent; for each grade, the use of content words from the task input ranged from 16 to 30%, with word use being the lowest at level 2 for each grade level. Statistics from phrase-level coding, however, did indicate different patterns of textual appropriation by grade level and for grade 3 students in particular. Table 40 shows the total number of exactly copied ([EC]) and minimally revised ([MINR]) strings for each grade level and the total number of responses which contained at least one borrowed string.

Table 40. Responses containing borrowed strings by grade level

	Level 2	Level 3	Level 4	Level 5	Total
Grade 3 (*n* = 400)					
Total N	32	26	109	105	272
Per 100 words	0.84 (1.87)	0.52 (1.47)	1.46 (1.68)	1.27 (1.44)	1.02 (1.67)
Percentage of texts	20.00	19.00	57.00	58.00	38.50
Grade 6 (*n* = 400)					
Total N	105	104	220	263	692
Per 100 words	1.72 (2.00)	1.29 (1.73)	2.07 (1.51)	2.01 (1.32)	1.78 (1.69)
Percentage of texts	58.00	53.00	84.00	91.00	71.50
Grade 9 (*n* = 400)					
Total N	27	76	212	305	620
Per 100 words	0.58 (1.78)	0.85 (1.29)	1.90 (1.52)	2.18 (1.50)	1.38 (1.68)
Percentage of texts	13.00	43.00	86.00	92.00	58.50

One notable difference is that the percentage of grade 6 level 2 and level 3 responses containing at least one instance of borrowing was higher than the same levels in grades 3 or 9. A review of the responses shows that these instances were mostly short chunks of language from a chart in the task input that contained information about the amount of water and the temperature needed to grow tomatoes. This phenomenon seems to be due to a feature of the task input that encouraged borrowing at all levels rather than a distinction based on grade level.

Borrowing was relatively infrequent in grades 3 and 9 at level 2. Beyond level 2, the grade 9 responses followed a similar pattern to grade 6. Data about the mean length of borrowed strings indicate further patterns by grade level. Table 41 presents the mean length of borrowed string for each grade level. This data is the mean for exactly copied and minimally revised strings combined.

Table 41. Mean length of borrowed string by grade level

	Level 2	Level 3	Level 4	Level 5	Total
Grade 3					
Mean (SD)	11.81 (3.91)	9.88 (5.06)	8.26 (3.85)	8.13 (3.80)	8.78 (4.15)
Min	5	5	4	4	4
Max	22	25	22	25	25
Grade 6					
Mean (SD)	6.75 (3.65)	6.80 (3.16)	6.57 (2.55)	6.60 (2.80)	6.65 (2.93)
Min	4	4	4	4	4
Max	25	21	17	20	25
Grade 9					
Mean (SD)	6.33 (3.38)	6.42 (3.11)	6.83 (3.21)	7.19 (3.86)	6.94 (3.46)
Min	4	4	4	4	4
Max	17	18	18	21	21

The mean length of string for grade 3, level 2 responses was much longer than other grade levels. While this may be due in part to the presence of sentence-length labels in the task graphic (rather than shorter labels), each task did include longer portions of text, so the opportunity for more extensive borrowing was presented at each grade level. A review of level 2 responses across grade levels shows that grade 3 responses distinctly tended to borrow complete sentences without integration into original text. The following grade 3, level 2 response illustrates this (minimally revised strings are underlined with a dotted line):

> Electricity can not flow through an incomplete circuit because its path is broken. When all parts in a circuit are connected it is a complete circuit. When a path is broken the lightbulb B doesn’t work. (3_093).

This example shows that the student borrowed two longer strings of text from the task graphic as complete sentences with minimal revision. This response did not directly address the task prompt (“describe how solving the problem with lightbulb B will change the flow of electricity”), but did show some awareness of it, as demonstrated by mentioning lightbulb B. In this case, the response consists mostly of language borrowed from the task input; this was used to address the prompt, and at a global level, to construct a response relevant to the prompt. Borrowing behavior can be viewed as a sign of emerging language proficiency. The borrowing of entire sentences as either exact copies or minimally revised strings does seem to be a distinct feature of low proficiency responses at grade 3. While these types of responses did occur at other grade levels, they do not typify the level in the way that they do for grade 3. In general, low proficiency writers in grade 3 who borrowed language from the task input used entire sentences rather than integrating task language into their own writing

The grade 3 example can be compared with a grade 6, level 2 response that also contains multiple borrowed strings of minimally revised text:

> Put the seeds in the holes. Make sure the temperature to be 50 degrees or higher. You have to put 15 - 20 centimeters of water. And now watch the plant do its job. (6_034)

An example grade 9, level 2 response shows a similar integration of short chunks of input language into original language:

> What happened to the toy car that the car got faster by going down the little ramp using the potential energy and kinetic energy. (9_081)

While this study cannot provide data about whether or not these patterns are due to developmental differences or to task differences, they are suggestive of differences and point to the need for future research.

4.3.2 Prompt rephrasing across grade levels

Across all grade levels, the qualitative coding included a code for prompt rephrasing. For this code, I marked all instances where a response contained a rephrasing of the task prompt. This could have occurred as exactly copied or minimally revised text, or as more extensive rephrasing not captured by phrase-level coding. This is one area where qualitative coding seemed to indicate different patterns by grade level. Table 42 summarizes the number of responses by score level which were coded as containing prompt rephrasing. These totals are out of 20 responses coded for each grade and score level.

Table 42. Responses containing prompt rephrasing by grade and score level

	Level 2	Level 3	Level 4	Level 5
Grade 3	3	0	1	4
Grade 6	1	3	3	6
Grade 9	6	6	8	10

Instances of prompt rephrasing were more frequent with each subsequent grade cluster, and half of the level 5 grade 9 responses included this feature. Prompt rephrasing can be interpreted as representing an awareness of the testing context and as a test-taking skill. It is unsurprising that older students, who presumably have had more experience with writing assessment in school and may also have received more instruction in test-taking skills, would use this strategy more frequently than younger students. As a rhetorical move, prompt rephrasing signals that a response directly addresses the demands of a task. The qualitative review also showed that at higher levels in each grades, students tended to fully rather than partially address the task prompt. The higher frequency of prompt rephrasing at level 5 corroborates this finding. While students who rephrased the task prompt did not necessarily address the task demands completely, they did show an awareness of what the task asked them to do and attempted to frame their response in relation to these task demands.

4.3.3 Summary of results for research question 3

The final research question looked at patterns of difference across grade levels. The data for the current study is limited in that it does not allow for direct comparison of data. Through analysis of data for research questions 1 and 2, two areas emerged as demonstrating patterns of difference by grade level. First, phrase-level coding showed that low proficiency grade 3 students tended to borrow longer strings of text when compared with grade 6 and 9 students. Second, grade 9 responses contained more instances of prompt rephrasing than lower grade levels, particularly at the higher proficiency levels. While these results should be interpreted with caution in terms of developmental differences between grade levels, they are suggestive of distinctions and provide some preliminary, exploratory information about developmental trajectories in writing.

4.4 Summary of results

This chapter presented results from the study by research question. As a whole, the results of this study point to the extent to which input-rich tasks constrain the language of student responses. Student writing is directly related to the language provided by the task, and writers at different proficiency levels took up and used this language in different ways. The consistency of findings across grade levels, particularly related to overall percentages of input language use in responses, suggests that input-rich tasks do represent a stable task type and that features of this task type relate systematically to response features. The next chapter discusses the implications of these findings for assessment research and practice.

5 DISCUSSION

This chapter discusses implications from the study. The discussion is organized by research question, followed by a discussion of general implications for assessment research and practice.

5.1 Discussion by research question

5.1.1 Research question 1

The first research questioned focused on the extent to which student use language from task input in their responses. The results show that across grade levels and score bands, approximately 50-70% of language in student responses came from the task input. Because task responses were relatively short, the use of original language was at times limited to only a few words.

The results show that the language provided in the task input was foundational to test takers responses, shaping much of what they write. As noted in the results section, the relative stability of word-level coding results suggests that input-rich tasks do represent a stable task type with consistent characteristics that shape responses in specific ways. This finding also provides evidence for the claim that input-rich tasks are able to assess academic language as a distinct from background knowledge, because the use of original language in responses in limited. This means that, as expected, students engaged with the language and content presented in the task input in their responses. According to Bachman and Palmer's (1996) framework of task characteristics, these results indicate a direct relationship between task input and response.

Results for research question 1 also included percentage of input language use according to whether it was directly copied or if words had been modified in any way (e.g., conjugation of a verb). Across all grades and score levels, the percentage of modified words as a proportion of the total responses ranged from averages of 1-4%. Although this seems to be an indication that students do not transform input words, it may instead be due to the kind of content words (e.g. “lightbulb,” “tomato”) provided in the task input. In other words, the small amount of modified words in responses may be a function of lack of opportunity rather than evidence of how writers engaged with task language. One consideration for task developers is to systematically include words that writers can manipulate and modify, since an ability to do so may distinguish writers at different score levels and to encourage creativity within the limited bounds of constrained task types such as these.

Research suggests that writing-only task types can place creativity demands on students (Read, 1990; Plakans, 2008). The results of this study show that input-rich tasks likely do not place these demands on test takers, as the language they produce is mostly limited to the task input. While a process-based approach to researching input-rich tasks is needed in order to understand the cognitive demands placed on test takers, these results do provide a starting point.

One important note deriving from research question 1 relates to scoring scale design. Given the extent to which the language of the task input shaped student responses, it would be best to create scoring scales that differentiate levels based on features of how students use this language. The WIDA writing rubric used in this study, which has since been updated as part of the new, computer-based writing test, does not differentiate score levels based on input language use beyond levels 1 and 2. The WIDA writing rubric is publicly available on the WIDA website (“Writing rubric of the WIDA Consortium, n.d.). The score level 1 descriptor for linguistic complexity states that, “varying amounts of text may be copied or adapted; adapted text contains original language.” The level 2 descriptor notes for Linguistic Complexity notes that, “varying amount of text may be copied or adapted" but copied or adapted text is not mentioned beyond level 2. The implication is that students at higher score levels use original language to responds to the tasks. The results of word-level coding indicate that lower proficiency writers use less language from the task input than writers at higher score levels. Thus, these scoring criteria do not seem to reflect how students use input language in their responses across score levels.

5.1.2 Research question 2

The results of research question 1 make it clear that the language of the task input is fundamental to how students respond to test tasks. Research question 2 focused on qualitative differences between different score levels. Data included information about which words are used at each score level and information about the use of borrowed strings from the task input.

The analysis of content words used by score level showed that particular words clearly distinguish between score levels; very few low-scoring responses used the word, while in some cases, almost all high-scoring students used the word. I listed the content words for each task based on COCA frequency data, but word use did not follow a clear pattern based on this categorization. Rather, word use patterns seemed to indicate which words from the task input were most central to task completion. For example, in the grade 9 task about growing tomatoes, over 90% of students at score levels 4 and 5 used the input words "yardstick," "thermometer" and "gauge" to respond to the prompt about the use of tools to grow tomatoes. At level 2, a mere 2-3% of responses used these words. The results suggest that task essentialness accounts for the use of different content words, and that higher-level students more frequently use the words most important to successful task completion.

This finding is supported by results from the qualitative coding, which showed that students at lower score levels often partially rather than fully addressed the task prompt, or provided off-task responses. Students at lower score levels tended to write generally about growing tomatoes but did not engage with the concept of using tools. In this example, the task was accessible to students at different score levels but task completion differentiated students. At the core of this difference is the cognitive task students were asked to engage with. To successfully complete the task, students had to apply and integrate language from two different charts presented in the task input. One important issue for test developers is to systematically review the cognitive functions that relate task input to expected responses via the task prompt. For example, test takers may be asked to synthesize, paraphrase, integrate, or evaluate information in the input in order to formulate a response. There may be a minimum language proficiency threshold necessary to engage with higher-order thinking skills demanded in tasks, beyond merely recounting task input.

Both the word-level data and qualitative coding seem to suggest a proficiency threshold for engaging fully with the task and understanding task input. Reading ability and input processing may be a relevant skill for the writing construct, and more research is needed here. Findings from studies related to integrated tasks are mixed. One study found that reading ability was not correlated with task scores (Grabe, 2003) while other process-based studies show that reading ability is a relevant factor in student performance (Plakans, 2009; Weigle, Yang & Montee, 2013). Because these studies were all conducted with university students, a basic level of first language literacy and English reading ability can be assumed. ELL students often have varying levels of first language literacy and in general have lower levels of reading development in English than participants in university studies. It would be worth exploring the role of reading ability in how students respond to input-rich tasks both in terms of general proficiency levels as well as cognitive processing perspectives.

Related to use of borrowed strings, the study shows that extensive use of language from the task input is not a major concern for input-rich tasks. Borrowed strings of language present a potential threat to scoring in that extensive use of source language may artificially inflate students' writing levels. In integrated tasks, concerns have been raised about how raters view source use and how this affects scoring processes and results (Weigle, Yang, & Montee, 2013; Gebril & Plakans, 2014; Cumming et al., 2001). Concerns include whether raters notice source borrowing and how their perceptions of acceptable and unacceptable source language use affect scoring decisions. Similar concerns apply to input-rich tasks, again with the caveat that the input-rich construct of writing is distinct from that of integrated tasks. While citation and quotation practices are not expected in this task type, it is possible that extensive use of source language could affect performance and scoring. While there were some cases in the data of extensively copied responses, this was rare and occurred most often at lower proficiency levels. Given the relatively brief input provided in the tasks, it is likely that raters could easily become familiar with input and consequently identify instances of input in responses. However, rater training materials may need to sensitize raters to this issue and provide explicit guidance about how to treat instances of task input use. Questions about whether raters should focus on scoring original language or focus on the integration of input language are an interesting issue for testing practice.

The use of both vocabulary and strings of phrases from the task input provide further indications about how students may process tasks while responding. As noted in the discussion for research question 1, input-rich tasks likely place minimal creativity demands on students. The integration of content language into original writing suggests that test takers are going back and forth from the task input to their writing during the composing process, and relying on this information to structure their responses. This finding reflects similar results from research on integrated tasks (Plakans, 2008, 2009; Weigle & Parker, 2012). For integrated task types, the use of source language provides evidence for skills integration, which is part of the task construct. For input-rich tasks, the construct is different, but similar skills and processes may be at work. Later in this chapter, I discuss implications for understanding the underlying writing construct for input-rich tasks. For example, information processing and uptake seem to be relevant skills.

One particularly interesting result from the research question 2 data relates to the use of math equations in student writing. Only the grade 9 task included this type of input, and the results showed that these were frequently used by higher-level writers who both rephrased these equations as prose and also integrated the equations into the text of their writing. Lower-proficiency writers tended not to use math equations from the task input in their writing. As language-based approaches to math become more widespread under the Common Core, it may be worth exploring how the domain of writing is addressed in math classes. How do students write about math in school? Do students regularly write about math equations, or are there other writing tasks that would better reflect authentic, classroom-based language use in this content area? Writing about math equations and processes may be a learned skill that students acquire through content instruction, and the patterns of use by score level suggest that this type of writing is unfamiliar to lower proficiency students.

5.1.3 Research question 3

The third research question compared patterns of task input use by grade-level cluster. One important limitation of the study is that students in each grade level completed different tasks. Thus, direct comparisons between grade levels are not possible and differences in performances may be either due to task features or due to developmental difference. However, data from the textual analysis combined with patterns in the qualitative coding phase did reveal patterns of difference that suggest developmental differences across the grade levels included in this study.

The clearest pattern of difference by grade level relates to the length of borrowed strings. In grade 3, the mean length of borrowed strings (for both exactly copied and minimally revised strings) was about 12 words at level 2. At grades 6 and 9, the mean string length was about six or seven words. Grade 3 students borrowed longer strings of text in generally with a clear pattern of extensive borrowing at the lowest score level for some papers. While the practice of appropriating large chunks of task input was not a particularly widespread feature across all responses, it did characterize a portion of low-proficiency responses in grade 3. There may be a developmental trajectory related to how students acquire skills related to integrating external sources and texts into their own writing. Source-based writing is a key academic skill, and one that becomes increasingly important as students progress from elementary schools to upper grades. It is also a key skill in the Common Core, and one area of potential research is in source-based writing practices and expectations for K-12 contexts. As noted in the literature review, these issues have been extensively researched and theorized in university contexts both within the U.S. and internationally, but limited work in applied linguistics has focused on younger students. Throughout this paper I have reiterated that integrated and input-rich tasks have different assessment purposes; however, there does seem to be an overlap in skills in terms of the tasks asking students to process and incorporate information into their responses.

Prompt rephrasing emerged from the data as an area of difference between grade levels. Previous research has shown prompt rephrasing is a typical feature of textual borrowing in response to integrated tasks (Weigle & Parker, 2012). While prompt rephrasing was not a particularly widespread test-taking strategy for students in this study, it did appear in grade 9 at higher proficiency levels. Grade 3 students did not use this strategy frequently. This may reflect different levels of awareness of the testing situation. Students in upper grade levels may have been more aware of the implied audience for their responses (test raters) and have had ideas about the ways their writing would be evaluated. Older students may also have had more familiarity with prompt rephrasing as a test-taking strategy based on experience with assessment.

In addition to its role as a test-taking strategy, prompt rephrasing can be interpreted an indication of test takers' task representation. Wolfersberger (2013) describes task representation as the mental conceptualization that students create of what they are supposed to do when responding to a task. In assessment design, it is important to align a task's intended demands with how students understand these task demands because the students' understanding of the task will mediate how they respond. In rephrasing the prompt, students indicated that they have a clear understanding of what they were being asked and how they should respond. This is another area where a process-based approach would be useful in extending these exploratory findings. Future research may consider including observational data and stimulated recall sessions to look at task representation.

5.1.4 Limitations

There are several important limitations to this work. First, as an exploratory study, the results seek to characterize responses to input-rich tasks generally. However, task characteristics are complex and dynamic. While I argue that the response features are consistent enough across grade clusters to support the conceptualization of input-rich tasks as a task type, it is not clear in this analysis how more fine-grained aspects of task features relate to responses. Studies which systematically vary task characteristics are needed in order to develop a deep understanding of this issue. The results of this study are suggestive and provide directions for future work.

Second, this study used operational test data with limited background information about students. While the goal of task design is to eliminate the need for background knowledge, the study did not include any information about time in school, general proficiency level in English, levels of academic achievement, or other student-level data that would be useful in understanding how test-taker characteristics and background knowledge affect performance.

A third limitation is that students across grade levels did not complete the same task. Thus, it is difficult to tease apart which differences may be due to cognitive development and which are related to task features. The results for research question 3 are limited and descriptive in nature.

5.2 Implications for research and practice

Taken as a whole, the results of the study have several implications for assessment research and practice. Reviewing the results of the study in light of Bachman and Palmer’s (1996) framework of task characteristics, I propose that input-rich tasks and responses to these tasks are characterized by the features summarized in Table 43.

Table 43. Proposed features of input-rich tasks and responses

Category	Features
Task input characteristics	• Linguistic input provides necessary background information • Graphic representations fundament to input • Graphic representation used to minimize linguistic input • Highly structured presentation of information • Topical vocabulary • Constrained prompts
Response characteristics	• 50% or more of vocabulary comes from the task input • Integration of vocabulary into original language structures and sentences • Appropriation of longer strings of input text relatively infrequent • Response structures reflects structure of input presentation • Proficiency threshold for fully comprehending input and responding to prompts
Relationship between the task and responses	• Input requires the integration of information from the task input to accomplish a cognitive task beyond merely rephrasing the input

These characteristics emerged from the analysis of response features, which showed clear patterns across grade levels in terms of how test takers engage with task input in their responses. A transparent construct of input-rich task should include a systematic framework for identifying the ways in which test takers interact with task input. Beyond this, the design of input-rich writing tasks should seek to standardize the parameters of language used in the task input and to manipulate this language based on desired response features. In previous phases of test development, task input was optimized for student understanding and to reflect the content demands of the task. However, this results of this study make it clear that the language of task input is the primary feature shaping student responses, particularly at the lexical level. Developers may consider providing shorter chunks of task input (e.g., through graphic labels or in bulleted lists) that would allow test takers to adapt and modify language in their responses.

While I am hesitant to draw direct implications from this study to classroom practice, there are some useful connections from this research to bigger issues of how language is assessed and how information from the test is used to make decisions about students. First, it is important to note that test tasks sample language from the target language use domain but are also limited in the scope and variety they can assess. Classroom writing often includes process-based approaches and interactive activities whereas test-based writing is a product-focused activity. Critics are right to point out the potentially negative and limiting effects of standardized testing on classroom practices. It is my hope that a better understanding of the types of language assessment tasks elicit will lead to improved decision-making based on test data as test users know the uses and limitations of test information.

The input-rich tasks analyzed in this study are different from classroom writing tasks in content areas. In those contexts, students can and should engage with content they know. Their writing demonstrates both content knowledge and language development. However, research on teacher preparation suggests that content-area teachers may not be prepared to address the language needs of ELLs, and may not receive training in academic language development (Anstrom et al., 2010). For these teachers, ACCESS for ELLs test scores, and corresponding interpretation tools, can be a useful starting point for understanding and addressing the needs of their ELL students as this information summarizes what they can do in English. In addition, while the test tasks and classroom writing tasks can and should differ in key features, the use of tasks that assess language in math and science may help support the idea that language development is embedded within these content areas, and that all educators, not just language teachers, have a role in teaching language.

Although ELD assessments play an important role in ensuring that ELLs have access to language services, large-scale standardized testing in K-12 has been highly criticized. For example, critics have argue that standardized assessments are limited in scope and do not adequately assess student learning (Jordan, Brown, & Guttiérez, 2010). The promise of standards-based reform, as Shepard (2000) has pointed out, is that "tests worth teaching to" can positively affect classroom instruction. However, the negative impact of large-scale assessment on student learning, teaching, and school culture has been widely documented and discussed (e.g., Darling-Hammond & Rustique-Forrester, 2005; Shepard, 2000; Madaus, Russell & Higgins, 2009). These criticisms often relate to content-area testing, which was instituted under NCLB and continues under ESSA. For ELLs, research has shown that standardized content achievement tests are not always good indicators of student learning because language proficiency serves as a source of construct-irrelevant variance (Abedi, 2002).

Throughout this study, I present large-scale ELD assessments as providing useful information for making decisions about student services. However, the widespread use of standardized assessment in K-12 education has been widely criticized. While ELD assessments have not been the focus of the same level of scrutiny as content tests, critiques about the negative affect of the testing movement on teaching and learning provide an important counterpoint to the argument that ELD assessments help ensure educational access and equity for ELLs. As Shepard (2000) points out, assessment practices are embedded within an overall culture of learning. She argues that accountability testing is not sufficient to diagnose individual student needs and likens their usefulness to that of a medical screening in that they have have some limited usefulness but are not sufficient to rive in-depth diagnosis and change (p. 13).

When looking at test score use, it is important that information from assessments be contextualized with information from other sources, including the observations and evaluations of classroom teachers. Additionally, assessment data should be used carefully for intended purposes, including federal reporting requirements, program-level review and evaluation, and decisions about exiting language programs. Proposals such as using test data to determine teacher compensation are a clear misuse of test scores. And while classroom teachers may find test scores useful for some aspects of instructional planning, both language and content teachers need more in-depth information that can only come from formative, classroom-based assessments focused on student learning outcomes. From this perspective, ELD test results are just one component of an overall system of effective instruction.

5.2.1 Recommendations

This section summarizes recommendations for assessment practice and outlines direction for further research. Recommendations for practice are organized by implications for the task construct, issues in task design, and scoring considerations.

In this study, I used the label “input-rich tasks” to describe the type of writing tasks used on the WIDA ACCESS writing test. Although this assessment has moved to a computer-based delivery format, the approach of using rich task input is relevant to WIDA as well as a number of ELD assessments.

Explore the role of reading comprehension in the construct: The results of this study indicate that comprehension of task input may be differentiated by proficiency level, and that responses from lower proficiency students often demonstrate a misunderstanding of task input. The task input was robust enough to sustain topically-relevant responses at all score levels, even if students did not fully understand the task input. And importantly, a misunderstanding of the task input was relatively rare at the higher score levels. Reading ability may have some role in how students respond, and the study suggests that the task design mitigates this factor by providing enough accessible input for lower proficiency students to engage with the task while still providing content that is robust and sophisticated enough to sustain higher-level responses from students who are able. Providing an hypothesis about how reading comprehension could function in the assessment is an area for further work.

Specify the role of comprehending and interpreting graphic information: Beyond reading ability, the input rich-tasks in this study required students to understand and interpret test graphics. These graphics are designed to reflect authentic academic graphical displays of information such as labeled diagrams and charts, and in order to respond to tasks students must understand this way of conveying information as an aspect of the academic genre. For students with limited schooling, this mode of presenting information may be unfamiliar. The current study provides some evidence for the efficacy of graphics-based tasks by demonstrating how these are used in student responses. For example, as a response strategy, lower-proficiency students sometimes described the task graphic. Thus, the task graphics succeeded in making the task accessible to these students and providing a way for them to respond.

Task graphics place information processing demands on test takers, and these processing demands interact with the language features of the task input to shape student responses. To date there has been limited research in the field of assessment about how graphic complexity affects task performance. Specifying the role of graphic information in tasks, and the underlying skills and experiences this modality requires from students, is a key part of refining the task construct.

Extend research about the characteristics of task input vocabulary: The finding that use of particular words from the task input differentiate students across score levels merits further exploration. A study by Crossley, Clevinger, and Kim (2014) found that in integrated TOEFL iBT speaking responses, that the repetition of words in the source text, the frequency of words in the source, and the use of words in positive connective clauses could accurately predict which words would be integrated into a test taker's response. They also found that the integration of language from the source in responses was predictive of human ratings.

In this study, I used information about word frequency as a way of looking at word characteristics, and found that this did not seem to explain word use. However, this study did not account for *how* words were used in the task input, including issues such as word repletion. Future research could leverage automated text analysis tools to look at other features of words and their context in order to build a more comprehensive understanding of how test takers use input language in their responses. As Crossley, Clevinger, and Kim's study suggests, this may be a key factor in rater judgments.

Systematically describe the cognitive function of the task and how input demands affect response features: Each of the tasks in this study asked students to go beyond merely recounting the input. Students had to perform some sort of cognitive task relating to synthesizing or applying information in order to respond appropriately to the prompt. Cognitive functions that may be relevant to input-rich tasks include analysis, synthesis, and application. Taxonomies of cognitive tasks may be helpful in systematically identifying these functions (e.g., Bloom, 1965). However, it is crucial to go beyond simply identifying cognitive functions and attempt to understand how they may affect responses. Do tasks that place higher cognitive demands on students affect their responses in particular ways? How do various components of task complexity affect response features? These issues would be a fruitful area for further research, and a potential area of cross-disciplinary work with second language acquisition. For example, Skehan (1996, 1998) and Robinson (2001, 2007) have developed task frameworks and hypothesized the role of cognitive processing in these frameworks. While this study has taken a construct-based approach to understanding task performance, insights from task-based SLA approaches could be a useful avenue of future research and do not necessarily conflict with a construct-based understanding of tasks.

Explore task-specific scoring materials: Alderson (1991) makes a distinction between scales oriented toward three stakeholder groups: test users, assessors, and test constructors. Each of these groups will have different uses for the scale. For test users, including educators and students, the scale may be used as for test score interpretation, while for raters (assessors), the scale will be used for scoring performances. Test developers (constructors) may use a scale to design or analyze test tasks and ensure that they elicit performances that are consistent with the features of the scale. Each stakeholder group has different needs and expertise that inform their understanding of a scale, and a single scale may not be sufficient to address the needs of each group. As noted earlier in this chapter, the scoring rubric used operationally for the study, which is now retired, did not account for the ways students at higher proficiency levels used language from task input. The writing rubric seems to be focused on external interpretation, or towards test users in Alderson's framework.

Scoring materials for input-rich tasks should ideally reflect a progression of proficiency based on how students use language from the task input. Even at the highest levels of writing, students in this study primarily used language from the task input. In fact, the use of original language was a potential indicator of an off-topic response and thus a failure to fulfill the task demands. Rather than raters simply looking at original language to indicate quality, it may be more useful to look at both original and varied language constructions. That is, higher proficiency writers will use task input language to formulate their own sentences and phrases and will have a greater variety of sentence types within their responses. The results of this study suggest that a scale oriented towards test raters should describe how writers typically use task input language at each level as this is both a dominant characteristic of student writing and one that can be differentiated by score level. While user-oriented scales may focus on general proficiency descriptors in order to help educators and other test users interpret results, rater-oriented scales for these tasks may require more narrow and focused descriptors. In addition, raters need explicit guidance about how to respond to input language use and borrowed strings of language when scoring. It should be clear that this language is expected in responses and, except in cases of extensive copying, should not negatively affect student scores. A scoring scale more closely tailored to student responses for this task type could address this.

Finally, the results of this study suggest that automated scoring may be useful area for further exploration. Results showing the usage of particular content words suggest that this information could be used for automated scoring purposes along with other criteria. Automated scoring is potentially controversial in practice, particularly for high-stakes testing, but this approach could be used for practice testing and diagnostic purposes.

5.3 Conclusion

The goal of this study was to explore how students at three different grade levels used language from the input when responding to input-rich test tasks. This study provides a first step in understanding the features of student responses to these tasks and how students across score levels engage with and use task input in their writing. The results raised a number of issues for future assessment research, and more work is needed to understand how students at different levels of cognitive and language development respond to writing tasks.

The limited use of extended strings of borrowed text suggest that copying language from the task is not a major threat to scoring. While more research is certainly needed, including process-based studies and in-depth analyses of response data, as a whole, the results contribute evidence for the validity of this task type by demonstrating that students across proficiency levels are able to respond within the constraints of the task input in ways that clearly differentiate ability levels.

I adopted the term "input-rich" tasks early on in the research process. While this label was useful within this study to describe the test tasks, upon further reflection and feedback this term may not be the most apt for what students are asked to do, particularly when the ACCESS for ELLs task are compared with performance tasks from content-based tests, which often include extensive and dynamic task input. A better term might be input-constrained or input-dependent tasks, although these terms may carry slightly negative connotations that I don't intend. When I selected the term input-rich tasks, I wanted to convent a positive, student-centered approach to task design which sees task input as a rich source of support for students as they demonstrate their language ability. This approach is in line with WIDA's philosophy, which I discussed in the methods section. This approach means that students at all levels should be able to show what they can do in response to the task. The task label I chose may not be exactly right, but I think the generalized task descriptions are appropriate and accurate.

One of my motivations in designing this study was to provide greater insight into how WIDA ACCESS tasks functions for the purpose of applying this information to my work. While I believe this study has use outside of the context of WIDA ACCESS, I do think the results have direct utility for the design of WIDA ACCESS task specifications. As I continue to work on the new, computer-based administration mode, I am interested in how delivery mode affects how students interact with test tasks. For example, do students use language from the task as extensively if it is not presented in front of them in a static format? How might the use of video or animation affect students' understanding of task input and their use of task language? As I have noted throughout the study, it is important to conceptualize tasks as variables in pursuit of a systematic understanding of their features. However, these variables often interact in complex and dynamic ways. This study provided a foundational understanding of how students use task input when responding to WIDA ACCESS writing tasks as well as many directions for future research.

REFERENCES

Abedi, J. (2002). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. *Educational Assessment*, 8, 231-257.

Abedi, J., Leon, S.,& Mirocha, J. (2000/2005). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data. In *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives* (CSE Tech. Rep. No. 663). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

About ELPA21: FAQs. (n.d.) Retrieved from http://www.elpa21.org/about/faqs

Anstrom, K., DiCerbo, P., Butler, F., Katz, A., Millet, J., & Rivera, C. (2010). *A review of the literature on academic English: implications for K-12 English language learners*. Arlington, VA. Retrieved from http://ceee.gwu.edu/Academic Lit Review_FINAL.pdf

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bachman, L. F. (2002). Some reflections on task-based language performance assessment, *19*(4), 453–476.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: designing and developing useful language tests*. New York: Oxford University Press.

Bakhtin, M. M. (1981). *The dialogic imagination*. Austin, TX: University of Texas Press.

Bakhtin, M. M. (1986). *Speech genres and other late essays.* (V.W. McGee, Trans.) Austin: University of Texas Press.

Bailey, A. L. (2006). Introduction: teaching and assessing students learning English in school. In A. L. Bailey (Ed.), *The language demands of school: putting academic English to the test* (pp. 1–26). New Haven: Yale University Press.

Bailey, A. L., & Wolf, M. K. (2012). The challenge of assessing language proficiency aligned to the Common Core State Standards and some possible solutions. In *Understanding Language*. Retrieved from http://ell.stanford.edu/sites/default/files/pdf/academic-papers/08-Bailey Wolf Challenges of Assessment Language Proficiency FINAL_0.pdf

Bauman, J., Boals, T., Cranley, E., Gottlieb, M., & Kenyon, D. (2007). Assessing comprehension and communication in English state to state for English language learners (ACCESS for ELLs). In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 81–92). Davis, CA: University of California, Davis School of Education.

Behizadeh, N., & Pang, M. E. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing, 29*, 25–41.

Biber, D., Conrad, S. & Leech, G. (2002). *Longman Student Grammar of Spoken and Written English.* Harlow, Essex: Longman.

Bunch, G. C. (2006). “Academic English” in the 7th grade: broadening the lens, expanding access. *Journal of English for Academic Purposes, 5*(4), 284–301.

Chamot, A. U., & O’Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.

Cheng, A. (2011). Language features as the pathways to genre: students’ attention to non-prototypical features and its implications. *Journal of Second Language Writing, 20*(1), 69–82.

Cook, G. H., Boals, T., Wilmes, C., & Santos, M. (2008). Issues in the development of Annual Measurable Achievement Objectives (AMAOs) for WIDA Consortium states. *WCER Working Paper, 2008–2*. Retrieved from http://wcer.wisc.edu/docs/working-papers/Working_Paper_No_2008_02.pdf

Council of Chief State School Officers. (2012). *Framework for English language proficiency development standards corresponding to the common core state standards and the next generation science standards*. Washington, D.C.

Crossley, S., Clevinger, A. & Kim, Y. (2014) The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency, *Language Assessment Quarterly*, 11(3), 250-270.

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: promises and perils. Language Assessment Quarterly, 10, 1–8.

Cumming, A., Kantor, R., & Powers, D.E. (2001). Scoring TOEFL essages and TOEFL 2000 prototype writing tasks: An investigation into raters decision making and development of a preliminary analytical framework (TOEFL Monograph No. 22). Princeton, NJ: Educatonal Testing Service.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for Next Generation TOEFL. *Assessing Writing, 10*, 5–43.

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14*(2), 175–187.

Darling-Hammond, L., & Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teacher quality. In J. Herman and E. Haertel (Eds.), T*he uses and misuses of data in accountability testing. The 104th Yearbook of the National Society for the Study of Education, Part II* (pp. 289-319). Malden, MA: Blackwell Publishing.

Davies, Mark. (2008-) The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Available online at http://corpus.byu.edu/coca/

Davis, M., & Morley, J. (2015). Phrasal intertextuality: the responses of academics from different disciplines to students' re-use of phrases. *Journal of Second Language Writing*, *28*, 20–35.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6, 241-252.

Enright, M. K., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., Nissa, S., Powers, D., and Schedl, M. (2008). Prototyping new assessment tasks. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson, (Eds.) (2007). *Building a validity argument for the Test of English as a Foreign Language* (1st ed.). New York: Routledge.

Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Journal*, *58*(4), 599–622.

ESSA (2015). Every Student Succeeds Act of 2015, Pub. L. No. 114-95 § 114 Stat. 1177. (2015-2016).

Fairclough, N. (1992). Intertextuality in critical discourse analysis. *Linguistics and Education, 4,* 269-293.

Fairclough, N. (2003). *Analysing discourse: textual analysis for social research*. London; New York: Routledge.

Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics, 28*, 440–465.

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73.

Gottlieb, M. (2004). *English language proficiency standards for English language learners in kindergarten through Grade 12: Framework for large-scale state and classroom assessment.* Madison, WI: WIDA Consortium.

Gottlieb, M., Cranley, M. E., & Oliver, A. (2007). U*nderstanding the WIDA English language proficiency standards: A resource guide.* Madison, WI: WIDA Consortium.

Grabe, W. (2003). Reading and writing relations: Second language perspectives on research and practice. In: B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 242-262). Cambridge: Cambridge University Press.

Haahr, M. (2006). Random.org: True random number service. Web resource, available at http://www.random.org

Hakuta, K. (2011). Educating language minority students and affirming their equal rights: research and practical perspectives. *Educational Researcher*, *40*(4), 163–174. http://doi.org/10.3102/0013189X11404943

Hu, G., & Lei, J. (2012). Investigating Chinese university students' knowledge of and attitudes toward plagiarism from an integrated perspective. *Language Learning*, *62*, 813–850.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.

Jordan, W.J., Brown, B., & Guttiérez, K. (2010). Defining equity: Multiple perspectives to analyzing the performance of diverse learners. *Review of Research in Education*, 34, 142-178.

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing*, *15*(4), 261–278.

Keck, C. (2014). Copying, paraphrasing , and academic writing development: a re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing*, *25*, 4–22.

Kenyon, D. M. (1992). *Rating scale symposium: Introductory remarks.* Language Testing Research Colloquium, Vancouver, Canada.

Kenyon, D. M., MacGregor, D., Li, D., & Cook, H. G. (2011). Issues in vertical scaling of a K-12 English language proficiency test. *Language Testing*, *28*(3), 383–400.

Klein, A. (2016, January 5). Under ESSA, states, districts to share more power. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2016/01/06/under-essa-states-districts-to-share-more.html

Kristeva, J. (1986). *The Kristeva Reader*. Oxford: Blackwell.

Kroll, B., & Reid, J. (1994). Guidelines for Designing Writing Prompts: Clarifications, Caveats, and Cautions. *Journal of Second Language Writing*, *3*(3), 231–255.

Lau v. Nichols, 414 U.S. 563. (1974).

Lemke, J. L. (1992). Intertextuality and educational research. *Linguistics and Education*, *4*, 257 267.

Lim, G. S. (2010). Investigating prompt effects in writing performance assessment. In J. S. Johnson, E. Lagergren, & I. Plough (Eds.), *Spaan fellow working papers in second or foreign language assessment* (Vol. 8, pp. 95–115). University of Michigan English Language Institute.

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: how they affect students, their parents, teachers, principals, schools and society.* Scottsdale, AZ: Information Age Publishing.

McHugh, M., & Pompa, D. (2016, January 21). Taking stock of ESSA's potential impact on immigrant and English- learner students [Webinar]. Retrieved from http://www.migrationpolicy.org/events/taking-stock-essa-potential-impact-immigrant-and-english-learner-students

McKay, P. (2000). On ESL standards for school-age learners. *Language Testing, 17*(2), 185–214.

Miller, J. & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Mission & the WIDA Story. (n.d.) Retried from https://www.wida.us/aboutus/mission.aspx

National Center for Education Statistics. (2016). Digest of Education Statistics. Retrieved March 5, 2017 from https://nces.ed.gov/programs/digest/d15/tables/dt15_204.27.asp

National Governors Association Center for Best Practices Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics*. *National Governors Association Center for Best Practices, Council of Chief State School Officers*. Washington, D.C. Retrieved from http://www.corestandards.org/

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By State*s. Washington, DC: The National Academies Press.

No Child Left Behind (NCLB) Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425. (2002).

Norris, J. M., Brown, J. D., Hudson, T. and Yoshioka, J. (1998). *Designing second language performance assessments.* (Vol. SLTCC Technical Report #18). Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Pecorari, D. (2008). *Academic writing and plagiarism: A linguistic analysis*. London: Continuum.

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. *Language Teaching, 47*, 269–302.

Pecorari, D., & Shaw, P. (2012). Types of student intertextuality and faculty attitudes. *Journal of Second Language Writing*, *21*, 149–164.

Petrić, B. (2012). Legitimate textual borrowing: direct quotation in L2 student writing. *Journal of Second Language Writing*, *21*(2), 102–117.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*(2), 111–129.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*(4), 561–587.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, *17*(1), 18–34.

Polio, C., & Shi, L. (2012). Perceptions and beliefs about textual appropriation and source use in second language writing. *Journal of Second Language Writing*, *21*(2), 95–101.

Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, *9*, 109–121.

Robinson, P. (2001). Tasks complexity, task difficulty and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. P. Garcia Mayo (Ed). I*nvestigating tasks in formal language learning* (pp. 7-27). Clevedon: Multilingual Matters.

Römhold, A., Kenyon, D., & MacGregor, D. (2011). Exploring domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency. *Language Assessment Quarterly*, *8*(3), 213–228.

Saldaña, J. (2013). *The coding manual for qualitative research.* London: Sage.

Schleppegrell, M. (2004). *The language of schooling: a functional linguistics perspective*. Mahwah, New Jersey: Lawrence Erlbaum.

Shepard, L. (2000). The role of assessment in learning culture. *Educational Researcher*, 29(7), 4 14.

Shi, L. (2004). Textual Borrowing in Second-Language Writing. *Written Communication*, *21*, 171-200.

Shi, L. (2012). Rewriting and paraphrasing source texts in second language writing. *Journal of Second Language Writing*, *21*, 134–148.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, 17, 36-62.

Skehan, P. (1998). A cognitive approach to language learning. New York: Oxford University Press.

The Every Student Succeeds Act: Explained. (2015, December 9). *Education Week*, p. 17. Retrieved from http://www.edweek.org/ew/articles/2015/12/07/the-every-student-succeeds-act-explained.html

Ujifusa, A. (2017a, February 7). House votes to overturn ESSA accountability, teacher-prep rules [Web log post]. *Education Week*. Retrieved from http://blogs.edweek.org/edweek/campaign-k-12/2017/02/house_votes_overturn_essa_accountability_teacher_rules.html

Ujifusa, A. (2017b, February 14). Uncertainties as congress takes aim at ESSA regulations. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2017/02/15/uncertainties-as-congress-takes-aim-at-essa.html?qs=essa

U.S. Department of Education. (2017). *Transitioning to the Every Student Succeeds Act (ESSA): frequently asked questions*. Retrieved from https://www2.ed.gov/policy/elsec/leg/essa/essatransitionfaqs11817.pdf

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

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, *21*(2), 118–133. http://doi.org/10.1016/j.jslw.2012.03.004

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly*, *10*(1), 28–48.

WIDA. (2012). *2012 amplification of the English language development standards, kindergarten-grade 12*. Retrieved from https://www.wida.us/get.aspx?id=540

Wolfersberger, M. (2013). Refining the construct of classroom-based writing-from-readings assessment: the role of task representation. *Language Assessment Quarterly*, *10*(1), 49–72.

Wong, A. (2015, December 9). The bloated rhetoric of No Child Left Behind’s demise. *The Atlantic*. Retrieved from http://www.theatlantic.com/education/archive/2015/12/the-bloated-rhetoric-of-no-child-left-behinds-demise/419688/

Writing rubric of the WIDA Consortium. (n.d.). Retrieved from https://www.wida.us/standards/eld.aspx

Yanosky, T., Amos, M., Cameron, C., Louguit, M., MacGregor, D., Yen, S. J., & Kenyon, D. (2013). *Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 203, 2011-2012 Administration: Annual Technical Report No. 8 Volume 1 of 3: Description, Validity, and Student Results*. Retrieved from https://www.wida.us/downloadLibrary.aspx

Yanosky, T., Chong, A., Louguit, M., Olson, E., Choi, Y., MacGregor, D.,Cameron, C., & Kenyon, D. (2012). *Annual Technical Report for ACCESS for ELLs® English Language Proficiency Test, Series 202, 2010-2011 Administration: Annual Technical Report No. 7 Volume 1 of 3: Description, Validity, and Student Results*. Retrieved from https://www.wida.us/downloadLibrary.aspx

APPENDICES

Appendix A: Transcription procedures

For each response file:

Open the .tiff image file

Open a new Microsoft word Document

> If more than one quarter of the words are illegible due to image scanning problems or student handwriting, place the image file in the “Unusable” folder. It will not be transcribed. Replace the file with a new file from the “Overage” folder.

If most of the response is legible, type the response using the following conventions:

Type the response as written.

In instances of **invented spelling or spelling errors**:

- If you are reasonably certain of the word(s), transcribe the word(s) using standard English spelling.
- If you are not certain of the word(s), use a single uppercase X to for each indecipherable word.

Note: Invented spelling is common for young learners. It is defined as the practice of spelling words using a “best guess” about the spelling based on sounds.

In cases where the letters are legible but the student has used a non-existent English word (e.g., an invented word), transcribe the word as written.

In instances where **text is illegible** because of a scanning error or student handwriting:

- If you are reasonably certain of the word(s), transcribe what it seems the student wrote.
- If you are not certain of the word(s), use a single uppercase X to for each indecipherable word.

If a writer uses **a language other than English** (typically Spanish):

- If possible, transcribe the non-English response.
- If the non-English response cannot be deciphered, use a single uppercase X to for each indecipherable word.

Punctuation

Generally maintain a writer's use of punctuation to mark sentence boundaries. If a writer uses punctuation where none is needed (e.g., periods that are not at the end of sentences), it is acceptable to remove this punctuation. The goal of transcription is to accurately transcribe the language the student produced while making adjustments that will allow for computer-based analysis.

Apostrophes can be added or deleted in order to standardize the use of language. For example, if a student writes "its" where "it's" should be used, this should be corrected in the transcription.

In some cases, students will use minimal or no punctuation. In these instances, **sentence-final punctuation should be added** using your best judgment about sentence boundaries.

Paragraph Breaks

In general, transcribe text as a **single paragraph** unless the student has clearly written a multi-paragraph response. For example, students may write several sentences and skip lines between each sentence. It is not necessary to maintain these line breaks in the transcription. In this case, transcribe the response as a single paragraph.

Appendix B: Word-level coding procedures and codes

Procedures for creating word-level codes (for each task)

1. Type the task input as a text file.
2. Import all text into Microsoft Excel. Enter each word in a separate cell.
    a. Delete any repeated words.
3. Use control+F to check the Excel document against the pdf of the task. Ensure that all task input words are listed.
4. Identify each word as an input or function word.
5. Identify the grammatical category of each word.
6. Identify all possible modifications of each word that might appear in the text.
7. Create the appropriate word-level code. Word-level codes should include the following information:
    a. IN for words copied from the input or MO for words modified from the input.
    b. C for content words or F for function words
    c. Grammatical category indicator (preceded by an underscore)

Procedures for Coding Texts (All grades)

1. Import all transcripts from .txt files into a single document in Microsoft Word. Each response should be on a single page.
2. For each response, insert the following header at the beginning of each text. This information will be used to process texts in SALT.

$ Student

+ ID:

+ Transcript:

+ Grade:

+ Gender:

+ State:

+ Task:

+ Score:

Note: The mail merge function in Word was used to automatically populate the text headers from the data spreadsheets.

3. Format texts so that each sentence begins after a line break. Begin each sentence with "S." This format is necessary for analysis in SALT.
4. Check that the data is properly formatted for analysis in SALT.
   a. Perform spell check and correct any spelling errors according to the guidelines described in the transcription conventions.
   b. Check to ensure that all sentences end in a period.
   c. Place a backslash symbol before any contractions (e.g., it's should be changed to "it/'s").
5. Open the spreadsheet listing all input words and codes for the grade level.
6. Use the find and replace function in Microsoft Word to ensure that all instances of a word are identified.
   a. Words that can have more than one grammatical category should be coded individually.

7. As a quality control measure: After all word-level codes have been identified for a grade-level, use control+F to ensure that all words have been identified and coded.

List of all codes (all grade levels)

How to treat special cases

Word Type	**How to code**
Preposition words that can also appear as adverbial participles as part of either phrasal or prepositional verbs	Coded as Input Function Words (Prepositions)
To: Can occur as a preposition or as part of infinitive marker	Coded as Input Function words; no grammatical category assigned
Wh-words can occur as several grammatical classes, including determiners, pronouns or adverbs.	Coded as Input Function, coded as “wh-words” category
It's	Coded as a single unit function word with code INF_ITS
Numerals	Code as Input Function words with “NUM” code; code lexical forms of numbers that appear in the input as Modified Function words (MOF).
Math symbols	Convert symbol to letters for the purposes of analysis in SALT. + → pls - → mns = → eqs x → vx (for variable x) Code as Math Symbol ([MS]). Do not code instances where students write out the math symbols (e.g., “plus”).

Appendix B.1 Grade 3 Codes

Category	**Word type**	**Part of speech**	**Code**
Input	Content	Adjective	[INC_ADJ]
Input	Content	Adverb	[INC_ADV]
Input	Content	Alphabetic symbol	[INC_ALS]
Input	Content	-ing participle	[INC_ING]
Input	Content	Noun	[INC_N]
Input	Content	Past participle	[INC_PP]
Input	Content	Verb	[INC_VB]
Indecipherable	n/a	n/a	[IND]
Input	Function	Article	[INF_ART]
Input	Function	Coordinating conjunction	[INF_COOR]
Input	Function	Determiner	[INF_DET]
Input	Function	“its”	[INF_ITS]
Input	Function	Modal	[INF_MOD]
Input	Function	Negative	[INF_NEG]
Input	Function	Number	[INF_NUM]
Input	Function	Possessive determiner	[INF_PD]
Input	Function	Personal pronoun	[INF_PPR]
Input	Function	Preposition	[INF_PREP]
Input	Function	Quantifier	[INF_QNT]
Input	Function	Subordinator	[INF_SUB]
Input	Function	“to”	[INF_TO]
Input	Function	Wh-word	[INF_WH]
Modified	Content	Adjective	[MOC_ADJ]
Modified	Content	Adverb	[MOC_ADV]
Modified	Content	-ing participle	[MOC_ING]
Modified	Content	Noun	[MOC_N]
Modified	Content	Past participle	[MOC_PP]
Modified	Content	Past participle (error)	[MOC_PPE]
Modified	Content	Verb	[MOC_VB]
Modified	Function	Pronoun	[MOF_PRN]

Appendix B.2 Grade 6 Codes

Category	**Word type**	**Part of speech**	**Code**
Input	Content	Adjective	[INC_ADJ]
Input	Content	Adverb	[INC_ADV]
Input	Content	-ing participle	[INC_ING]
Input	Content	Noun	[INC_N]
Input	Content	Verb	[INC_VB]
Indecipherable	n/a	n/a	[IND]
Input	Function	Article	[INF_ART]
Input	Function	Coordinating conjunction	[INF_COOR]
Input	Function	Determiner	[INF_DET]
Input	Function	“its”	[INF_ITS]
Input	Function	Modal	[INF_MOD]
Input	Function	Number	[INF_PPR]
Input	Function	Possessive determiner	[INF_PD]
Input	Function	Preposition	[INF_PREP]
Input	Function	Personal pronoun	[INF_TO]
Input	Function	Wh-word	[INF_WH]
Modified	Content	Adjective	[MOC_ADJ]
Modified	Content	Adverb	[MOC_ADV]
Modified	Content	-ing participle	[MOC_ING]
Modified	Content	Noun	[MOC_N]
Modified	Content	Past participle	[MOC_PP]
Modified	Content	Verb	[MOC_VB]
Modified	Function	Number	[MOF_NUM]
Modified	Function	Pronoun	[MOF_PRN]

Appendix B.3 Grade 9 Codes

Category	**Word type**	**Part of speech**	**Code**
Input	Content	Adjective	[INC_ADJ]
Input	Content	Adverb	[INC_ADV]
Input	Content	Alphabetic symbol	[INC_ALS]
Input	Content	Noun	[INC_N]
Input	Content	Noun with possessive marker	[INC_NPO]
Input	Content	Verb	[INC_VB]
Indecipherable	n/a	n/a	[IND]
Input	Function	Article	[INF_ART]
Input	Function	Auxiliary Verb	[INF_AUX]
Input	Function	Coordinating conjunction	[INF_COOR]
Input	Function	"its"	[INF_ITS]
Input	Function	Modal	[INF_MOD]
Input	Function	Number	[INF_NUM]
Input	Function	Possessive determiner	[INF_PD]
Input	Function	Personal pronoun	[INF_PPR]
Input	Function	Preposition	[INF_PREP]
Input	Function	Quantifier	[INF_QNT]
Input	Function	"to"	[INF_TO]
Input	Function	Wh-word	[INF_WH]
Input	Math Symbol	n/a	[INM_MS]
Modified	Content	Abbreviation	[MOC_ABR]
Modified	Content	-ing participle	[MOC_ING]
Modified	Content	Noun	[MOC_N]
Modified	Content	Past participle	[MOC_PP]
Modified	Content	Verb	[MOC_VB]
Modified	Function	Number	[MOF_NUM]

Appendix C: Procedures for transcript analysis in SALT

To analyze transcripts in SALT:

1. Copy each text as a single entry in SALT.
2. Save each entry using the following naming convention:

   Grade_Text number (e.g., 3_001, 3_002)

3. Use the error check button to check the text for problems. Correct any issues.
4. Once all responses have been entered for a grade level, enter the list of codes used in the dataset as a code list.
5. Use “Rectangular Data File” under the “Tools” menu to conduct a batch analysis of all texts with the grade level.
   a. Under Standard Measures Report, select Total Completed Words, MLU in words, Number of Different Words, Type Token Ratio.
   b. Under Explore, load the code list for the grade level. Check boxes to count: Number of Occurrences.
6. Once the rectangular data file has been generated as a .csv file, save as a Microsoft Excel file (.xlsx).

Appendix D: Phrase-level coding procedures

Phrase-level Coding identifies the following codes:

Code	Definition
[EC]	Strings of four or more words exactly copied from the task input
[MINR]	Strings of four or more words from the task input with minimal revisions. Minimal revisions are defined as approximately one change every four words. Strings of text with more extensive changes should not be coded.

Note: A revision is defined as a the modification of a word, the substitution of a word or phrase (e.g., synonym replacement), or the addition or deletion of a word or short phrase. For example, the deletion of a prepositional phrase in a string of text would count as one change rather than counting the deletion of each word as a separate change. Strings coded as [MINR] should have a clear link to the task input. Strings of text with more substantive revisions were not coded in this phase.

1. Open an electronic copy of the task input.
2. Open a copy of student response transcripts for one task in a Microsoft word document.
3. Using the text search feature in MS Word, identify strings of input text. Key content words or phrases were used to identify input text strings.
4. Each input text string was coded as an Exact Copy (EC) or Minimal Revision (MINR) along with the number of words in the string. Strings of moderately or substantially revised text were not coded but in some instances were highlighted and marked for later review during qualitative analysis of text strings.
5. Transcripts were loaded into SALT to tabulate code totals.

6. Note: For grade 9 samples, Exact Copies or Minimally Revised strings of input taken from math equations were also marked as "ME" for Math Equation so that these instances could be counted an analyzed separately.
7. After coding, all instances of MINR codes were checked a second time to ensure that they met the criteria for minimal revision.

Appendix E: Qualitative coding

1. I randomly selected 25 texts at each score point using www.random.org, a random number generator.
2. I printed out copies of all texts identified for analysis (80 per grade level) and a copy of the relevant task.
3. I first read through all tasks at a score point, beginning with 2-2-2 and moving upward. During a first read-through, the researcher made notes about revalent features.
4. During a second read-through, I created a set of codes and began applying these to each text. Coding was an iterative process, with new codes being added as needed. If this occurred, the researcher would begin at the start of the set.
5. After coding a set of responses at a score point, I made general notes describing key features and observations about the texts.
6. Codes were used as a basis for the next score point review. As needed, new codes were added to reflect changes in text features by score point.
7. After all texts at a grade level were reviewed and coded, I tabulated codes for each text in Excel and typed any notes. This allowed for easy review and quantification.