ABSTRACT

ALLEN, LYDIA ANGELEA. Exploring the Use of Oral Formative Assessment in the Secondary Mathematics Classroom. (Under the direction of Dr. Karen Hollebrands).

The purpose of this study is to explore the use of oral formative assessment in secondary mathematics classrooms as a link between conceptual understanding and procedural fluency. While oral formative assessment is not new to mathematics education, it has not been widely studied or implemented at the secondary level. The present study specifically investigates the use of oral formative assessment in an Advanced Placement (AP) Statistics course during a unit of study on the normal distribution. The study examines conceptual knowledge as demonstrated by an oral assessment, and then explores how that knowledge translates to procedural proficiency on a written assessment.

The participants in this study were fifteen AP Statistics students in a public high school in North Carolina. The data collected in this study consisted of oral formative assessments recorded individually by each student and written formative assessments taken by each student during class time.

The conceptual framework for the study considers Star's (2005) and Baroody et al.'s (2007) work on types and qualities of demonstrated knowledge in mathematics. Students' oral assessment data was analyzed for evidence of superficial procedural knowledge, deep procedural knowledge, superficial conceptual knowledge, and deep conceptual knowledge. Students' demonstrated knowledge was then compared to their proficiency on the written assessment.

Findings from the study suggest that students who demonstrate deep conceptual knowledge in an oral assessment also tend to demonstrate high proficiency on written assessments. That being said, students who demonstrate superficial procedural knowledge in an oral assessment do not necessarily demonstrate low proficiency on written assessments. While

demonstrated knowledge on oral formative assessments is not determinative of proficiency on written assessments, it does provide teachers with a unique perspective as to how their students are thinking and reasoning mathematically. The mathematics education community can use studies such as this for insights into the way that oral formative assessment can reveal and promote students' conceptual understanding.

Exploring the Use of Oral Formative Assessment in the Secondary Mathematics Classroom

by
Lydia Angelea Allen

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Master of Science

Mathematics Education

Raleigh, North Carolina

2022

APPROVED BY:

Dr. Karen Hollebrands
Committee Chair

Dr. Robin Anderson

Dr. Molly Fenn

DEDICATION

To my parents.

BIOGRAPHY

Lydia Allen was born in Stamford, Connecticut on June 8, 1992. Soon afterwards, she and her family moved to Burlington, North Carolina, her father's hometown. After being homeschooled by her parents through first grade, Lydia was a student in the Alamance-Burlington School System through her sophomore year of high school, attending Elon Elementary School, Turrentine Middle School, and Walter M. Williams High School. For her junior and senior years of high school, Lydia attended the North Carolina School of Science and Mathematics in Durham, graduating in 2010.

After receiving a North Carolina Teaching Fellows scholarship, Lydia enrolled at North Carolina State University. At NC State, Lydia was a Caldwell Fellow and a member of the University Honors Program, and she was inducted into Phi Beta Kappa in 2012. She spent the 2013-2014 academic year studying at the University of Guanajuato in Guanajuato, Mexico, where she completed coursework in both mathematics and Spanish language, literature, and culture. Lydia graduated from NC State in 2015 with a Bachelor of Science in Mathematics, a Bachelor of Science in Mathematics Education, and a Bachelor of Arts in Spanish Language and Literature, summa cum laude in each degree.

In 2015, Lydia moved back to Alamance County and began her public school teaching career at Western Alamance High School, where she taught AP Calculus AB, AP Calculus BC, NC Math 2, Discrete Math, and Essentials for College Math. In 2018, Lydia moved to Walter M. Williams High School, where she currently teaches AP Calculus AB, AP Calculus BC, AP Statistics, and NC Math 3. In 2020, Lydia began to pursue a Master of Science in Mathematics Education at NC State while continuing to teach full time at Williams High School. After receiving her master's degree, Lydia plans to continue teaching high school mathematics.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my family. To my parents, Philip and Melissa Allen, for all you have taught me and for all the opportunities you have provided me. Thank you for being my very first teachers. To my sister, fellow teacher, and best friend Veronica, whom I've looked up to since I was born. And to my brother Garret, who has one of the best mathematical minds I know.

I would like to thank Dr. Karen Hollebrands, who advised me throughout my graduate career and throughout the process of writing this thesis. It has been an honor to work with you. I would also like to thank my committee members. Thank you to Dr. Molly Fenn, whose innovative oral assessment practices inspired this thesis, and thank you to Dr. Robin Anderson, whose course provided me the first opportunity to implement oral assessments in my classroom.

I would also like to acknowledge a few of my earlier mathematics teachers. In particular, thank you to Ms. Jennie Beedle at Elon Elementary School for first recognizing my inclination towards mathematics, to Ms. Becky Caison at Williams High School for providing me with solid foundational skills, and to Dr. Dan Teague at NCSSM for inspiring me to study mathematics education at NC State.

I am the teacher I am today because of two phenomenal educators – Ms. Tracey Weigold and Dr. Tina Starling. Tracey, thank you for being the best cooperating teacher I could have imagined, and Tina, thank you for connecting me with Tracey and mentoring me along the way. I am honored to now call you both colleagues and friends.

Lastly, thank you to my students, both past and present. I did not know when I started teaching just how much I would grow to care about you, both as mathematicians and as human beings. I may be your teacher, but you all have taught me more than you know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

The mathematical teaching practice of building procedural fluency through conceptual understanding is foundational to mathematics education (NCTM, 2014). Rittle-Johnson et al. (2001) define procedural knowledge as "the ability to execute action sequences to solve problems" and conceptual knowledge as "implicit or explicit understanding of the principles that govern a domain and of the interrelations between units of knowledge in a domain" (p. 346-347). In other words, procedural knowledge tends to be focused on the ability to produce a correct response, whereas conceptual knowledge focuses on a correct understanding of the process that leads to a response. Deep conceptual understanding allows students to use their knowledge flexibly, using critical-thinking and problem-solving skills to find solutions to contextual and mathematical problems (NCTM, 2014).

While it is clear that conceptual understanding is important in mathematics, most assessment tools are lacking in their ability to provide meaningful information about students' conceptual knowledge (Niemi, 1996). Many assessments only look at the final product, meaning the student's thinking process is either obscured or completely hidden. As a potential solution to this, Kazemi (1998) suggests implementing the alternative assessment practice of having students explain their thinking and share their problem-solving strategies, which allows them to engage in and demonstrate conceptual thinking.

Brown (1987) offers the belief that rigorous understanding of a concept implies one's ability to consciously draw forth knowledge and reflect on it, specifically through a verbal representation. However, verbal representation can be both written and oral. This begs the question of whether oral explanations are inherently different from written explanations, especially in the context of a mathematics classroom. Is it just as conducive to learning for students to write a paragraph explaining their thinking as it is for them to explain their thinking aloud? The literature suggests that oral explanations offer unique benefits.

Oral assessments, sometimes referred to as interview-based assessments (Ellemor-Collins & Wright, 2008) or clinical interviews (Ginsburg, 1981) allow teachers “to see problems through the eyes of the students, to respond to each student's particular needs, and to focus on stages of learning rather than answers” (Long & Ben-Hur, 1991, p. 46). In other words, teachers are able to observe the process behind the product, better understanding the conceptual strengths and weaknesses of a student. Through this, the teacher is better able to engage in the iterative process of helping students connect conceptual knowledge to procedural knowledge and vice versa (Rittle-Johnson et al., 2001). Furthermore, oral explanations can also offer teachers nonverbal clues as to how a student engages with math, including body language and processing time (Ellemor-Collins & Wright, 2008).

Oral explanations can provide intrinsic benefits for students as well as teachers. For example, studies have shown that gesturing can improve a person’s cognitive capacity. When students are able to represent their ideas in a visuospatial manner, it may enhance their ability to process information (Goldin-Meadow et al., 2001). The ability for students to use gestures in oral explanations might help strengthen their cognitive abilities in a way that written explanations cannot.

Studies have also shown that confidence in mathematics is strongly correlated with achievement for students, and teacher-student interactions can play a role in developing that confidence (Reyes, 1984). Through interactions like clinical interviews, teachers have a unique medium to understand a student’s self-concept and self-confidence regarding learning mathematics (Long & Ben-Hur, 1991). With paper-and-pencil assessments alone, the intangible variable of student confidence is difficult to observe or measure.

Statement of the Problem

With the sudden advent of the COVID-19 pandemic in March 2020, the education system was forced to make instant, drastic changes. As terms like "remote learning" and "asynchronous instruction" flooded the educational lexicon, teachers all over the world were faced with the daunting challenge of facilitating student learning through a computer screen.

One of the biggest losses for mathematics teachers in this new context was the ability to circulate throughout the classroom. As Cole (1999) states, "walking around is the most immediate, efficient assessment method" that teachers have available. Instead of being able to observe and facilitate dynamic student thinking, teachers often quite literally had only a "snapshot" of a student's work during remote instruction. Furthermore, with so much problem-solving technology available to students, it became increasingly difficult for teachers to gauge their students' true understanding of mathematical concepts.

The silver lining, however, is that unprecedented educational challenges can serve as an impetus for creative educational advancements. As teachers lost the ability to walk around their classrooms, they were forced to find new ways to assess their students' thinking and conceptual understanding. This opened the door for teachers to explore the use of video recording software to allow students to share their ideas orally. Such technology allows teachers to have immediate, convenient access to student information, which can allow for more individualized learning (Baran, 2014).

However, as teachers began to find innovative ways to use oral formative assessments in their instruction, the lack of research in this field became apparent, especially for teachers of secondary students. While literature does exist regarding the benefits of verbal justifications and clinical interviews in mathematics education, much of it has been conducted in the realm of elementary school classrooms (Rittle-Johnson et al., 2001; Kazemi, 1998; Niemi, 1996). There is a need for exploration and discovery into how oral formative assessments can enhance secondary mathematics teaching and learning as well.

Furthermore, while there is literature promoting the effectiveness of dynamic, one-on-one interviews between a teacher and a student, interview-based assessments are not routinely used due to their time-consuming nature (Ellemor-Collins & Wright, 2008). The advent of technologies such as *VoiceThread* (https://voicethread.com/) and *Flipgrid* (https://info.flipgrid.com/) have made it possible for students to document their thinking in a medium that is less time-consuming for teachers to analyze. However, there is a lack of research into how such tools can be used effectively in a secondary setting.

This thesis intends to explore the use of oral formative assessments in a secondary mathematics classroom. The present study specifically examines how the use of oral formative assessments during lessons on the normal distribution in an AP Statistics classroom can reveal and promote students' conceptual understanding and procedural fluency.

CHAPTER 2: REVIEW OF LITERATURE

In this chapter, literature related to assessment in mathematics education will be reviewed in order to provide a basis for consideration of the role of oral formative assessment in the secondary mathematics classroom. This chapter will start by looking broadly at assessment in mathematics education before focusing on formative assessment. Then, oral formative assessment will be addressed specifically, first by reviewing the different types of oral formative assessments, and next by exploring how these assessments can be used in the mathematics classroom by both teachers and students. Particular attention will be given to the definitions of conceptual and procedural knowledge, and a framework for understanding the types and quality of knowledge that students may demonstrate in oral formative assessments will be presented. The research questions investigated in the present study will be posed at the end of this chapter.

Assessment in the Mathematics Classroom

Assessment can be broadly analyzed as a facet of general education, but it is important to consider its unique manifestation in a mathematics classroom. In *Assessment Standards for School Mathematics* (1995), NCTM describes assessment as "the process of gathering evidence about a student's knowledge of, ability to use, and disposition toward mathematics, and of making inferences from that evidence for a variety of purposes" (p. 3). This comprehensive definition points to three important assessment metrics:

- *A student's knowledge of mathematics* is analogous to the student's conceptual understanding of a mathematical topic.
- *A student's ability to use mathematics* is closely tied to the student's procedural fluency in a mathematical topic.

- *A student's disposition toward mathematics* is related to the student's motivation, self-concept, and self-confidence in mathematics.

However, just collecting evidence around these metrics does not inherently count as assessment. As the definition explains, true assessment lies in analyzing the collected evidence and using it for some purpose. This purpose is the defining characteristic which separates mathematical assessments into two categories: summative assessments, whose purpose is to judge students' mastery of content, and formative assessments, whose purpose is to provide feedback that improves student learning (McIntosh, 1997). It is interesting to note that the difference between formative and summative assessments does not lie in the method of assessment, but rather in the interpretation of the assessment data and the decisions that come about as a result of that interpretation (Shermis & Di Vesta, 2011).

Summative assessments, which frequently occur at the conclusion of an instructional cycle, interpret assessment results as some sort of final judgment on a student's learning. Especially at the secondary level, this final judgment is often expressed in the form of a grade (Senk et al., 1997). In many mathematics classrooms, this grade is "calculated" in large part by considering the percentage of objective questions a student answers correctly on a paper-and-pencil (or digital point-and-click) assessment. Especially when these assessments are externally created at the district or state level, students are frequently asked to demonstrate nothing more than low-level procedural skills through multiple-choice questions (NCTM, 2014).

Summative assessments are often used for accountability purposes, assigning a concrete measure to the achievement status of students. These high-stakes assessments might serve to evaluate the fitness of students to advance to subsequent mathematics courses, while also evaluating teachers and administrators on their ability to appropriately prepare students for those courses (Shermis & Di Vesta, 2011). In *Principles to Action* (2014), NCTM laments the importance given to this function of summative assessments, stating, "in the name of accountability, the rich potential for using assessment processes to strengthen student learning and improve instruction has been diminished" (p. 90).

In contrast, formative assessment implies the belief that assessment is not an end in itself, but rather the means through which teachers and students identify areas for improvement in teaching and learning (Shermis & Di Vesta, 2011). This challenges the notion that assessment is something done *to* students, suggesting instead that assessment is done *for* students as a way to improve their learning (NCTM, 2000). In short, formative assessment is "assessment for the purpose of instruction" (Ginsburg, 2009, p. 110).

Instead of occurring at the end of a learning cycle, formative assessments should be integrated throughout the instructional process (Sherman & Di Vesta, 2011). Senk et al. (1997) note that some forms of assessment, especially those with the purpose of assigning grades, can consume a large portion of instructional time in the classroom. Well-executed formative assessments, on the other hand, are an ongoing, integral piece of classroom instruction. Formative assessments should be part of the learning process, not an interruption to it (NCTM, 2000).

Ginsburg (2009) proposes that, through formative assessment, teachers should be seeking information in four different categories to appropriately inform instruction. These categories are: performance, thinking/knowledge, learning potential, and affect/motivation.

- *Performance* is the most traditional assessment measure, in which teachers consider the overt achievement of the student during a specific mathematical task.
- *Thinking/knowledge* alludes to the cognitive processes and conceptual understanding

contributing to a student's ability (or inability) to perform the mathematical task.

- *Learning potential* refers to the readiness of the student to engage in the mathematical task, namely whether or not the student has the appropriate foundational skills and processes necessary for the mathematical task.
- *Affect/motivation*, while often not inherently tied to the mathematical task at hand, is an important facet of formative assessment. Supporting student learning requires teachers to understand and attend to students' mathematical dispositions.

The information that the teacher gains in these categories should be actionable, leading to improved instructional practices that support student learning.

NCTM (2014) suggests that "shifting the primary focus and function of assessment from accountability to effective instructional practice is an essential component of ensuring mathematical success for all students" (p. 98). This call for greater attention to formative assessment exposes the need for a renewed analysis of traditional formative assessment practices, as well as innovative formative assessment methods that leverage the use of technology as a tool for students to express their mathematical thinking.

Types of Formative Assessments

The power of formative assessments lies in their diversity; almost any action taken by the teacher has the potential to serve as a source of formative assessment. McIntosh (1997) presents the belief that good teaching and good assessing are one and the same. From this belief, it can be argued that good teaching involves constantly assessing students in a variety of ways, both informally and formally.

By definition, informal assessments require little to no planning, but rather happen continuously throughout the learning cycle as teachers are observing and interacting with their students (Shermis & Di Vesta, 2011). For example, teachers can informally assess student performance through a quick look at students' work while circulating throughout the room. Students' thinking/knowledge can be informally assessed through a short conversation between a student and teacher. Finally, learning potential and affect/motivation can be assessed through the informal observation of student body language. These are just a few examples of the diverse ways in which informal assessment can and should permeate a mathematics classroom.

Formal formative assessments, on the other hand, require a certain amount of planning. With this more formal type of assessment, the teacher designs opportunities or instruments that elicit specific information from the student (Shermis & Di Vesta, 2011). Traditional forms of formal assessments are quizzes and tests, which can be considered formative provided that the information gathered from these assessments is used to adjust instruction to better support students (NCTM, 2014). Less traditional forms of formal formative assessments include rubrics and exit tickets that correspond to pre-planned student outcomes (Baron, 2016). Again, all four assessment categories (performance, thinking/knowledge, learning potential, and affect/motivation) can be addressed through means of formal formative assessment.

As shown, there are countless ways that teachers can engage in formative assessment. But while formative assessment opportunities are plentiful, they are not all equal in their usefulness to teachers in supporting student learning. In order to maximize the effectiveness of formative assessments to improve instruction, teachers must look deeper than the superficial analysis of whether or not student work is "right or wrong." They should instead focus on how students are thinking about mathematical ideas (NCTM, 2000). The problem, however, is that many assessments are very limited in their ability to provide real insight into student thinking (Niemi, 1996). Analysis of student thinking through assessment cannot be an afterthought, but instead necessitates that the teacher engage in conscientious assessment planning. As NCTM (2014) explains, "obtaining evidence about understanding and reasoning requires the use of tasks and methods designed for those purposes" (p. 92). The challenge, then, is for teachers to develop and adapt formative assessments that are purposefully designed to elicit clear evidence of student thinking.

Oral Formative Assessments

In this quest for teachers to obtain evidence about student thinking, asking students for written explanations of their work has been a popular method. There have been a variety of studies over the past two decades suggesting the benefits of written explanations in the mathematics classroom (Santos & Semana, 2015; Back et al., 2009; Albert, 2000).

In addition to written explanations, there has also been some evidence that oral processing can provide students with unique advantages. A study by Carraher et al. (1987) asked sixteen third graders in Brazil, ranging in age from 8 to 13 years, to solve a series of arithmetic problems using the four basic operations - addition, subtraction, multiplication, and division. The students' success at solving these problems was then analyzed with regard to whether they chose to solve the problem using oral or written procedures. The study found that oral computation was linked to a greater probability of success in all four operations. In conclusion, the study explains that "oral mathematics can no longer be treated merely as idiosyncratic procedures nor inconsequential curiosities" (p. 96).

Joyner and Bright (2016) also encourage the use of oral formative assessment practices, using the following justification:

> Hearing students explain why and how they arrive at a given response gives us as

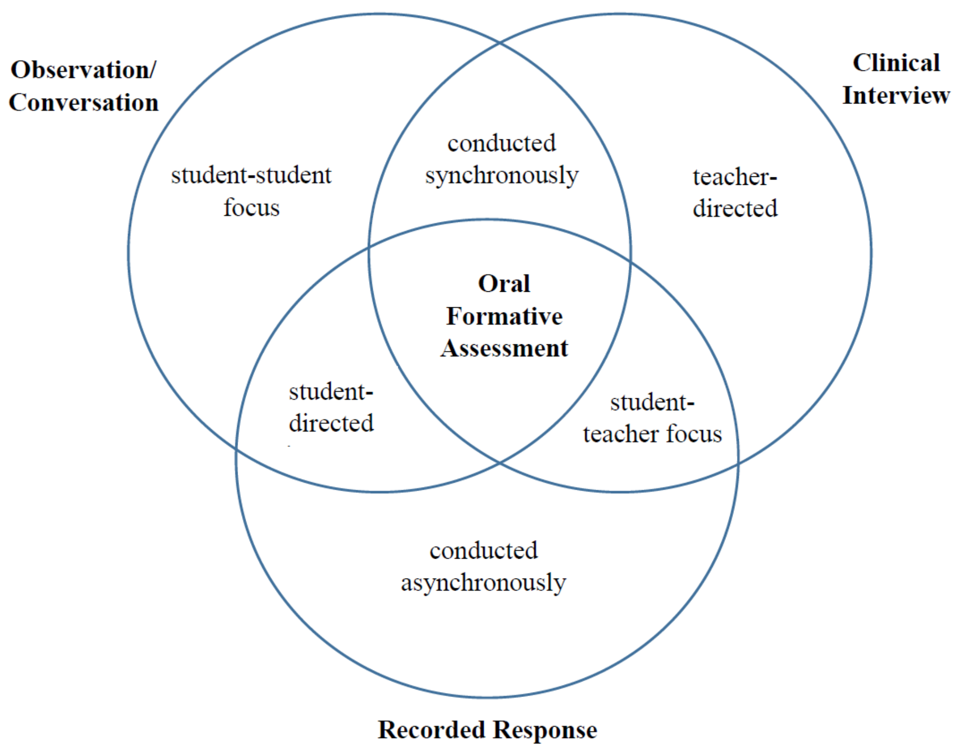
teachers a chance to recognize incorrect understanding as well as correct but incomplete understanding. For example, one question we might ask ourselves during a conversation is, "Are students reasoning in a valid way, or are they simply repeating steps in a procedure without understanding what those steps mean?" (p. 130)

Oral formative assessments gives teachers a clearer picture of what students really understand, not just what they can produce.

In heeding the call to explore oral mathematics more deeply, this chapter continues with a discussion of well-known forms of oral formative assessments, observations/conversations and clinical interviews, as well as a suggestion of a less common form of formative assessment, which will be termed as recorded response. Figure 1 summarizes the relationship between these forms, and then each is discussed in greater detail.

Figure 1

Types of Oral Formative Assessments



Observation/Conversation

As the terms "math talk" and "discourse" have become more prevalent, it is clear that having students talk to each other about mathematical concepts is an important aspect of mathematics education (Walter, 2018; Ghousseini et. al, 2017; Wagganer, 2015). NCTM (2014) counts the facilitation of meaningful mathematical discourse as one of its eight Mathematics Teaching Practices, explaining that "discourse in the mathematics classroom gives students opportunities to share ideas and clarify understandings, construct convincing arguments regarding why and how things work, develop a language for expressing mathematical ideas, and learn to see things from other perspectives" (p. 29).

The teacher, then, has the opportunity to use observation of student discussions as a method of formative assessment. If students are presented with a task to work on in groups, the teacher can circulate throughout the classroom and listen in on student conversations. Walking around and hearing discourse between students is the most immediate way for teachers to figure out how students are engaging with mathematical content in the moment (Cole, 1999). It is a form of active formative assessment, meaning the assessment is taking place during instruction and can instantly be used by teachers to inform their instruction in the moment (Reinholz & Gillingham, 2017).

An observation of student discussion can also turn into a conversation if the teacher decides to intervene in any way, whether this be through whole class discussions or a quick conference with a group or an individual student. This allows the teacher to ask clarifying questions or request that students explain certain aspects of their work (Cole, 1999).

One difficulty of observation is that catching short snippets of student conversations does not often lead the teacher to deep knowledge of the students' reasoning and understanding (Cole, 1999; Ginsburg 2009). Cole (1999) suggests that teachers “bring [a] chair and consider watching a group for ten minutes or more” (p. 225). While this would be a wonderful thing to do if teachers had the time, “observation is simply a very hard thing to do in a classroom of 25 or 30 students” (Ginsburg, 2009, p. 112).

Ginsburg (2009) does point out that researchers often use video recordings to capture longer observations of students. He admits, however, that this is impractical for teachers to do with any regularity. Collecting hours and hours of footage of students working and conversing in groups is not useful, as no classroom teacher would have time to routinely assess such lengthy recordings.

Clinical Interview

When teachers are purely observing students interact with each other without much external input, they have the opportunity to hear student thinking expressed more “naturally.” While this might provide interesting information, simply observing students or having a brief conversation does not allow teachers much opportunity to elicit specific information from students with regards to their reasoning and understanding. One potential solution to this is for teachers to engage with students in what is known as a “clinical interview.”

In general, a clinical interview takes place between an interviewer and interviewee (generally, a teacher/researcher and a student), and the interviewer presents tasks, observes, and asks clarifying questions designed to elicit student thinking and better understand the thought process of the student (Ginsburg, 2009). Long and Ben-Hur (1991) offer this insight into clinical interviews:

> The clinical interview yields information not easily available from other sources. It gives insights into students’ experiences by permitting the teacher to understand the meanings

> that students find in mathematical problems and to appreciate their feelings and confidence about learning mathematics. It furnishes the teacher with information about students' backgrounds and learning styles, their strengths and weaknesses, and about cultural differences that may affect their views of, or success in, mathematics. It also allows teachers to stress process over product, encourages talking about mathematics, and gives teachers and students instant feedback. (p. 44)

Long and Ben-Hur go on to explain that clinical interviews allow teachers to probe if students understand that different methods can lead to identical results, giving insight into a students' conceptual understanding of the topic at hand. As Ginsburg (2009) explains, the clinical interview is "rewarding because it can provide contact between the teacher's mind and the student's" (p. 115).

These interviews can be recorded so the teacher can play back the video at a later time. Ellemor-Collins and Wright (2008) tout the advantages of making video recordings of clinical interviews, creating what they term a videotaped interview-based assessment. This allows the teacher to not take any notes during the interview, but instead focus on observing the student and working to probe the student's understanding.

Clinical interviews are not uncommon in elementary schools, but they are rarely used in secondary school settings (Joyner & Bright, 2016). Again, time is the biggest constraint. Especially when these assessments are recorded for later viewing, the time required to both conduct interviews and review videos for even a few students makes it impractical to routinely implement (Ellemor-Collins & Wright, 2008). Furthermore, conducting clinical interviews is no easy feat. Interviewing "depends on human skill, may not produce reliable results, and indeed done badly can distort student thinking and elicit only what the student thinks the interviewer wants to hear" (Ginsburg, 2009, p. 115).

Recorded Response

As described above, observation/conversation is a widely-used form of formative assessment that can provide quick insight into student thinking. The downfalls of this type of assessment, however, are that not all students will be able to share their voice, and the teacher might only be able to observe student thinking at a surface level. Clinical interviews, on the other hand, allow teachers to delve much deeper into student thinking, but they are time consuming and impractical to do with any regularity.

One solution to these problems is to implement a third type of oral formative assessment: a recorded response. This type of assessment asks students to orally explain their thinking through recording a short video of themselves talking through a problem. For example, instead of assigning twenty factoring problems for homework, a teacher might ask each student to record a brief video explaining how to factor a given quadratic trinomial. This allows every student voice to be heard, stresses the importance of student thinking over correct responses, and gives teachers the ability to quickly and asynchronously assess students.

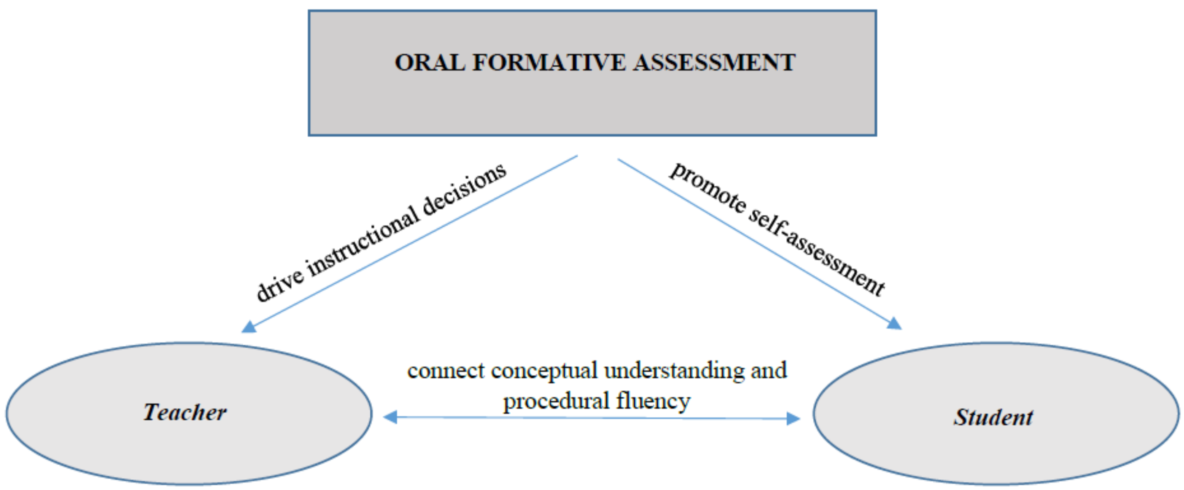
This type of formative assessment is relatively new, as it has only recently been made possible through consistent student access to personal recording devices. As such, there is very limited evidence of and research into this practice being implemented in secondary mathematics classrooms.

Using Oral Formative Assessments Effectively

Ginsburg (2009) expresses the core idea of formative assessment in this powerful statement: "The bottom line attribute of formative assessment is its *actionable* character" (p. 111). In other words, what makes an assessment formative is not any concrete characteristic, but rather the abstract ability to use the information gained to improve student learning. As stated previously, oral assessments can provide unique benefits for both teachers and students. Teachers can use oral formative assessments to drive instructional decisions, students can use oral formative assessments to promote self-assessment, and students and teachers can jointly use oral formative assessments to increase conceptual understanding. The interconnectedness of these uses is shown in Figure 2, and then each use is considered more deeply.

Figure 2

Teacher and Student Uses of Oral Formative Assessment



To Drive Instructional Decisions

Regardless of what form of oral formative assessment is implemented, one major purpose is for teachers to use the results to drive instructional decisions in the classroom. NCTM (2000) explains that

> instructional decisions made by teachers - such as how and when to review prerequisite material, how to revisit a difficult concept, or how to adapt tasks for students who are

> struggling or for those who need enrichment - are based on inferences about what students know and what they need to learn. Assessment is a primary source of the evidence on which these inferences are based, and the decisions teachers make will only be as good as that evidence. (p. 23)

The challenge, then, becomes for teachers to find assessments that provide good evidence with which to make inferences about student understanding.

With paper-and-pencil tasks, teachers often get a one-dimensional viewpoint of student performance (NCTM, 2000). Sometimes teachers might erroneously assume that a student who arrives at a correct response is doing so through correct reasoning. Joyner and Bright (2016) caution that if "instruction assumes too much … about students' knowledge, students may internalize incorrect ideas" (p. 312).

Oral assessments specifically can provide unique insight into how students are thinking and reasoning. Instead of having teachers assume what students do and do not understand, students can explain their understanding in their own words. Then, these formative assessments can serve their purpose of increasing "precision in how instructional time is used in class" and assisting "teachers in identifying specific instructional needs" (NMAP, 2008, p. 48).

To Promote Self-Assessment

While some definitions of formative assessment seem to primarily focus on impacting instructional decisions (NCTM, 2000; Ginsburg, 2009), formative assessment is shortchanged when meant only for use by the teacher. Students, not teachers, should be at the center of the assessment process. In *Principles to Action* (2014), NCTM states that "an important goal of assessment should be to make students effective self-assessors, teaching them how to recognize the strengths and weaknesses of past performance and use them to improve their future work"

(p. 95). Teachers have the responsibility to help students learn how to monitor their own learning (Joyner & Bright, 2016).

Oral formative assessment lends itself well to student self-assessment. As found in the study by Carraher et al. (1987) mentioned previously, when problems focused on written algorithmic solutions, students tended to lose track of the meaning of the problem itself. Carraher et al. conclude that the

> loss of meaning is likely to be the explanation for children's willingness to accept results such as a remainder larger than the minuend - a result that is promptly rejected if the child refers back to the meaning encoded by the particular computation. (p. 95)

In other words, the meaning and context of a problem seems to be better preserved when explained orally rather than when simply written down. And when meaning is preserved, students are better able to assess the validity of their answer in the context of that meaning, helping them self-correct when they arrive at an impossible or improbable solution.

Furthermore, self-assessment is not limited to a student's analysis of his or her mathematical accuracy. In the process of reflecting on what they know, students gain confidence in their ability to learn (Joyner & Bright, 2016).

To Connect Conceptual Understanding and Procedural Fluency

Finally, as stated previously, building procedural fluency through conceptual understanding is foundational to mathematics education (NCTM, 2014). The primary focus of this thesis is to investigate the connection between procedural fluency and conceptual understanding through the lens of oral formative assessments. As such, before analyzing the utility of oral formative assessments for this purpose, the definitions of conceptual understanding and procedural fluency, as well as their interconnectedness, will be discussed in greater detail.

Framework for Understanding Procedural and Conceptual Knowledge

As Murray & Sharp (1986) readily admit, "The relationship between our ability to perform a task and to understand both the task and why our action is appropriate appears - even with only a little reflection - to be complicated" (p. xi). This relationship has been regularly studied in the general field of cognition, but its analysis occupies a special place in mathematics, in particular. Resnick & Ford (1981) explain, "the relationship between computational skill and mathematical understanding is one of the oldest concerns in the psychology of mathematics" (p. 246). Since the mid-1980s, the terms *procedural knowledge* and *conceptual knowledge* have been widely used by mathematics education researchers in the exploration of this relationship (Star & Stylianides, 2013).

In their seminal work on conceptual and procedural knowledge in mathematics, Hiebert & Lefevre (1986) define conceptual knowledge as

> … knowledge that is rich in relationships. It can be thought of as a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information. Relationships pervade the individual facts and propositions, so that all pieces of information are linked to some network. (pp. 3-4).

This definition is closely related to the previously presented definition by Rittle-Johnson et al. (2001), who suggest that conceptual knowledge is concerned with the "interrelations between units of knowledge" (p. 346). For example, in a high school trigonometry class, conceptual knowledge might consist of relating the ratios of special right triangles to certain coordinates on the unit circle.

In contrast to conceptual knowledge, Hiebert & Lefevre (1986) describe procedural knowledge in two parts:

> One kind of procedural knowledge is a familiarity with the individual symbols of the system and with the syntactic conventions for acceptable configurations of symbols. The second kind of procedural knowledge consists of rules or procedures for solving mathematical problems. Many of the procedures that students possess probably are chains of prescriptions for manipulating symbols. (pp. 7-8).

Again, this definition is closely related to the previously presented definition by Rittle-Johnson et al. (2001), who suggest that procedural knowledge is typically manifested in a student's "ability to execute action sequences to solve problems" (p. 346). Continuing with the trigonometry class example, procedural knowledge might consist of using coordinates on the unit circle to evaluate trigonometric functions.

While these two definitions appear to highlight the differences between conceptual knowledge and procedural knowledge, many mathematics education researchers have suggested that the interconnectedness and ambiguity between these two types of knowledge is of primary importance (Nilsson, 2020; Star, 2005; Silver, 1986). Instead of separating procedural knowledge and conceptual knowledge - and in doing so, inherently causing debate as to which is more important for students - it is more beneficial to explore how the two types of knowledge work together. In particular, Rittle-Johnson et al. (2015) suggest a bidirectional relationship between procedural knowledge and conceptual knowledge, where improvements in either form of knowledge support development in the other form.

Star (2005) suggests amending the distinction in forms of knowledge to include not just knowledge type (procedural versus conceptual) but knowledge quality (superficial versus deep), as shown in Figure 3.

Figure 3

Types and Qualities of Procedural and Conceptual Knowledge (Star, 2005)

Types and Qualities of Procedural and Conceptual Knowledge

Knowledge type	Knowledge quality	
	Superficial	Deep
Procedural	Common usage of *procedural knowledge*	?
Conceptual	?	Common usage of *conceptual knowledge*

Baroody et al., (2007) built on this proposed amendment in a more detailed framework, shown in Figure 4. This framework helps to elaborate on the differences between superficial and deep knowledge quality, as well as introducing a knowledge type that highlights the intersection of procedural and conceptual knowledge.

Figure 4

Efforts to Define Types and Qualities (Continua) of Procedural and Conceptual Knowledge (Baroody et al., 2007).

Efforts to Define Types and Qualities (Continua) of Procedural and Conceptual Knowledge

Knowledge type	Knowledge quality		Reference
	Superficial	Deep	
A. Procedural only	Surface-level rules (superficial step-by-step knowledge)	Deeper-level rules (serve to create or modify surface-level rules)	Matz, 1980
	Task-performing procedures	Procedural "operations that involve stepping outside of the system" (take place in a planning space that involves conceptual knowledge)	Davis, 1983
	Weak scheme (disembodied procedural knowledge)	Strong scheme (integrated procedural knowledge)	Baroody, 2003; Baroody & Ginsburg, 1986; Brownell, 1935; Moursund, 2002; Paden, n.d.
B. Conceptual only	Weak schema	Strong schema	Baroody, Cibulsksis, Lai,& Li, 2004; Baroody & Ginsburg, 1986; Baroody, Wilkins, & Tiilikainen, 2003; Lunkenbein, 1985
	Primary-level concept (less abstract)—tied to a specific context	Reflective-level concept (more abstract)—tied to multiple contexts	Hiebert & Lefevre, 1986
C. Both procedural and conceptual	Drill theory	Meaning theory	Brownell, 1935
	Instrumental understanding (rules without reason)	Relational understanding	Skemp, 1987
	Procedural or conceptual	Proceptual	Gray & Tall, 1994
	Knowing about—includes knowing how, that, and why	Knowing to—"active, practical knowledge that nables people to act creatively"	Mason & Spence (1999); cf. Skemp, 1979
	Routine expertise	Adaptive expertise	Hatano, 1988, 2003

In light of the importance of the interconnectedness between procedural knowledge and conceptual knowledge, NCTM (2014) promotes the building of procedural fluency from conceptual understanding as one of its central mathematics teaching practices, claiming that "effective teaching of mathematics builds fluency with procedures on a foundation of conceptual understanding so that students, over time, become skillful in using procedures flexibly as they solve contextual and mathematical problems" (p. 10).

Using Oral Formative Assessments to Reveal and Develop Conceptual Understanding

Rittle-Johnson et al. (2015) suggest that "mathematical competence rests on developing both conceptual and procedural knowledge, and it is widely agreed that conceptual knowledge often supports and leads to procedural knowledge" (p. 594). If students are to have a foundational conceptual understanding on which to build procedural fluency, as proposed by NCTM (2014), assessing a student's conceptual understanding is of utmost importance. That being said, assessing students' conceptual understanding is no easy task. As Niemi (1996) laments, "The need for improved assessments of students' conceptual understanding and ability to use mathematical representations is inescapable" (p. 361). Thankfully, oral formative assessments can aid both teachers and students in the pursuit of increasing students' conceptual understanding, and creating a foundation for procedural fluency.

For teachers, oral formative assessments provide a unique opportunity to hear authentic student thinking. According to Schoenfeld (2007),

> The more that teachers can 'get inside their students' heads' in an ongoing way … the more they will be able to tailor their instruction to students' needs … The more teachers know what students know, the more they will be able to build on their strengths and address their needs. (p. 277).

In order for teachers to build on student understanding and address misconceptions, they must have regular access to how students are thinking conceptually, not just what they can produce procedurally.

Oral formative assessments are not solely beneficial for teachers. For students, the very act of oral explanations helps develop conceptual understanding. According to Carpenter, Franke, and Levi (2003),

> Students who learn to articulate and justify their own mathematical ideas, reason through their own and others' mathematical explanations, and provide a rationale for their answers develop a deep understanding that is critical to their future success in mathematics and related fields. Students not only need to learn the bg ideas of mathematics; they need to learn the mathematical ways of thinking that are entailed in generating these ideas, in deciding how to express them, in justifying that they are true, and in using them to justify the mathematical procedures they are learning. (p. 6)

As suggested above, it is in the act of articulating and justifying their work that students make connections and develop conceptual understanding.

Research Questions

This study investigates the use of recorded responses as a form of oral formative assessment in the researcher's AP Statistics classroom. Building upon Hiebert & Levefre's (1986) definitions of conceptual and procedural knowledge, the work of Star (2005) and Baroody et al. (2007) will guide the analysis of the recorded responses. The questions of interest are:

1. What is the nature of students' conceptual understanding of the normal distribution as revealed by oral formative assessments?
2. How does a student's conceptual understanding translate to procedural fluency on written assessments?

CHAPTER 3: METHODOLOGY

This study is a qualitative study of fifteen students who are all enrolled in an Advanced Placement Statistics course at a high school in North Carolina. The purpose of this chapter is to describe the context of the study, the participants, the sources of data, and the methods used in analyzing the data.

Context for the Study

Participants in the current study are students enrolled at a traditional public high school in North Carolina. According to the National Center for Education Statistics (http://nces.ed.gov), during the previous school year (2020-2021), this high school was classified as a Title I school, meaning students from low-income families accounted for at least 40% of enrollment. This high school had an enrollment of roughly 1200 students during the previous school year, with the largest racial/ethnic identities being White (34%), Black (32%), and Hispanic (25%).

The Advanced Placement Statistics Course

Currently there are 43 students enrolled in Advanced Placement (AP) Statistics at this high school, divided among two sections. Both of these sections are taught by the same teacher, who is in her seventh year of public school teaching and in her fourth year of teaching AP Statistics. The curriculum for AP Statistics is set forth by the College Board, and focuses on the four specific skills outlined in Table 1.

Table 1

Advanced Placement Statistics Course Skills (College Board, 2020)

Skill Category	Description
Selecting Statistical Methods	Select methods for collecting and/or analyzing data for statistical inference.
Data Analysis	Describe patterns, trends, associations, and relationships in data.
Using Probability and Simulation	Explore random phenomena.
Statistical Argumentation	Develop an explanation or justify a conclusion using evidence from data, definitions, or statistical inference.

The Normal Distribution

This study took place while the teacher and her students were engaged in a unit of study on exploring one-variable data, specifically with regard to data that is normally distributed. The normal distribution is a foundational aspect of introductory statistics courses as it can serve as a model for many datasets from the physical, biological, and psychological realms (Batanero et al., 2004). For most students in this course, this unit of study is their first formal exposure to the normal distribution.

According to the College Board (2020), students should be developing the following specific skills while studying the normal distribution:

- Determine proportions and percentiles from a normal distribution.
- Compare measures of relative position in data sets. (p. 49)

The first of these specific skills relates to the general skill of using probability and simulation, while the second specific skill relates to the general skill of data analysis.

Furthermore, while not specifically tied to the topic of the normal distribution, the College Board (2020) promotes the following skill while analyzing one-variable data, relating to the general skill of statistical argumentation:

- Interpret statistical calculations and findings to assign meaning or assess a claim. (p. 41)

These three specific skills served as the objectives for this unit of study. Instruction for this unit took place over the course of five 85-minute class periods during the month of October 2021.

Study Design

The researcher explored the use of oral formative assessments during this unit of study on the normal distribution. Instead of completing the usual paper-and-pencil homework problems associated with determining proportions and percentiles from a normal distribution, students who volunteered to participate in the study recorded short videos of themselves explaining how such problems were solved. These recorded responses were designed with the intent of eliciting students' conceptual understanding of the normal distribution. Following submission of these videos, the participants took short written assessments, designed with the intent of measuring students' procedural fluency with normal distribution calculations.

Participants

From the 43 students enrolled in the two sections of AP Statistics at the school, 15 students volunteered to participate in this study. Parental consent was obtained for participants who are under 18 years of age. The participating students are majority female in terms of gender identity (10 females and 5 males) and majority White in terms of racial identity (13 White, 1 Black, 1 Hispanic). Fourteen of the students are upperclassmen (8 seniors and 6 juniors), and one student is a sophomore.

Sources of Data

Two sources of data were collected for this study: video recordings of participants explaining normal distribution calculations and written assessments of normal distribution calculations. Each source of data will be described, and specific data collection instructions and instruments can be referenced in the appendix. All tasks and assessments included in this study were created entirely by the researcher, modeled on types of problems seen in the AP Statistics curriculum set forth by the College Board.

Using *Flipgrid*, each participant submitted two self-recorded videos engaging in two separate tasks (included in Appendix A). These videos were recorded by the participants alone and on their own time; the researcher was not present for any recording session. The first task provided students with a worked-out solution to a problem involving a normal distribution calculation, and students were asked to create a *Flipgrid* video describing how the solution was reached, explaining each step (how it was done and what it means) as clearly as possible. This task was presented to students on the day that they had encountered similar problems during class, and students had at least two days to submit their *Flipgrid* video. The second task was given during the next class period. Again, this task provided students with a worked-out solution, but this time to a problem involving a different type of normal distribution calculation which had been covered in class on that day. Once again, students were asked to create a *Flipgrid* video describing how the solution was reached as clearly as possible, and they had at least two days to submit their video.

During the class period after each video submission, participants completed a written assessment consisting of multiple parts (included in Appendix B). The first part of each written assessment was identical in form to the type of problem students had described in their video recordings. The subsequent part(s) of each written assessment required students to deviate slightly from the procedures that they had described in their videos.

Analysis of Data

The two sources of data were analyzed in three separate phases - analysis of oral assessments, analysis of written assessments, and a comparative analysis of students' oral and written work. The three phases are described in detail below.

Phase 1: Analysis of Oral Assessments

Before beginning analysis of the oral assessments, the researcher created frameworks for the analysis of each oral assessment with respect to Star's (2005) and Baroody et al.'s (2007) work on types and qualities of procedural and conceptual knowledge. Table 2 shows the framework for analysis of the first oral assessment, and Table 3 shows the framework for analysis of the second oral assessment.

Table 2

Framework of Analysis for Oral Assessment #1

	Demonstrated Conceptual Knowledge		
	Level A	Level B	Level C
Component 1: description of z-score calculation	procedural description of z-score calculation - no interpretation (superficial procedural knowledge)	explanation of z-score as the number of standard deviations above/below the mean (superficial conceptual knowledge)	interpretation of z-score in context of SAT scores (deep conceptual knowledge)
Component 2: description of shaded normal density curve	procedural description of calculating area under a normal curve, using calculator or Table A (superficial procedural knowledge)	explanation of plotting z-score and finding area or percentage to the right of the z-score (deep procedural knowledge)	interpretation of area as the proportion or percentage of observations in the shaded region (deep conceptual knowledge)

Table 3

Framework of Analysis for Oral Assessment #2

	Demonstrated Conceptual Knowledge		
	Level A	Level B	Level C
Component 1: description of obtaining z-score from normal curve	procedural description of obtaining z-score using calculator or Table A (superficial procedural knowledge)	explanation of finding z-score with area of 0.10 above (or 0.90 below) (deep procedural knowledge)	interpretation that the proportion of data above this z-score is 0.10 (deep conceptual knowledge)
Component 2: description of converting from z-score to SAT score	procedural description of converting from z-score to SAT score with no reference to mean or standard deviation (superficial procedural knowledge)	explanation of multiplying the z-score by the standard deviation and then adding the mean (superficial conceptual knowledge)	interpretation of multiplying z-score by the standard deviation to get the total number of points above the mean score (deep conceptual knowledge)

The frameworks help specifically define what type of explanation would demonstrate superficial procedural knowledge (Level A), deep procedural knowledge and/or superficial conceptual knowledge (Level B), or deep conceptual knowledge (Level C) for each component of each oral assessment.

Once these frameworks were established, the video recordings of the oral assessments were downloaded from *Flipgrid*, and the researcher transcribed all of the videos with the aid of *Temi* transcription services. Then, the researcher watched each video multiple times, annotating the transcript of each video with evidence of Level A, Level B, or Level C knowledge for each component of each oral assessment.

Phase 2: Analysis of Written Assessments

The written assessments were analyzed for procedural proficiency. To begin analysis of the written assessments, each part of each assessment was split into two major procedural components, similar to those components evaluated in the oral assessment. A rubric outlining the delineation between low proficiency, medium proficiency, and high proficiency for each component in the first written assessment is shown in Table 4. A similar delineation between proficiency levels for each component in the second written assessment is shown in Table 5.

Table 4

Levels of Proficiency for Written Assessment #1

	Low Proficiency	Medium Proficiency	High Proficiency
Component 1: calculates z-score	does not attempt to calculate z-score	incorrectly calculates z-score	correctly calculates z-score
Component 2: finds area under normal curve	labels and/or shades normal curve incorrectly	correct sketch of normal curve but incorrect percentage	correct percentage (with or without sketch)

Table 5

Levels of Proficiency for Written Assessment #2

	Low Proficiency	Medium Proficiency	High Proficiency
Component 1: obtains z-score	labels and/or shades normal curve incorrectly	correct sketch of normal curve but incorrect z-score	correct z-score (with or without sketch)
Component 2: converts z-score	does not attempt to set up equation	sets up incorrect equation and/or arrives at incorrect value	correct value

Once these rubrics were established, the researcher made copies of all written assessments and annotated each part of each assessment to mark evidence of low, medium, or high proficiency in each component.

Phase 3: Comparative Analysis of Oral and Written Assessments

Once the oral and written assessments had been analyzed separately, participants' oral and written assessments were paired with each other. First, the researcher grouped together all participants who demonstrated Level A knowledge of Component 1 for the first oral assessment and compared their proficiency scores on Component 1 for the first written assessment. The same was done for all participants who demonstrated Level B knowledge and then again for all participants who demonstrated Level C knowledge. The researcher repeated this process for Component 2 of the first set of assessments, grouping the participants by demonstrated knowledge level on the first oral assessment and comparing knowledge level to proficiency score on the first written assessment. This same process was repeated for the second set of assessments.

CHAPTER 4: FINDINGS

This chapter presents the findings of the data analysis. As described in the previous chapter, there are three sources of data: oral assessments, written assessments, and participant surveys. First, the knowledge types and levels exhibited by students in the first oral assessment will be described, attending to Star's (2005) and Baroody et al.'s (2007) work on procedural and conceptual knowledge. Next, the proficiency demonstrated by participants on the first written assessment will be described. Finally, a comparison of participants' oral and written assessment will be presented. This same process will be repeated for the second set of oral and written assessments.

As stated at the end of Chapter 2, this study seeks to answer the following research questions:

1. What is the nature of students' conceptual understanding of the normal distribution as revealed by oral formative assessments?
2. How does a student's conceptual understanding translate to procedural fluency on written assessments?

Oral Assessment #1

The first oral assessment (included in full in Appendix A) provided students with real data about the distribution of SAT scores from the 2019 SAT Suite of Assessments Annual Report, showing that the scores were approximately normally distributed with a mean score of 1059 and a standard deviation of 210. Then, students were informed that the average SAT score of admitted freshmen at North Carolina State University in the fall of 2020 was 1344. The question of interest was as follows: Approximately what percentage of all SAT takers scored higher than 1344?

A worked-out solution to this question was presented to students, whose task was as follows: "Create a *Flipgrid* video describing how this solution was reached. Explain each step (how it was done and what it means) as clearly as possible." The solution presented to students was composed of two major parts. The first part was the calculation of a standardized z-score. The second part was a diagram of a standardized normal density curve with the area of interest shaded in. As such, two major components were analyzed in the students' videos: their description of the z-score calculation and their description of finding the area of the shaded region of the normal density curve. It is important to note that the analysis is based on the knowledge elicited from students in their verbal explanations, which is not necessarily indicative of all a student may know about the topic.

Component #1: Description of Z-Score Calculation

As described in Chapter 3, a student's description of the z-score calculation was classified as Level A (superficial procedural knowledge), Level B (superficial conceptual knowledge), or Level C (deep conceptual knowledge). Table 6 summarizes the results of these classifications, followed by an in-depth description of each level.

Table 6

Results from Component #1 of Oral Assessment #1

	Level A: Superficial Procedural Knowledge	Level B: Superficial Conceptual Knowledge	Level C: Deep Conceptual Knowledge
Number of Students	5	6	4

Level A: Superficial Procedural Knowledge

For the first component, superficial procedural knowledge was identified as a procedural description of the z-score calculation with no interpretation as to what the z-score actually means. Student H demonstrated superficial procedural knowledge in her oral assessment.

> Student H: So, first what we have to do is find our z-score, which they did, which is value minus mean over standard deviation. So they took the value of 1344 minus 1059, and then divided it by 210, which gave us 1.36, approximately.

As shown, Student H correctly provides the steps used to calculate the z-score, but she does not demonstrate any conceptual knowledge of what a z-score actually measures.

Student L also demonstrated superficial procedural knowledge, but with slightly more details that hinted at an underlying conceptual understanding.

> Student L: Because, uh, 1344 doesn't fall, like, exactly on a standard deviation, uh, we have to find a z-score in order to find where that falls on, like, a standardized curve. Um, so to find the z-score you use the equation value minus the mean divided by the standard deviation. And we're given all of these, so, uh, 1344 minus 1059 divided by 210, uh, equals 1.36. And, uh, now we have our z-score, which we see on the graph […] it falls in between 1 and 2, obviously.

In saying that 1344 doesn't "fall … exactly on a standard deviation," Student L is hinting at the idea that a z-score measures how many standard deviations above or below the mean a certain value lies, but she does not state this explicitly.

Student B had a description of the z-score calculation similar to that of Student L, with glimmers of recognition that a z-score is a number of standard deviations.

> Student B: So, in this problem you're trying to find the area under the curve in this specific shaded region. However, where it's located, um, on the curve is not at a specific standard deviation point. So it's not at, like, 1 or 2. It's kind of in the middle between 1 and 2. So you have to find the z-score to be able to find

exactly where that is located. So the average SAT score was 1344. So that's gonna be your first value and you're gonna subtract it from, uh, you're going to subtract the mean from that value. And, um, then you're gonna divide it by the standard deviation to be able to get your z-score, which gives you 1.36.

Similar to when Student L stated that 1344 doesn't "fall … exactly on a standard deviation," Student B notes that "where it's located … is not at a specific standard deviation point." In these comments, both students imply at least a cursory understanding that a z-score measures the number of standard deviations that a data point is away from the mean. However, neither student stated this understanding explicitly, and as such, did not demonstrate sufficient conceptual knowledge to be considered at Level B.

Level B: Superficial Conceptual Knowledge

Next, superficial conceptual knowledge of the first component was identified as an explicit explanation of the z-score being the number of standard deviations that a data point is above or below the mean. Student K provided a Level B explanation in her oral assessment.

> Student K: Okay, so for this equation, we're gonna start by finding the z-score, which is the value minus the mean over the standard deviation. And the value in this equation is 1344, minus 1059, which is the mean, over 210, which is the standard deviation. And then you're gonna calculate that and do 1344 minus 1059, which is 285, over 210. And that is gonna give you 1.35714 and so on, so we're just gonna round that to 1.36. So the z-score is about 1.36, and the z-score is how many standard deviations from the mean, um, it is, so, or the value is.

Student K began her explanation very procedurally, going into great detail about how the number 1.36 was calculated. However, her ending statement that "the z-score is how many standard deviations from the mean … the value is" suggested a superficial conceptual understanding of a z-score.

Student M gave less procedural details than Student K, but still demonstrated superficial conceptual knowledge of z-scores.

> Student M: So to find this out, you want to start by calculating the z-score. So what you do is take the, this score that you're trying to account for, so, like, who scored higher than, um, a certain score. You subtract the mean from it. And then you divide all that by the standard deviation. What that tells you is how many standard deviations away from the mean you are.

In saying "what this tells you is how many standard deviations away from the mean you are," Student M provided a conceptual explanation of the z-score that went past purely procedural calculations.

With regard to both Student K and Student M, there is the possibility that in claiming that a z-score measures the number of standard deviations that a data point is from the mean, they were simply repeating a phrase that they had heard in class. Student G, on the other hand, provided a little more detail in her display of superficial conceptual knowledge.

> Student G: And the first step we need to take is to calculate the z-score. And the z-score is essentially the number of standard deviations something is from the mean. And on this scale, the mean is equal to 0 and a single standard deviation is equal to 1. So in order to find this value, we do 1344, our value, minus 1059, the mean, to get the difference between the two. And then we divide by our standard deviation of 210 to find the number of standard deviations we are away. And when we do this, we get 1.36.

Student K describes each step of the calculation in a little more detail, tying together the procedural steps and their conceptual meaning. She first describes subtracting the mean from the given value "to get the difference between the two" and then describes dividing by the standard deviation "to find the number of standard deviations we are away."

Level C: Deep Conceptual Knowledge

Finally, in order to demonstrate deep conceptual knowledge for the first component, students had to interpret the z-score calculation not just generally, but in the context of SAT scores. Student C provided this explanation in her oral assessment.

> Student C: You first need to find the z-score, which, so, you would do 1344 minus the mean of 1059, which gives you 285. And then you divide by the standard deviation of 210, which gives you 1.357, and then some more. And it rounds to 1.36.

At first, Student C appeared to demonstrate Level A knowledge in her description of the z-score, as after this statement, she went on to describe the shaded area under the curve. However, at the very end of her oral assessment, Student C ended with this statement:

> Student C: But that z-score of 1.36 means that students who scored a 1344 are 1.36 standard deviations away from the mean of 1059.

This one simple statement revealed a deep contextual and conceptual understanding of the meaning of the calculated value of 1.36.

Student E was more detailed in her description of the z-score, and infused context throughout her explanation.

> Student E: The person first started by finding the z-score, and the z-score is the calculation for the number of standard deviations a value is away from the

> mean. So to find the z-score, they started by taking the average SAT score of freshmen in the fall of 2020, which was 1344. And they subtracted from the mean SAT score of all SAT scores in 2019. [...] Then they divided that by 210, which is the standard deviation of all 2019 SAT scores. And that was approximately 1.36. And to reiterate, 1.36 means that the NC State University average SAT score of freshman was 1.36 standard deviations away from the mean 2019 SAT score, uh, of all people who took it.

Through her ability to continually refer back to the context of the question in her conceptual explanation of the z-score calculation, Student E demonstrated deep conceptual knowledge.

Student F also continually referred back to context in his description of the z-score calculation.

> Student F: Um, so to start out, they just worked on, um, finding the z-score, which is how many standard deviations away from the mean you are. Um, so to do that, first they found how far away from the mean they were, um, the average SAT of the admitted freshmen at NC State was, um, from the mean of all SAT scores. So that was 1344 minus 1059. And then they divide by 210, which is the standard deviation of SAT scores. So, the first part is finding how far away you are. Then, by dividing by the standard deviation, you get it in units of the standard deviation. Um, so they found that they were 1.36 standard deviations away from the mean, or they had a z-score of 1.36.

In addition to his strong attention to context, Student F also tied together the procedural steps and their conceptual meaning, just as Student G did. His statement that "by dividing by the standard deviation, you get [the z-score] in units of the standard deviation" portrays a deep conceptual understanding of how and why a z-score is a measure of the number of standard deviations a value is away from the mean.

Component #2: Description of Shaded Normal Density Curve

As described in Chapter 3, students' descriptions of finding the area of the shaded region on the normal density curve were classified as Level A (superficial procedural knowledge), Level B (deep procedural knowledge), or Level C (deep conceptual knowledge). Table 7 summarizes the results of these classifications, followed by an in-depth description of each level.

Table 7

Results from Component #2 of Oral Assessment #1

	Level A: Superficial Procedural Knowledge	Level B: Deep Procedural Knowledge	Level C: Deep Conceptual Knowledge
Number of Students	6	4	5

Level A: Superficial Procedural Knowledge

For the second component, superficial procedural knowledge was identified as a purely procedural description of calculating the area under a normal curve using a calculator (all students use TI-84 Plus CE graphing calculators in class) or a standard normal table (referred to as "Table A" in the AP Statistics curriculum). There was quite a bit of variation among students' demonstration of superficial procedural knowledge. For example, Student I and Student N gave the following explanations:

> Student I: And so that gets you the z-score of 1.36. And then you're able to put that into the bell curve and get 8.7, which you can do on the calculator.
>
> Student N: Then by using Table A, we find that only 8.7% of all SAT scores were higher than the NC State average mean score of 1344.

These two students did at least mention the tools that could be used to find the area under the normal distribution curve, but neither one described the procedure of how to use these tools to obtain the value 0.087.

Student D, on the other hand, attempted to describe the procedure of using Table A, going so far as to project a standard normal table on his video screen during his explanation.

> Student D: So then when they, when they graphed the bell curve, they drew a line at 1.36 z-scores above the mean of zero. And then [...] using Table A, they went to 1.36 and they looked for the decimal of the area, the decimal point that applies to the area for 1.36 z-scores above the mean. And 1.3 is here, somewhere [points to 1.3 in the left column of standard normal table on screen], 1.3 and 6 [points to 0.06 in the top row, then points to the cell with the intersection of the row and column], and this is roughly 8.7.

It is interesting to note that while Student D did correctly find the cell associated with a z-score of 1.36, that cell did not contain the number 0.087. Since the standard normal table he was using provides areas to the left of given z-scores, the value in the cell read 0.9131, not 0.087. Student D did not address this discrepancy.

Although they were classified as demonstrating Level A knowledge, neither Student I nor Student N nor Student D described the procedure of finding an area under a normal curve in enough detail for their procedure to be replicated. Student K, on the other hand, went into great detail during her procedural description of how to use a TI-84 Plus CE graphing calculator to find the area.

> Student K: And so we're gonna use our 1.36, and we are going to do on our calculator "2nd" then "vars" [...] then we're gonna press "2." Then we're gonna

> enter, um, in the lower area 1.36. And then for the upper, we're gonna give it a big value, and I'm just gonna give it a million because that's a big number, and then "paste." And that is gonna give you 0.0869 and so on. And then we're gonna take that number and we're gonna round it to 0.087. And then you're just gonna take the decimal and move it over two, and that is gonna give you 8.7%. And that is the answer. So approximately what percent of all SAT takers scored higher than 1344? Approximately 8.7% scored higher than 1344. So that's gonna be the answer.

While Student K did describe an in-depth procedure that could be replicated to find the correct answer, her description was purely procedural with no indication of any underlying conceptual understanding.

Despite the great variation among these student responses, all were deemed to demonstrate superficial procedural knowledge, with no indication of flexible understanding of the procedures in place.

Level B: Deep Procedural Knowledge

Baroody et al. (2007) suggest that students who exhibit flexibility in their procedures are demonstrating deep procedural knowledge as opposed to superficial procedural knowledge. Considering this distinction, students could demonstrate deep procedural knowledge on the second component by explaining the process of plotting the z-score and finding the area to the right (as opposed to the left) of that z-score. For example, Student C explained the following:

> Student C: And then on the graph, it's a drawing of the standardized normal distribution so the mean is at zero. And then you plot 1.36 and shade to the right because you're looking for what percent of SAT takers scored higher. Then you

do, um, to find the answer of 8.7% you do “2nd,” “vars,” “2” to get to “normalcdf.” And then you put the lower is 1.36, and the upper, you put, like, a big number. I put 10,000. And then “enter.” And then you get 0.0869, which is equal to 8.7% when rounded. Um, and then we know once we find 8.7% that, um, 8.7% of all SAT takers scored higher than a 1344.

While Student C described a similar calculator process as Student K, Student C demonstrated understanding of how this process was connected to the problem by explaining that the area was shaded to the right in order to find the area above a certain z-score.

Student H also demonstrated deep procedural knowledge in her explanation of how to appropriately use Table A to arrive at the correct answer.

> Student H: So what I did, I went to my Table A, which you can also use your graphing calculator, but I used Table A, and you will go to where is 1.3, and you have to go over to where your decimal place is. So ours is 1.36, so we’ll go over to 0.06. So 1.3 over to 0.06 is 0.9131. So since our normal distribution is the area that we’re trying to find is on the right and not the left, Table A only shows us the left side. So what we have to do is subtract that number we get from Table A from 1. So 1 minus 0.9131 is 0.0869. So roughly 8.7%. So the answer to our problem is approximately 8.7% of all SAT takers scored higher than a 1344.

Student H verbalized flexibility in her procedural knowledge in a way that Student D did not. It can be inferred that Student H knows that the entire area under a density curve is 1, and is able to use that knowledge to find areas above a certain z-score when given the area below a certain z-score.

Student L also demonstrated flexibility in her procedural knowledge in real-time, as evidenced by the following excerpt:

> Student L: And, uh, now we have our z-score, which we see on the graph. You can't see it, but it falls in between 1 and 2 obviously on, um, and we're trying to find what's, uh, the scores greater than that. So we're shading to the right [...] So because we have our z-score, we're able to plug that into Table A [...] So when we, when we find 1.3 and then we go to the hundredth spot, which is 6, so 1.36… um… perhaps I'm doing this wrong. [Student pauses video.] I'm not, I forgot that this shades to the left. So, uh, on the, on Table A, it will give us 0.9131, which is, um, 91.31%. But because they're shading to the left, we need to shade to the right, 'cause we're trying to find what's greater than, uh, 1344. Uh, you basically just subtract that from 1, or the 0.9131 from 1, and you get our, uh, a percentage for shading to the right, which is 8.7%.

While Student L did not demonstrate much conceptual knowledge in her explanation, she demonstrated deep procedural knowledge through her flexibility, particularly in her ability to identify her procedural mistake and correct it in the moment.

Level C: Deep Conceptual Knowledge

To demonstrate deep conceptual knowledge on this component students had to explicitly interpret the area under the normal density curve as the proportion or percentage of observations in the shaded region. Student M presents this interpretation immediately after calculating the z-score.

> Student M: And then after you find out how many standard deviations you are away from the mean, then you can calculate what percent of the bell curve is

accounted for, uh, in an area above or below that standard deviation away from the mean. So in this case, you resulted in a z-score of 1.36, which means that, um, you are 1.36 standard deviations above the mean. And then you can, uh, look at one of the, the charts we have [...] so there are charts in stats that you can use to calculate percentages based off of standard deviations you are away from the mean. So, in this case, we would be 1.36. So you find 1.3, then you go over to 0.6 [points to Table A]… hold it, I can't find it… okay. And then once you find yourself on the chart, then that's how you reach a conclusion of the percent roughly.

While Student M does not clearly exhibit procedural understanding of how to use Table A correctly (just like Student D, Student M does not acknowledge that the cell he points to does not contain the correct answer), his first sentence indicates a conceptual understanding of Table A. He is able to clearly state that Table A provides the "percent of the bell curve [that] is accounted for" in an area below or above a certain z-score.

Though Student M only explicitly states this conceptual understanding one time, Student O weaves this understanding throughout his explanation.

Student O: And since this, we've been told that this data is, has an approximately normal distribution, we can use our Table A to figure out what percent of the data would be above, uh, would be above 1.36 standard deviations from the mean. Now, normally you use this side [points to side of Table A with positive z-scores] 'cause it's a positive amount of standard deviations away. And if you go here, but… if you go here and you go to 1.36 standard deviations away on the positive side here, uh, it tells you that you are, that 91% of the data falls… below

> it. And of course, that's below it. So you could take 100 and subtract 91%, or this number, this four decimal number from it. But, uh also, you can know that since it's a uniform distribution [draws a bell curve in the air with hand], there's the same amount of data on one side as there is on the other. There's 50% on the left and 50% on the right from the middle. So you can instead flip over to this side [points to side of Table A with negative z-scores], the side that shows negative standard distributions away. And you could go to 1, negative 1.36 standard deviations away. And if you do that, you get your remainder of data, which is 8.69% of the data falls above it, because 8.69% falls below it. And it's the same on both sides. And so if you round your data up, you would get 8.7%

Although Student O makes a handful of vocabulary mistakes, such as referring to a "uniform" distribution instead of a "normal" distribution and referring to "standard distributions" instead of "standard deviations," his multiple references to the numbers representing the percentage of data above or below a certain z-score imply deep conceptual knowledge.

Written Assessment #1

After submitting their first oral assessment, students participated in a written assessment in the following class period. The first written assessment (included in full in Appendix B) asked students to answer three different questions using the distribution of SAT Math scores from the 2019 SAT Suite of Assessments Annual Report, showing that the scores were approximately normally distributed with a mean score of 528 and a standard deviation of 117. Students were instructed to clearly show their work as they answered the questions.

The first question, labeled as part (a), was identical in procedure to the problem provided in the first oral assessment - students were asked to calculate approximately what percent of all

SAT takers scored higher than a 630. Then, the last two questions, labeled as parts (b) and (c), required students to use slightly different procedures. In part (b), students were asked to calculate a percentile, requiring them to find the area to the left of a certain z-score as opposed to the right. In part (c), students were asked to find the percent of all SAT takers who scored in a certain range of score, requiring them to find the area between two z-scores. All three parts were assessed separately.

The same two major components that were analyzed in the students' videos were assessed in each part of the written assessment, but this time purely in terms of procedural proficiency. The first component assessed students' proficiency in calculating a z-score, and the second component assessed students' proficiency in finding the appropriate area under a normal density curve. As described in Chapter 3, for each component, students were rated as demonstrating low proficiency, medium proficiency, or high proficiency.

Component #1: Calculates Z-Score

As described above, the first component of the analysis was to assess students' proficiency in calculating z-scores. Students who did not attempt to calculate a z-score were classified as demonstrating low proficiency, students who incorrectly calculated a z-score were classified as demonstrating medium proficiency, and students who correctly calculated a z-score were classified as demonstrating high proficiency. Table 8 summarizes the number of students who demonstrated each level of proficiency of the first component for each part of the first written assessment.

Table 8

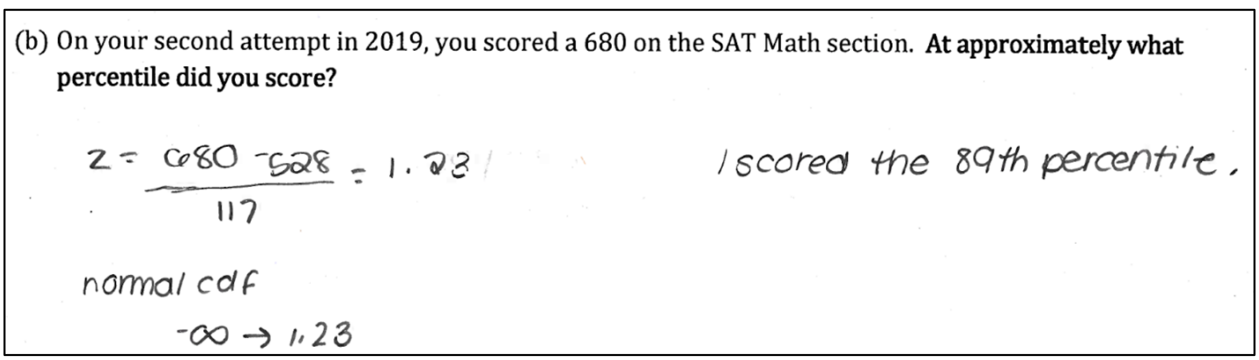
Results from Component #1 of Written Assessment #1

	Low Proficiency	Medium Proficiency	High Proficiency
Part (a)	0	0	15
Part (b)	0	1	14
Part (c)	2	1	12

In part (a), all fifteen students correctly calculated the appropriate z-score, demonstrating high proficiency in this component. In part (b), Student J was the only student who incorrectly calculated the z-score, but her error was in the computation, not the set-up, as shown in Figure 5.

Figure 5

Student J, Written Assessment #1, Part (b)


(b) On your second attempt in 2019, you scored a 680 on the SAT Math section. **At approximately what percentile did you score?**

$z = \frac{680 - 528}{117} = 1.23$

I scored the 89th percentile.

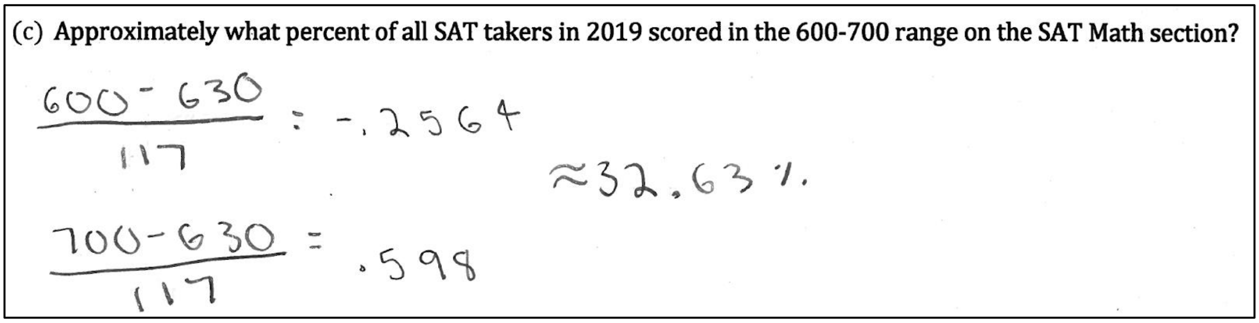
normal cdf

$-\infty \rightarrow 1.23$

In part (c), students had to calculate two z-scores in order to find the proportion of observations in a bounded range of scores. Neither Student K nor Student A attempted to calculate z-scores on this part, while Student G was the only student who incorrectly calculated these z-scores, using the wrong value for the mean score, as shown in Figure 6.

Figure 6

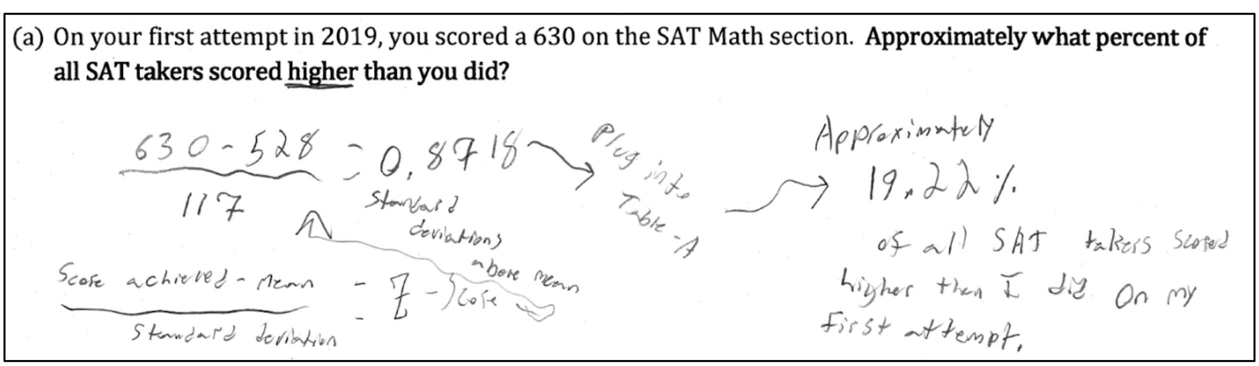
Student G, Written Assessment #1, Part (c)



As shown in Table 8, all students demonstrated high proficiency in calculating a z-score in at least two parts of the first written assessment, and the vast majority of students demonstrated high proficiency in calculating a z-score in all three parts. As shown in Figure 7, Student O did an especially thorough job of showing and explaining his calculation of the z-score in part (a). He included annotations indicating the general formula for a z-score as well as a label indicating that his numerical value measured the number of standard deviations above the mean.

Figure 7

Student O, Written Assessment #1, Part (a)



Component #2: Finds Area Under Normal Curve

The second component of the analysis was to assess students' proficiency in finding a desired area, or percentage, under a normal density curve. Students who labeled and/or shaded the normal curve incorrectly were classified as demonstrating low proficiency, students who included a correct sketch of the normal curve but provided an incorrect percentage were classified as demonstrating medium proficiency, and students who provided the correct percentage, with or without a sketch of the normal curve, were classified as demonstrating high proficiency. Table 9 summarizes the number of students who demonstrated each level of proficiency of the second component for each part of the first written assessment.

Table 9

Results from Component #2 of Written Assessment #1

	Low Proficiency	Medium Proficiency	High Proficiency
Part (a)	2	0	13
Part (b)	3	3	9
Part (c)	2	2	11

In part (a), all but two students correctly provided the area to the right of their calculated z-score. There was variability in the amount of work shown on this question among students who demonstrated high proficiency. Student L sketched a standardized normal curve (as shown in Figure 8), Student K sketched a non-standardized normal curve (as shown in Figure 9), and Student J did not sketch a curve at all (as shown in Figure 10).

Figure 8

Student L, Written Assessment #1, Part (a)

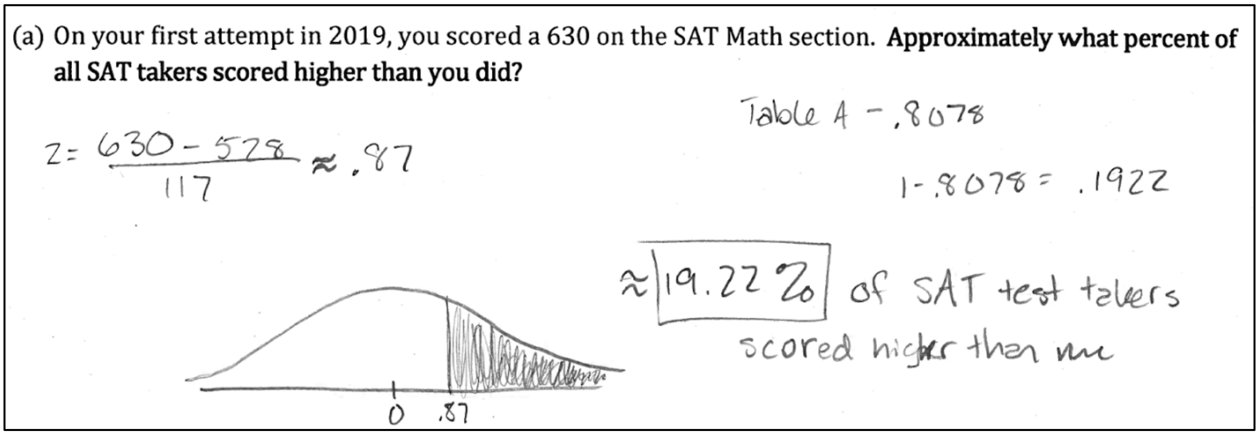


Figure 9

Student K, Written Assessment #1, Part (a)

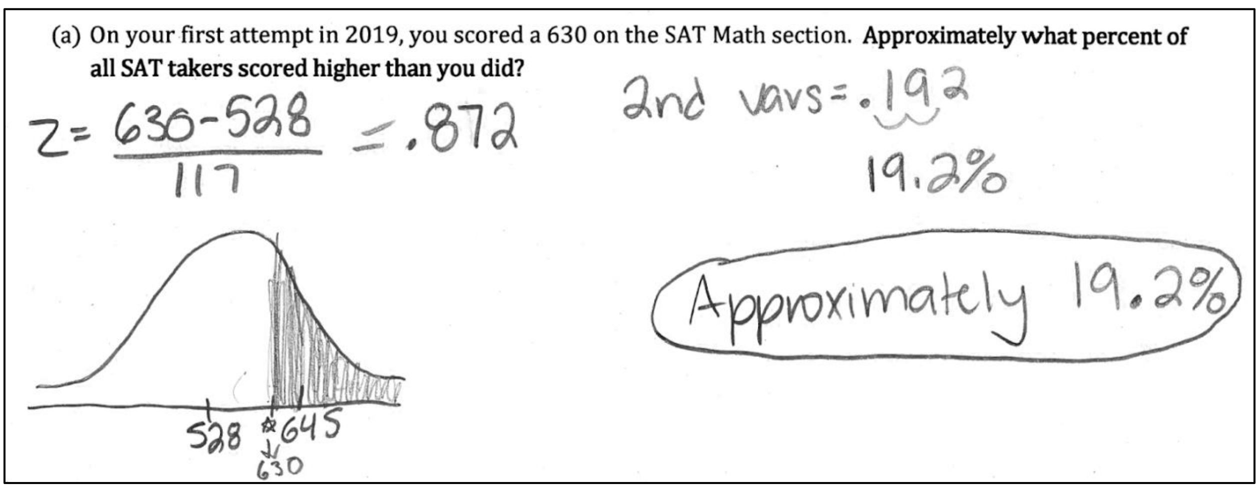
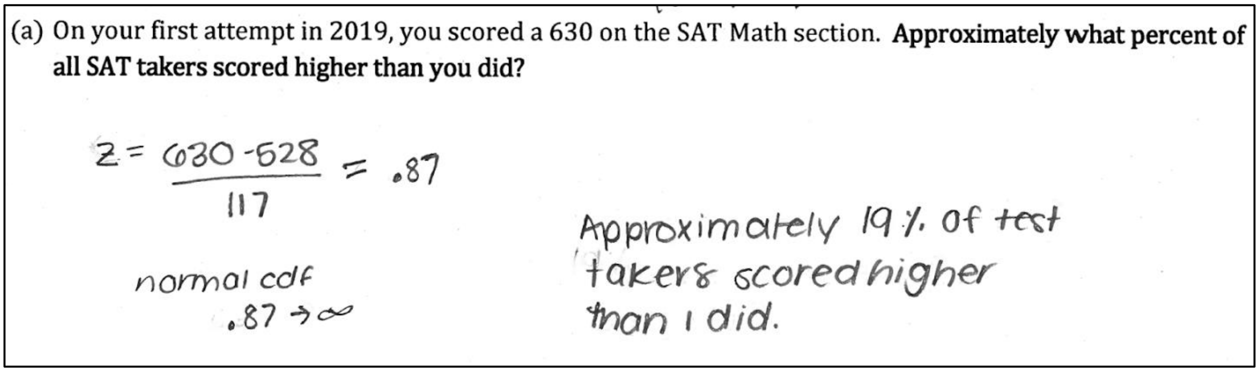


Figure 10

Student J, Written Assessment #1, Part (a)



Two students, Student A and Student B, demonstrated low proficiency in calculating the area under the normal distribution curve in part (a). Student A correctly drew a non-standardized normal curve, but did not translate that to a correct drawing of the standardized normal curve, as shown in Figure 11. Student B, on the other hand, showed no work and simply provided an incorrect percentage with no indication of how she arrived at the answer, as shown in Figure 12.

Figure 11

Student A, Written Assessment #1, Part (a)

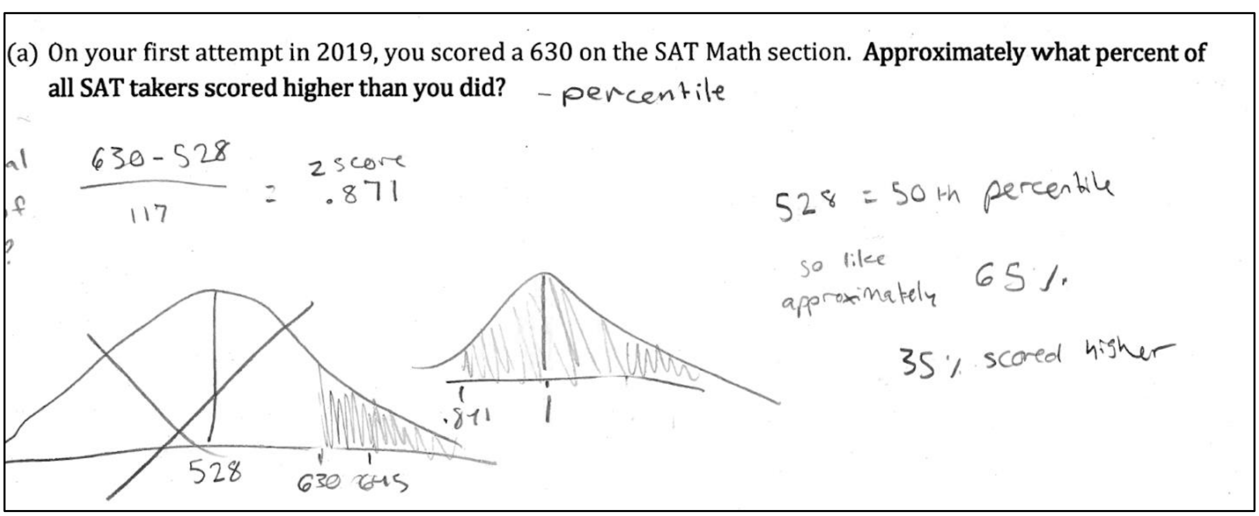
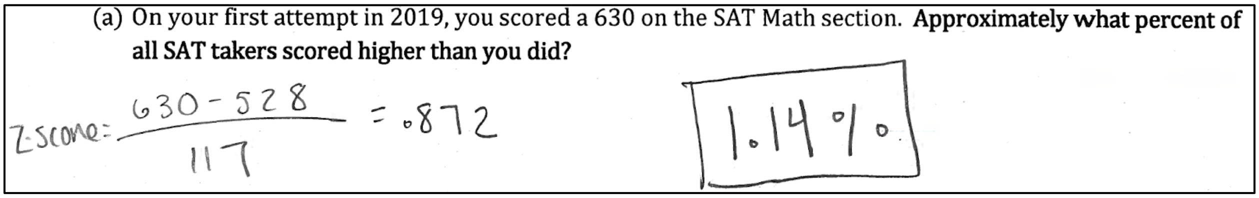


Figure 12

Student B, Written Assessment #1, Part (a)



As stated previously, part (b) asked students to calculate a percentile, requiring them to both know the definition of a percentile (which had been covered in class) and know how to find area bounded above by the z-score (shading to the left) rather than bounded below by the z-score (shading to the right). Nine students calculated this percentile correctly, but again there was variation in the work shown. Student F drew a standardized normal curve and calculated the area to the left of the calculated z-score, as shown in Figure 13. Student E, on the other hand, drew a standardized normal curve and calculated the percentage to the right of the calculated z-score. Then, she seems to have subtracted this percentage from 100% (though no work is shown) to arrive at the correct percentile, as shown in Figure 14.

Figure 13

Student F, Written Assessment #1, Part (b)

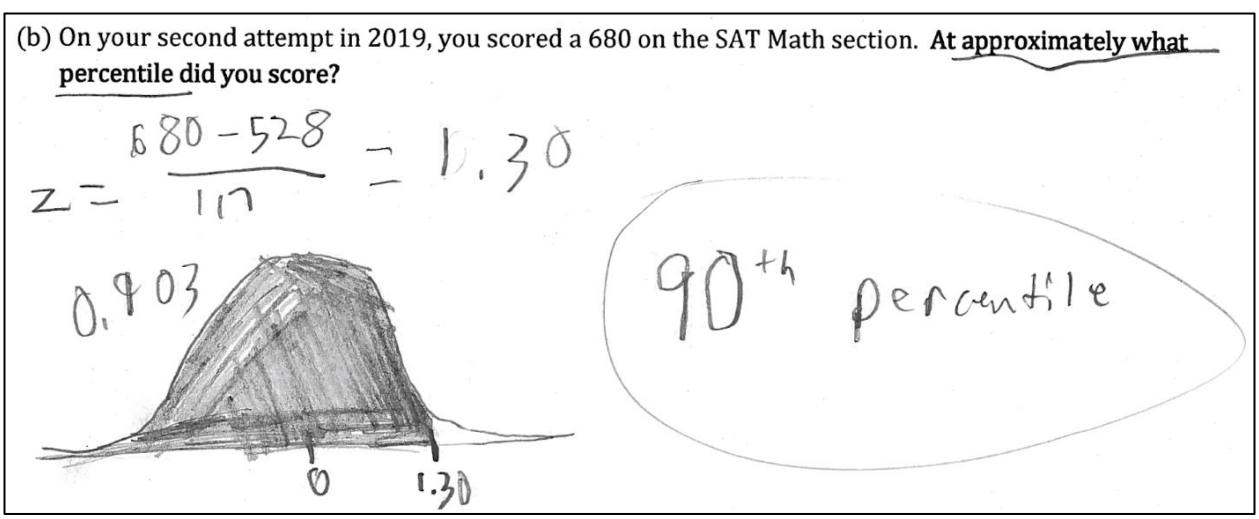
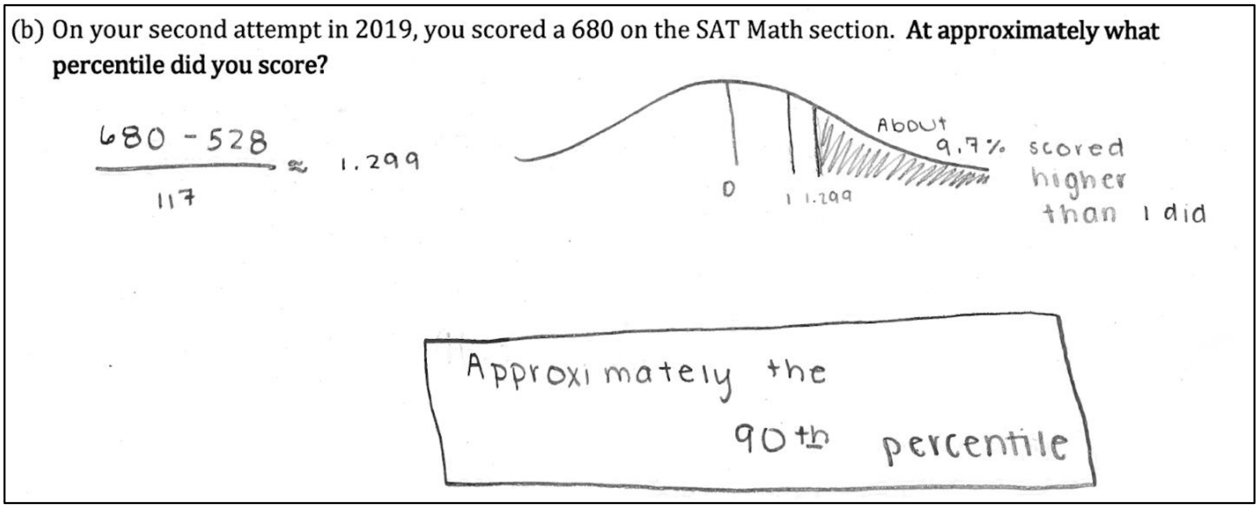


Figure 14

Student E, Written Assessment #1, Part (b)



Three students demonstrated medium proficiency on the second component of part (b). Student H calculated the area above the z-score as opposed to below the z-score, implying that she did not know or recognize the definition of a percentile. Student L and Student B, on the other hand, both seem to have known the definition of percentile, as indicated by their correct shaded areas on the normal curve. Student L correctly shaded a standardized normal curve (as shown in Figure 15), and Student B correctly shaded a non-standardized normal curve (as shown in Figure 16). However, neither Student L nor Student B provided a percentage associated with their shaded areas.

Figure 15

Student L, Written Assessment #1, Part (b)

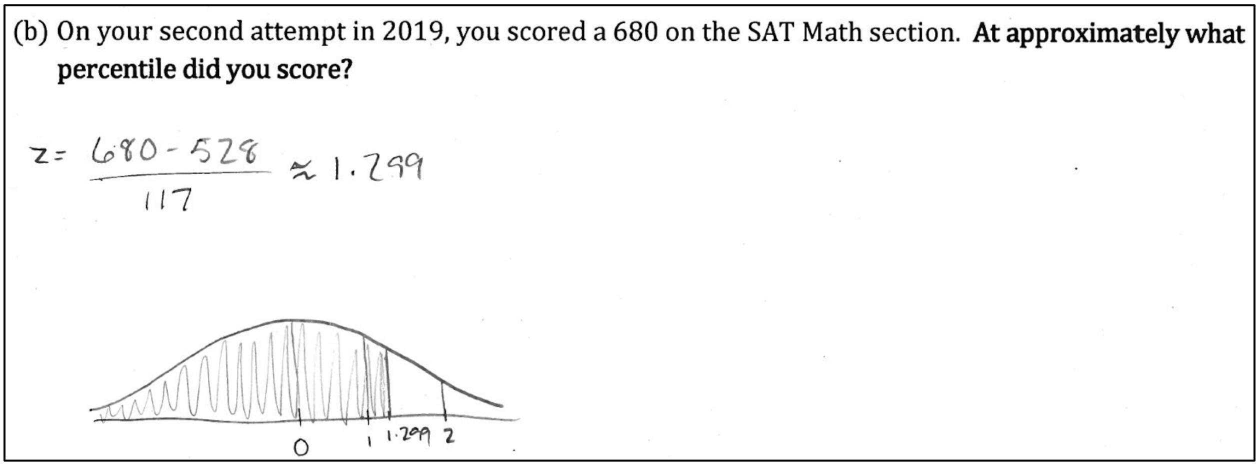
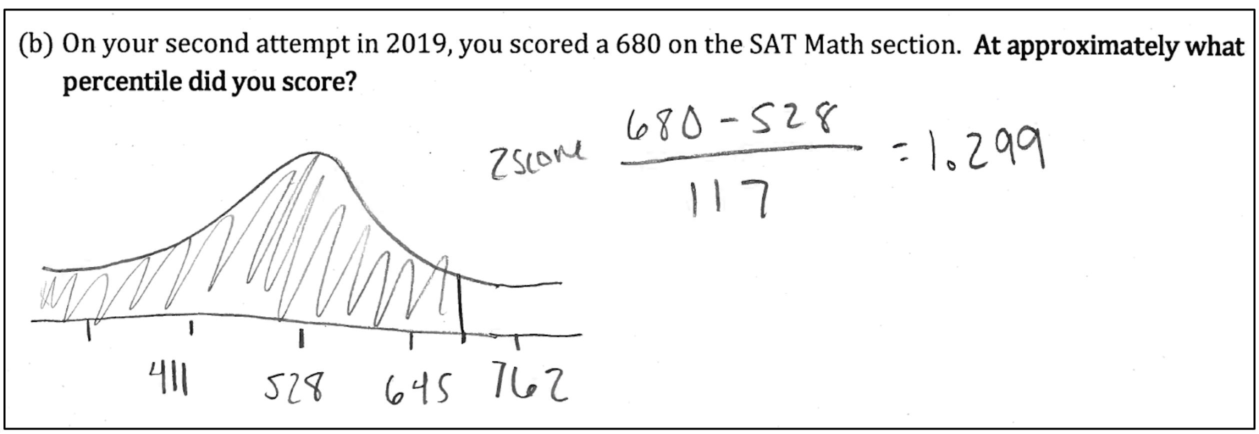


Figure 16

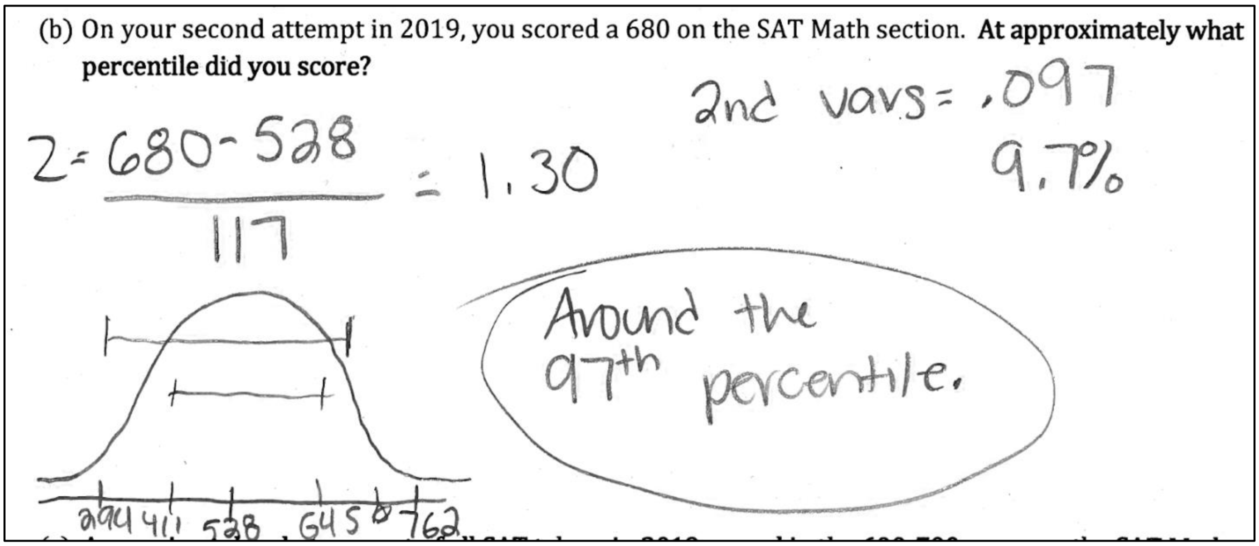
Student B, Written Assessment #1, Part (b)



Three students demonstrated low proficiency on the second component of part (b). Interestingly, although Student K did not shade any portion of the normal curve, she appears to have correctly calculated the area above the appropriate z-score, just as Student E did (see Figure 17). Then, however, instead of subtracting this percentage from 100% as Student E appears to have done, Student K appears to have increased her percentage by a factor of 10, to arrive at an answer of the 97th percentile. Student A and Student C, on the other hand, did not correctly calculate the area above or below their calculated z-scores.

Figure 17

Student K, Written Assessment #1, Part (b)



Finally, in part (c), students were asked to calculate the percentage of all SAT takers who scored between 600 and 700 on the SAT Math section. This required students to find the area between two z-scores on the standardized normal curve. Eleven students demonstrated high proficiency on the second component of part (c). Student C showed how she used her calculator to arrive at the correct answer (see Figure 18) and Student D showed how he used Table A to arrive at the correct answer (see Figure 19).

Figure 18

Student C, Written Assessment #1, Part (c)

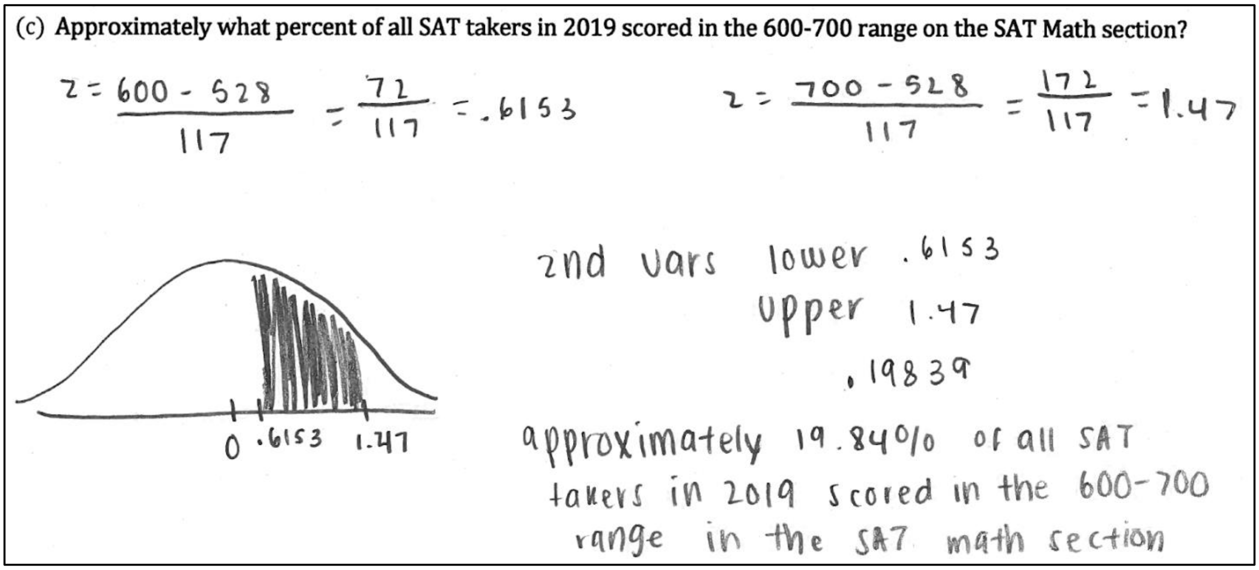
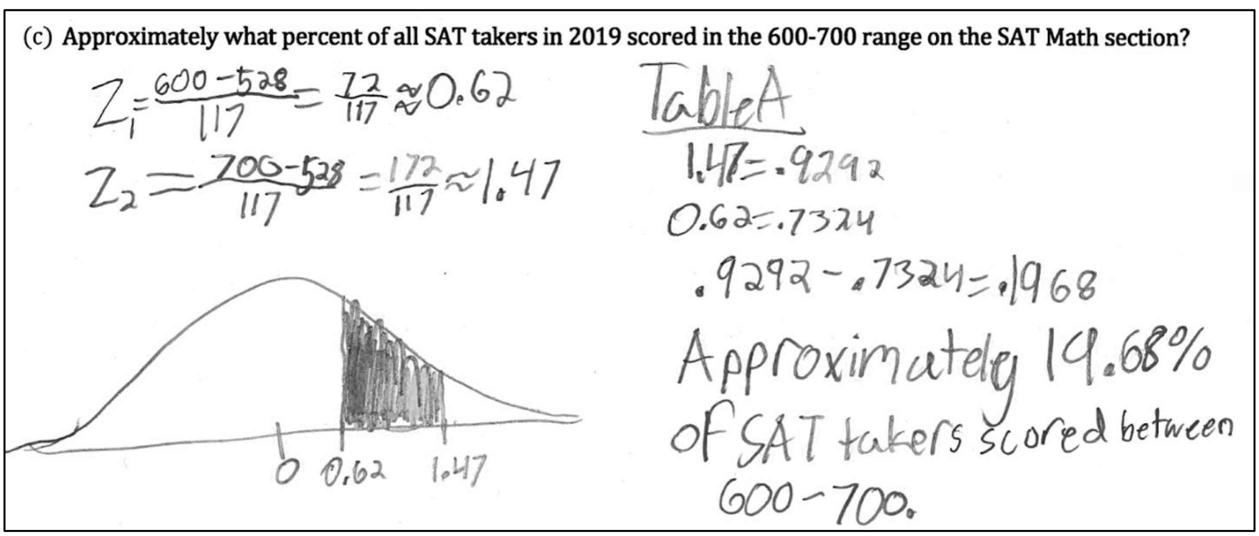


Figure 19

Student D, Written Assessment #1, Part (c)



Student A was one of the two students who demonstrated medium proficiency, as she correctly shaded the area corresponding to the correct percentage, but was unsuccessful in calculating that area (see Figure 20). Student H demonstrated low proficiency, as she did not clearly indicate what area she was looking for. That being said, it is interesting to note that

Student H has the correct answer written on her paper, though she does not claim it as her final answer (see Figure 21).

Figure 20

Student A, Written Assessment #1, Part (c)

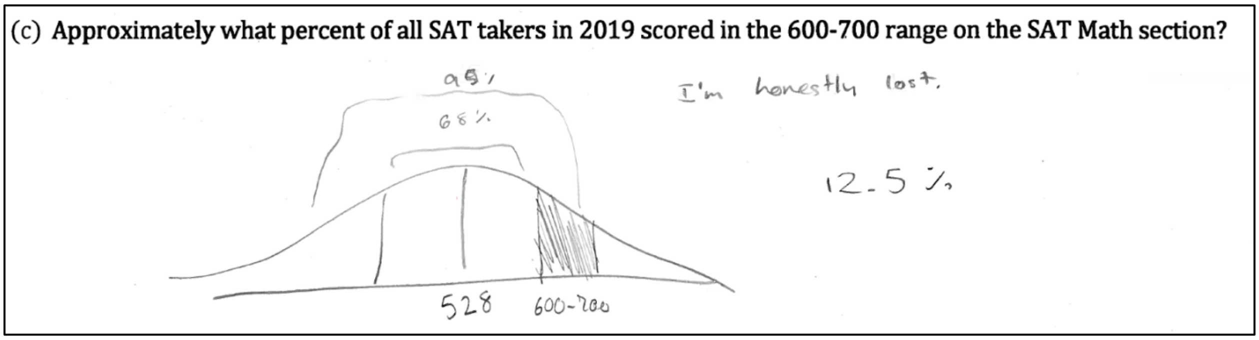
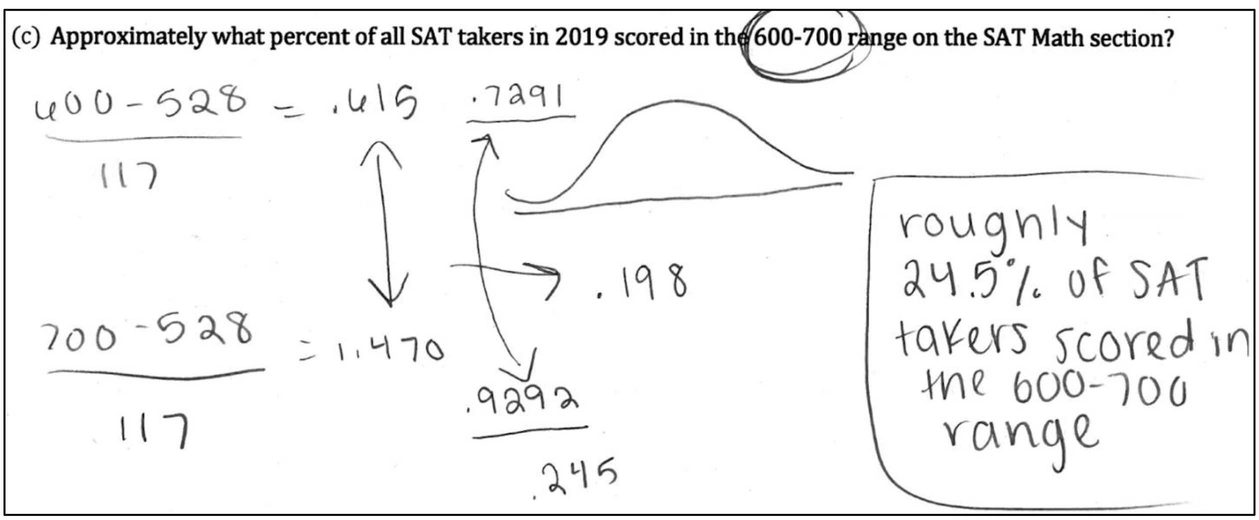


Figure 21

Student H, Written Assessment #1, Part (c)



Comparison of Oral Assessment #1 and Written Assessment #1

After analyzing the oral and written assessments separately, students' oral assessment classifications were compared with their written assessment results. Again, each overarching component, z-scores and normal distribution curves, was analyzed separately.

Component #1: Calculates Z-Scores

To begin the analysis, the distribution of students' proficiency on z-score calculations on each part of the written assessment was compared among the three levels of students' demonstrated knowledge about z-scores on the oral assessment. Figure 22, Figure 23, and Figure 24 show the distributions for part (a), part (b), and part (c), respectively.

Figure 22

Distribution of Component #1 Proficiency Ratings, Written Assessment #1, Part (a)

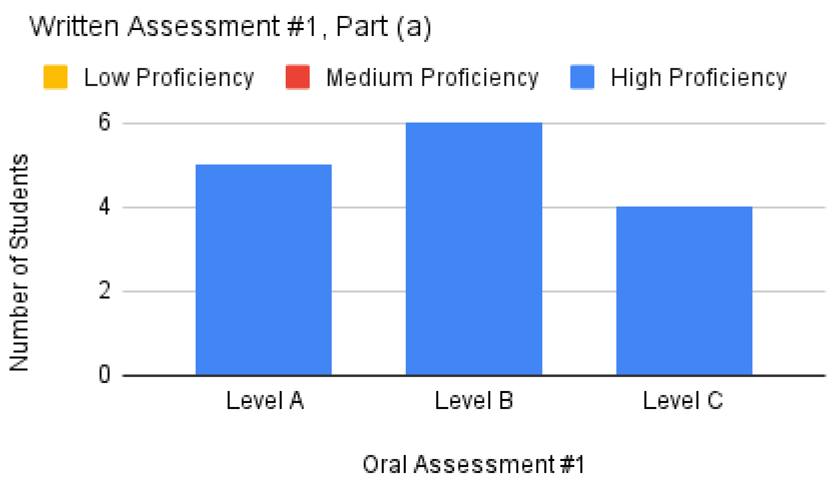


Figure 23

Distribution of Component #1 Proficiency Ratings, Written Assessment #1, Part (b)

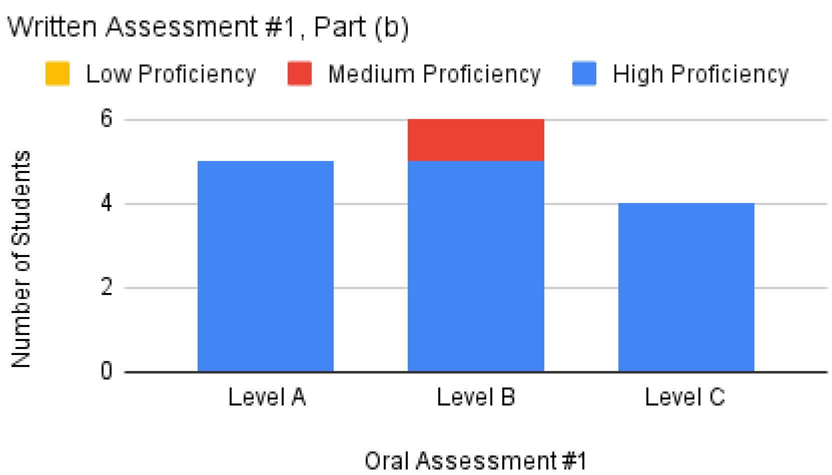
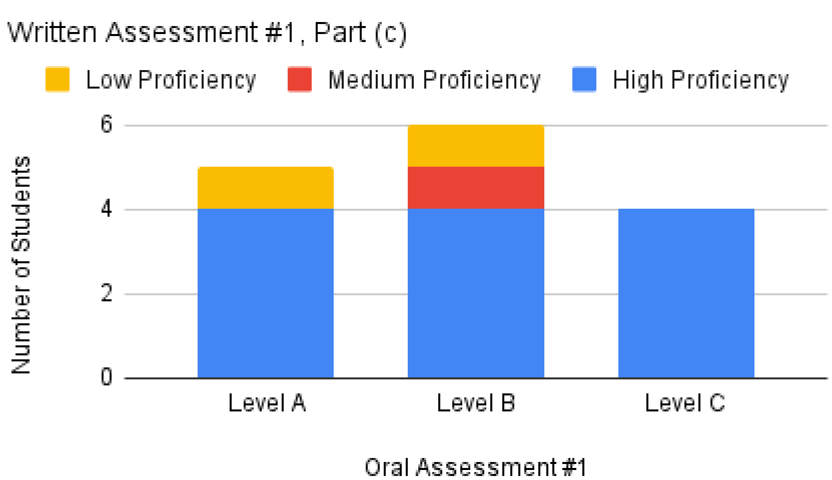


Figure 24

Distribution of Component #1 Proficiency Ratings, Written Assessment #1, Part (c)



Firstly, it is important to note that the overwhelming majority of students demonstrated high proficiency on z-score calculations on the written assessment, regardless of level of knowledge demonstrated on the oral assessment. Still, it is interesting to note that all four students who demonstrated Level C knowledge (deep conceptual knowledge) on the oral assessment demonstrated high proficiency on all three parts of the written assessment.

Component #2: Finds Area Under Normal Curve

Next, the distribution of students' proficiency on finding area under a normal curve on each part of the written assessment was compared among the three levels of students' demonstrated knowledge about finding area under a normal curve on the oral assessment. Figure 25, Figure 26, and Figure 27 show the distributions for part (a), part (b), and part (c), respectively.

Figure 25

Distribution of Component #2 Proficiency Ratings, Written Assessment #1, Part (a)

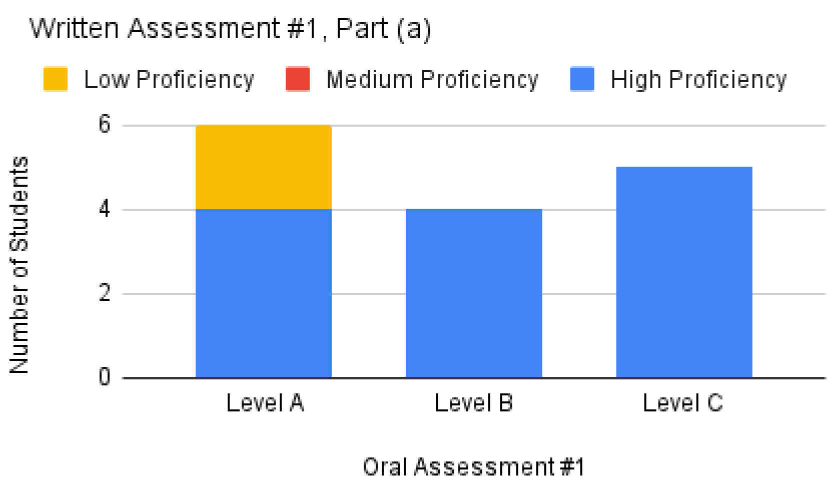


Figure 26

Distribution of Component #2 Proficiency Ratings, Written Assessment #1, Part (b)

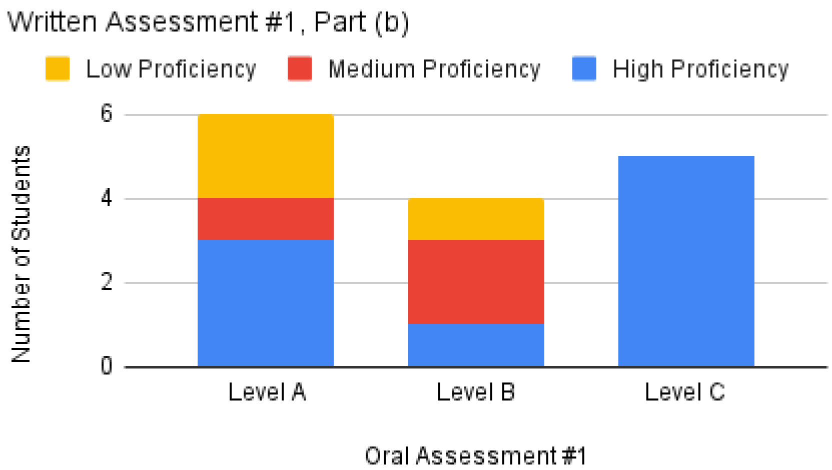
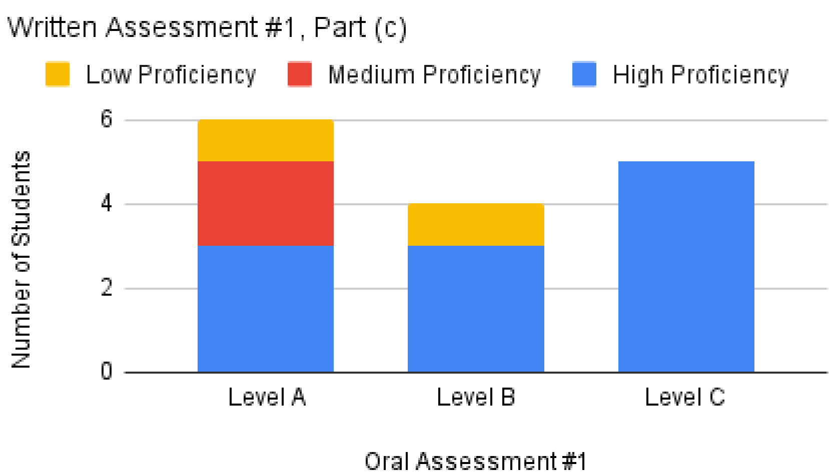


Figure 27

Distribution of Component #2 Proficiency Ratings, Written Assessment #1, Part (c)



Again, it is important to note that the majority of students demonstrated high proficiency on finding the area under normal curves on the written assessment. However, it appears that level of knowledge is more closely associated with proficiency level for normal distribution calculations than it is for z-score calculations. Especially in parts (b) and (c), when the questions were not identical in procedure to the question given on the oral assessment, students who demonstrated Level C knowledge (deep conceptual knowledge), were more likely to demonstrate high proficiency on finding the area under a normal curve than those students who demonstrated Level A or Level B knowledge.

Oral Assessment #2

The second oral assessment (included in full in Appendix A) provided students with the same data as in the first oral assessment, showing that SAT scores in 2019 were approximately normally distributed with a mean score of 1059 and a standard deviation of 210. The question of interest was as follows: To be in the top 10% of all test takers, approximately what score do you need to earn?

A worked-out solution to this question was presented to students, who again had the following task: "Create a *Flipgrid* video describing how this solution was reached. Explain each step (how it was done and what it means) as clearly as possible." The solution presented to students was composed of two major parts. The first part was a diagram of the standardized normal curve with the top 10% of the area shaded in. The second part was an algebraic conversion of the corresponding z-score to an SAT score. As such, two major components were analyzed in the students' videos: their description of obtaining the appropriate z-score based on the area of interest and their description of converting the standardized z-score back to an SAT score. Again, it is important to note that the analysis is based on the knowledge elicited from students in their verbal explanations, which is not necessarily indicative of all a student may know about the topic.

Component #1: Description of Obtaining Z-Score

As described in Chapter 3, a student's description of obtaining the appropriate z-score was classified as Level A (superficial procedural knowledge), Level B (deep procedural knowledge), or Level C (deep conceptual knowledge). Table 10 summarizes the results of these classifications, followed by an in-depth description of each level.

Table 10

Results from Component #1 of Oral Assessment #2

	Level A: Superficial Procedural Knowledge	Level B: Deep Procedural Knowledge	Level C: Deep Conceptual Knowledge
Number of Students	3	10	2

Level A: Superficial Procedural Knowledge

For the first component, superficial procedural knowledge was identified as a purely procedural description of obtaining the z-score using either a calculator or Table A. All three students who demonstrated Level A knowledge failed to provide a procedural description that could be replicated by others. For example, Student I gave the following explanation:

> Student I: And so in order to find the z-score, they drew a graph and then they filled in the top 10%, calculated that, and they got 1.28.

It is unclear what Student I meant by "calculated that," and whether she knows how to find the z-score using either a calculator or Table A.

Student B also gave a purely procedural description without much detail, as shown in this excerpt:

> Student B: We were not given, um, our x-value to plug into our z-score equation because we're not solving from a score to see what percentage you're in. We're solving from the percentage to see what score you would have, basically. So the first thing you do is you draw your graph, and obviously it would be on the right

> side because the higher the score the better, so right side. And then you're gonna plug it into your calculator using "second", "vars," and then the inverse. And that will give you your z-score of 1.28.

While Student B does begin by acknowledging the conceptual difference between this problem and the one presented on the first oral assessment, her description of finding the z-score of 1.28 is not very specific. The meaning of "plug it in your calculator" does not clearly communicate the process of obtaining a z-score from a given area under a normal distribution curve.

Level B: Deep Procedural Knowledge

Students who demonstrated deep procedural knowledge made the connection between the area under the normal curve and the corresponding z-score more clear in their explanations. Student N was brief but explicit in identifying this connection.

> Student N: Since we know we are looking for the top 10%, we can use our calculator to find the z-score. We use "inverse norm" with an area of 0.9 to the left, because that would leave 10% of the area to the right. This gives us a z-score of 1.28.

Student E gave a more detailed, yet similar, explanation as compared to Student N, as shown in the following excerpt:

> Student E: So the first thing that this person did, um, to find the score to be in the top 10% of all test takers is they drew a normal distribution graph and they found, uh, where the top 10% would be. And they shaded that. Um, and as you can see, it is on the right, and the area of that shaded area is 0.1. [...] So to figure out the z-score, uh, you would use inverse normal calculations on the calculator. And because we know that the area of the shaded part is 0.1, the person would input

> 0.1 and then highlight “right” because we know that 10% is on the right side of the normal distribution graph. And that comes out to be a z-score of 1.28.

In this description, Student E is clearly explaining what she is entering into her calculator and how those inputs correspond to the shaded area on the normal distribution graph.

Student O opted not to use technology to find the z-score, but instead described how it could be done using Table A.

> Student O: So to set up this one, we already have the percent. We’d have to find the z-score that would correlate with it. Uh, with our A Table, you’d have to go… you’d have to go to the negative side because it gives what’s below. And it’s the same on each side. So you can go to the negative side, and you have to find the number closest to 0.1 as possible, which in this case would be negative 1.28. Which you would convert to being positive, ‘cause you’re finding the top part. So that would be the z-score that you’d use.

While Student O fails to make explicit references to “area” at or below a certain z-score, his flexible use of Table A demonstrates deep procedural knowledge.

Level C: Deep Conceptual Knowledge

Only two students demonstrated what was considered deep conceptual knowledge, which required that they interpret the area of 0.1 as representing the proportion of data above the z-score on a normal distribution curve. Student F does so in the following excerpt:

> Student F: They start off by drawing the curve. It’s just a normal curve. And then they shade in the top 10 percent-ish. Um, so they just shade in to the right of the curve. Um, ‘cause that’s where the high values are. They shade in around 10%. Um, and they’ll put the actual values later when they know them. Um, so

> in order to find the score, first you have to find, um, what z-score the 10% will be at, um, 10% will be above. Um, so in order to do that they, um, just plug in with their calculator and use "inverse norm." Um, and if you had, and they would plug in an area of 0.1, um, since 10% is the same thing as 0.1 out of 1. Um, and they'd have the tail, they'd have it face to the right since they shaded to the right. [...] Um, so when they do that they get a z-score around 1.28. [...] And then they have a z-score, um, which means they can, you know, put where the line actually is to shade in after the 10%.

In the midst of his procedural explanation of how to draw the curve and plug corresponding values into the calculator, Student F clearly states that "you have to find [...] what z-score the [...] 10% will be above." While it may have been an even stronger description if Student F had stated that 10% of all data or observations would be above that z-score, his explanation was sufficient to demonstrate deep conceptual knowledge of obtaining a z-score from a corresponding area.

Student J also demonstrated deep conceptual knowledge, as shown in the following excerpt:

> Student J: So in the previous problem [...] we were trying to find the proportion of, um, of values that were above it or were below it. But this time it's the inverse of it. They're giving us a proportion and we're trying to figure out at what value do you have to be for that proportion to be true. So in this case, in order for… for your proportion to be 10%, or 0.10, what value do you have to be? [...] So we went from percentage to z-score because we used, since we were given an area, we used the inverse… "inverse norm" function, which is the opposite of "normalcdf." Since we're… we start off with an area instead of finding the

area. So using that, it gives us a z-score of 1.28.

Student J really emphasizes the relationship between the first and second oral assessment problems - that instead of finding the proportion of values that were above or below a certain value, she is now finding the value that has a certain proportion of values above or below it.

It is interesting to note that both Student F and Student J, the only two students who were classified as demonstrating deep conceptual knowledge, both appeared to engage in a productive struggle while explaining their thinking. Their explanations contain many pauses and interjections, and their thinking is exhibited in "real time."

Component #2: Description of Converting from Z-Score to SAT Score

As described in Chapter 3, students' descriptions of converting from the standardized z-score to an SAT score wer classified as Level A (superficial procedural knowledge), Level B (superficial conceptual knowledge), or Level C (deep conceptual knowledge). Table 11 summarizes the results of these classifications, followed by an in-depth description of each level.

Table 11

Results from Component #2 of Oral Assessment #2

	Level A: Superficial Procedural Knowledge	Level B: Superficial Conceptual Knowledge	Level C: Deep Conceptual Knowledge
Number of Students	2	13	0

Level A: Superficial Procedural Knowledge

For the second component, superficial procedural knowledge was identified as a purely procedural description of converting from z-score to SAT score with no acknowledgment of what the numbers represented, namely with no reference to the mean or standard deviation of the SAT score distribution. Student H was one of two students who provided the following cursory description:

> Student H: So we do 1.28 equals X minus 1059 over 210. So we just carry that out, which gives us X minus 1059 equals 268.8. So then we just subtract from here. And then, so we add to here, sorry. And so we get 1327.8. So to be in the top percent of 10% of all test takers, you must earn a score of approximately 1330 or higher, which is just us rounding.

Student B went a little further than Student H in acknowledging that the equation came from the formula for a z-score, but her explanation still did not indicate any understanding of what the numbers represented.

> Student B: And that will give you a z-score of 1.28. And then you're gonna set that equal to the equation to find your z-score, which will give you 1.28 equals X minus 1059 over 210. And then you're gonna multiply both sides by 210, and then add 1059 to both sides. And then that is how they got their score of 1330.

Baroody et al. (2007) describe superficial procedural knowledge as "superficial step-by-step knowledge" or "disembodied procedural knowledge." Both Student H and Student B provide step-by-step procedures, but their procedures are completely disembodied from the recognition of the numbers as the mean and standard deviation of the distribution of SAT scores.

Level B: Superficial Conceptual Knowledge

Students could demonstrate superficial conceptual knowledge by "embodying" their procedures, namely in identifying the numbers they were using in their calculation as the mean and standard deviation of the distribution of SAT scores. All of the thirteen remaining students did this in some form, though there was variability in both the length and clarity of their explanations. Student D described the procedure as follows:

> Student D: And then if you take 1.28 and then plug in the others. So you have the z-score equals X minus the mean divided by the standard deviation. So in order to solve for X, you have to multiply by the standard deviation. So 1.28 times 210 gives us 268.8, which they have here. And then to further solve for X, you have to add 1059 to both sides, which when we do that, we get 1327.8.

While Student D does not explain why he is choosing to use this equation, he at least references that he is using the mean and standard deviation in his calculation.

Student G goes into slightly more detail about why she is using this equation to solve for X, as seen in the excerpt below:

> Student G: So we'll go with 1.28 as, um, as our z-score for that exact point. And then we just need to plug it into our equation, which would be the z-score we need minus the mean, and we already know that's 1059. So we subtract X, which is a variable since we don't know it. And we subtract X minus 1059 over the standard deviation of 210, which will give us how many standard deviations away if we were working it out that way and we were solving for the z-score. But we have it. We just need to solve for X, which first we multiply both sides by 210. We get 268.8 equals X minus 1059. And then we add that 1059 back to get X equals 1327.8. And that is our answer.

Student G goes a bit deeper than Student D when she explains that the z-score "will give us how many standard deviations away we are." However, she does not clearly state how "solving for X" results in the desired SAT score.

Level C: Deep Conceptual Knowledge

None of the fifteen students demonstrated deep conceptual knowledge of converting from a standardized z-score into the desired SAT score. To demonstrate this level of knowledge, students were expected to explain the meaning behind the steps of their calculations. In particular, the researcher was listening for an explanation that since a z-score is the number of standard deviations that a value is away from the mean, multiplying the z-score by the standard deviation of the distribution of SAT scores results in the total number of points that the student's test score is away from the mean test score. In other words, deep conceptual knowledge could be displayed by interpreting the value 268.8 as the number of points a student would need to score above the mean in order to be in the top 10% of all test takers. As no student attempted to interpret this number in context, no student was classified as demonstrating deep conceptual knowledge.

Written Assessment #2

After submitting their second oral assessment, students participated in a second written assessment in the following class period. The second written assessment (included in full in Appendix B) again provided students with the distribution of SAT Math scores from 2019, which were normally distributed with a mean score of 528 and a standard deviation of 117. Students were given two different questions relating to this distribution and once again were instructed to clearly show their work as they answered the questions.

The first question, labeled as part (a), was identical in procedure to the problem provided in the second oral assessment - students were asked to calculate the lowest score they would have needed to earn in order to score in the top 5% of all test takers. In part (b), students were asked to calculate the interquartile range of SAT Math scores in 2019. Students had previous experience calculating the interquartile range from a discrete set of data points and from a boxplot, but they had never before calculated the interquartile range of normally distributed data. Part (a) and part (b) were assessed separately.

The same two major components that were analyzed in the students' videos were assessed in each part of the written assessment, but again, only in terms of procedural proficiency. The first component assessed students' proficiency in obtaining the appropriate z-score from a given area under the normal distribution curve, and the second component assessed students' proficiency in converting from a standardized z-score into an SAT Math score. As described in Chapter 3, for each component, students were rated as demonstrating low proficiency, medium proficiency, or high proficiency.

Component #1: Obtains Z-Score

As stated, the first component of the analysis was to assess students' proficiency in obtaining a z-score corresponding to a certain location on the normal distribution curve. Students who labeled and/or shaded a normal distribution curve incorrectly were classified as demonstrating low proficiency, students who had a correct sketch of a normal distribution curve but did not obtain the correct z-score were classified as demonstrating medium proficiency, and students who obtained the correct z-score (with or without a sketch of the normal distribution), were classified as demonstrating high proficiency. Table 12 summarizes the number of students who demonstrated each level of proficiency of the first component for each part of the second written assessment.

Table 12

Results from Component #1 of Written Assessment #2

	Low Proficiency	Medium Proficiency	High Proficiency
Part (a)	0	2	13
Part (b)	5	0	10

In part (a), all students were able to at least draw an accurate normal distribution curve with the appropriate shaded area. However, two students were not able to use their diagram to aid them in obtaining the correct z-score, as shown in Figure 28 and Figure 29, resulting in a demonstration of medium proficiency.

Figure 28

Student M, Written Assessment #2, Part (a)

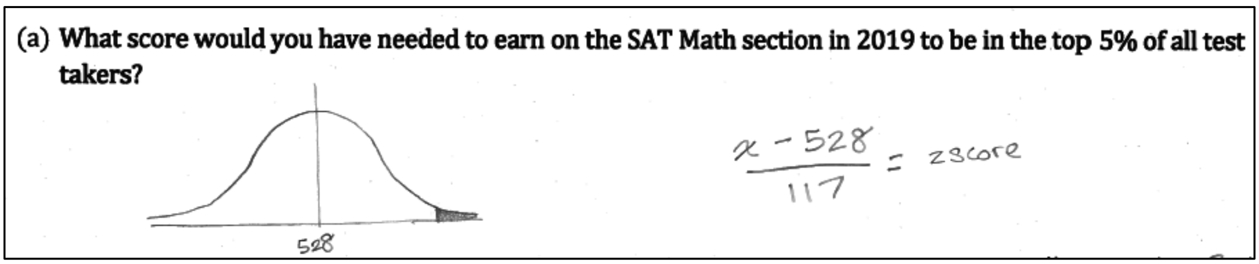
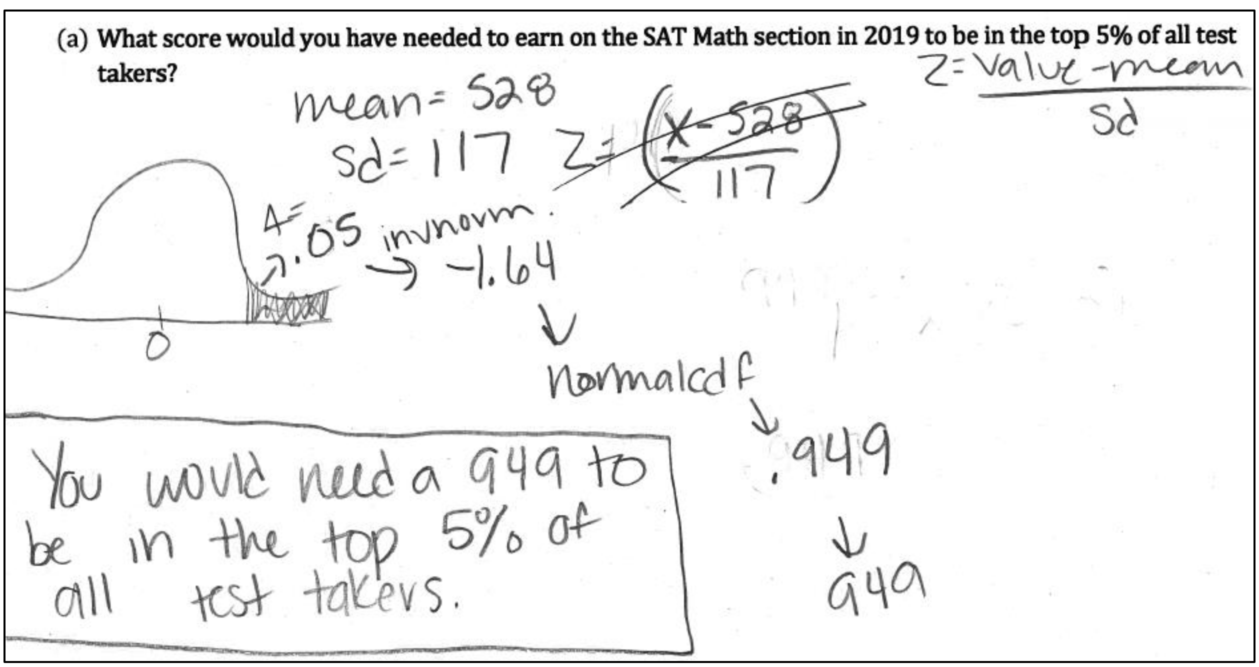


Figure 29

Student K, Written Assessment #2, Part (a)



Student M drew an unstandardized normal curve and shaded a small region located in the right-hand tail of the curve. He then set up a correct z-score equation, but he was unable to find the specific z-score associated with his shaded area.

Student K, on the other hand, drew a standardized normal curve, and clearly labeled that she had shaded in an area of 0.05 in the right-hand tail of the curve. Student K then indicates that she used the calculator function "InvNorm" to find a z-score of negative 1.64. This is the z-score that corresponds to the 5th percentile, not the 95th percentile. It is interesting to note that while Student K's diagram is correct, she provides a z-score that would clearly lie below the mean, not above the mean as indicated by her diagram.

Of the students who demonstrated high proficiency of this component in part (a), there was a range of work shown. Student H did not demonstrate her process of finding the z-score other than annotating her diagram (as shown in Figure 30), Student C wrote down the steps that she performed on her calculator (as shown in Figure 31), and Student O acknowledged that either Table A or a calculator could be used to find the z-score, and annotated his diagram to draw attention to the fact that the lower bound of the shaded area was located "1.64 standard deviations away" (as shown in Figure 32).

Figure 30

Student H, Written Assessment #2, Part (a)

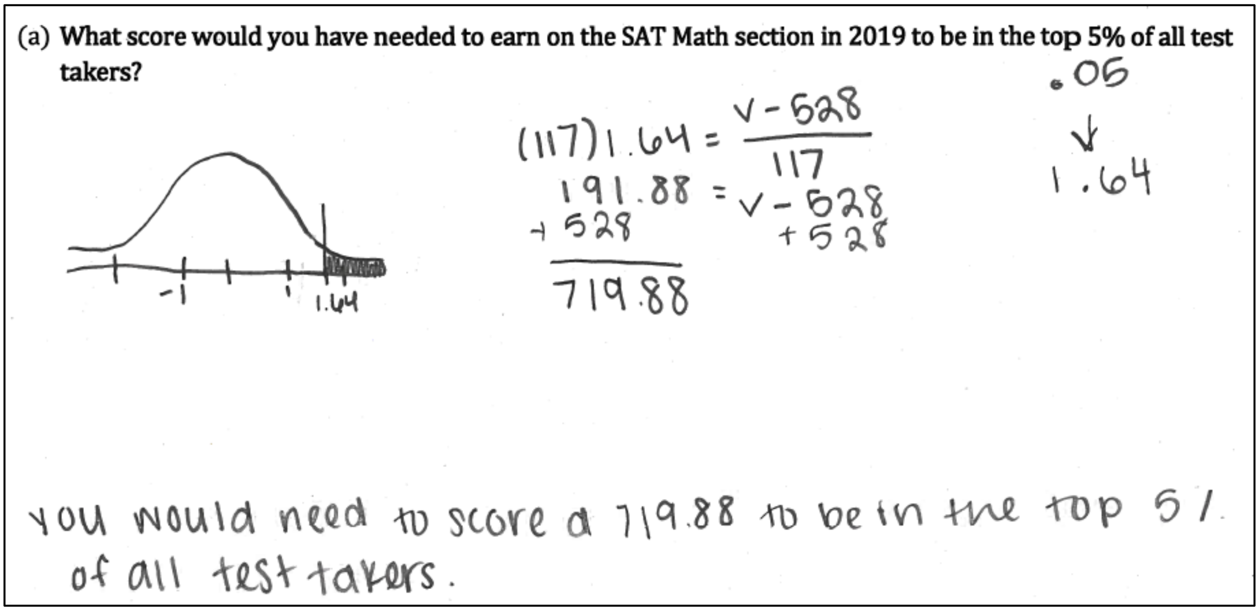


Figure 31

Student C, Written Assessment #2, Part (a)

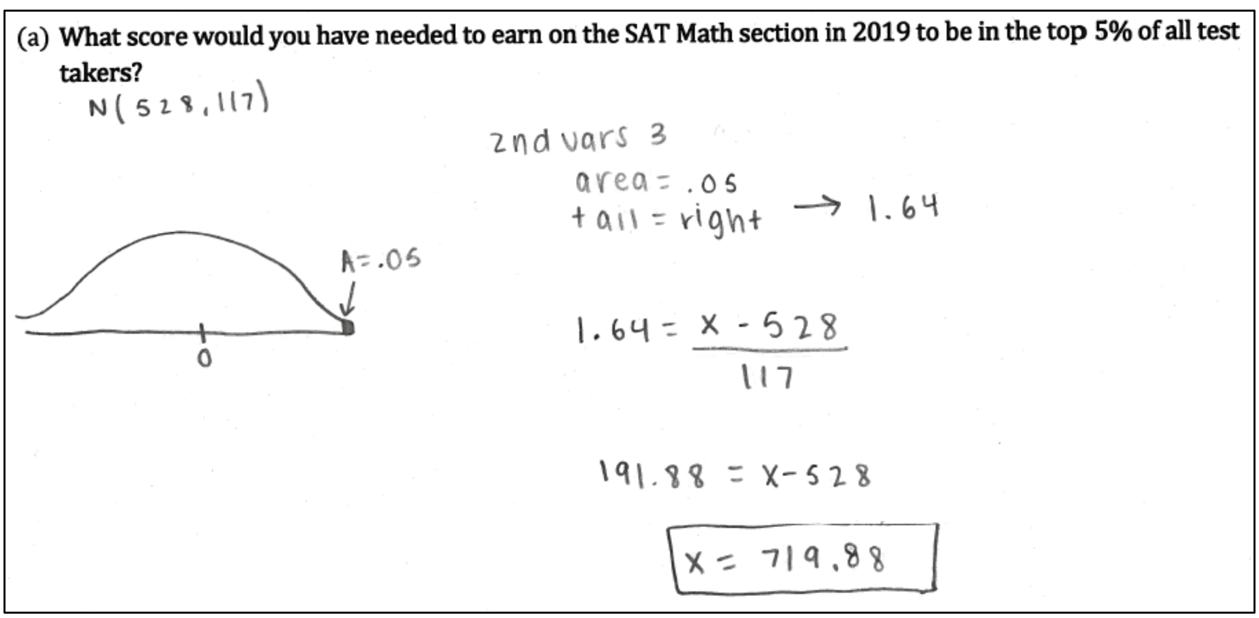
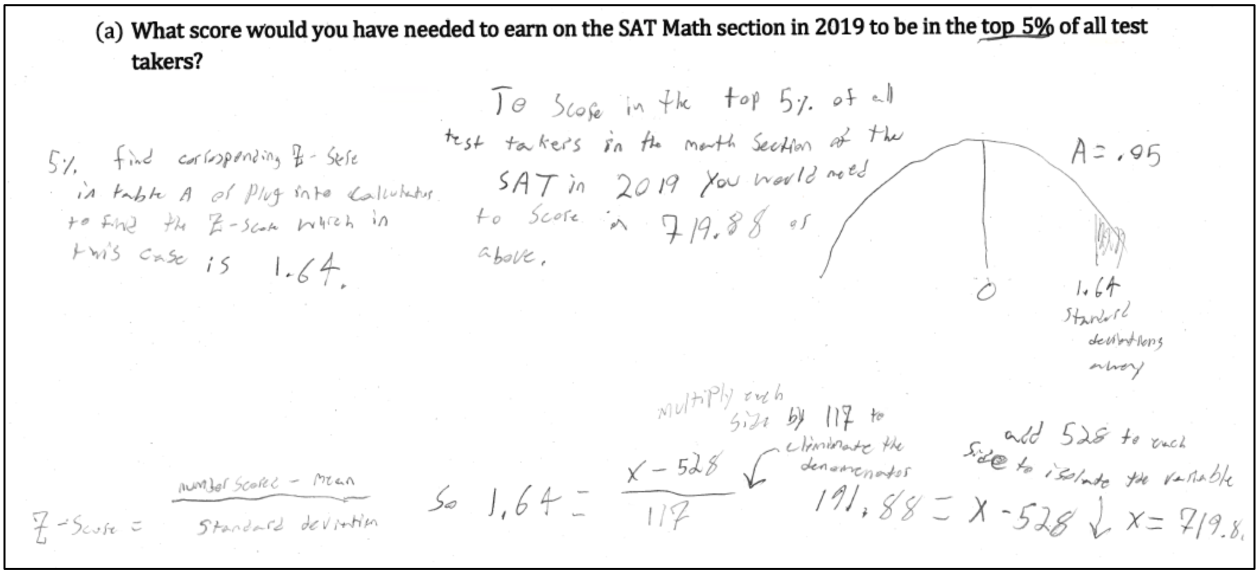


Figure 32

Student O, Written Assessment #2, Part (a)

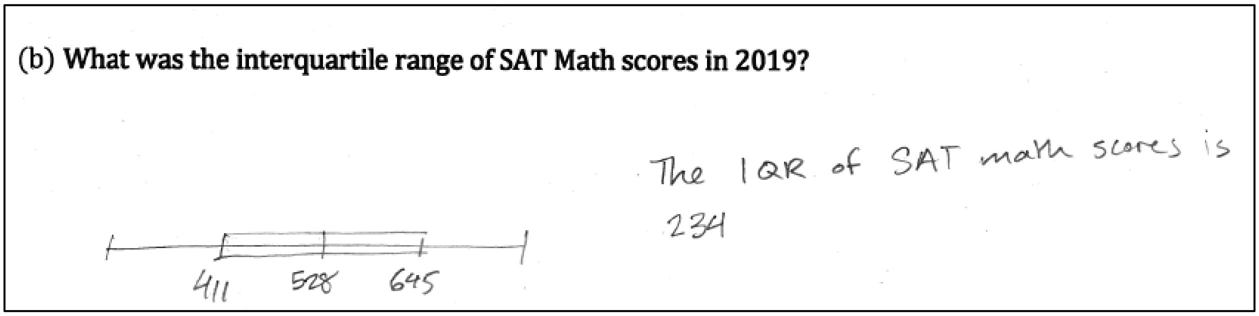


In part (b), students were asked to find the interquartile range of SAT Math Scores in 2019. Five students were unable to correctly shade and label a normal distribution curve, demonstrating low proficiency on this component.

As stated previously, students had previous experience with calculating the interquartile range of data summarized by a boxplot, but they had never been prompted to calculate the interquartile range of data summarized by a normal distribution. As shown in Figure 33, student M attempted to draw a boxplot to summarize the SAT Math score data, and while he did demonstrate understanding that the mean and median of symmetric distributions are roughly equal, he did not demonstrate understanding that the first quartile and third quartile are not one standard deviation away from the mean for normally distributed data.

Figure 33

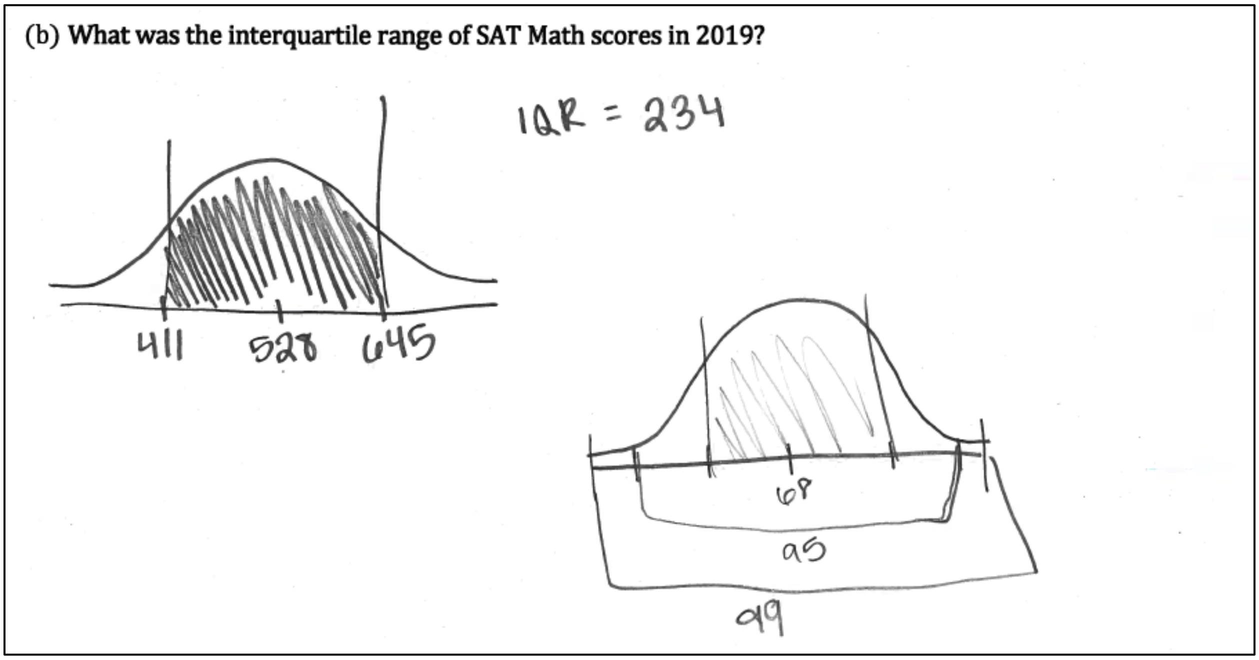
Student M, Written Assessment #2, Part (b)



As shown in Figure 34, Student H arrived at the same incorrect conclusion as Student M did. She also calculated the range of values within one standard deviation of the mean and declared that to be the interquartile range. Based on her second diagram, however, this student seemed to notice that this area should correspond to the middle 68% of the data. It is not clear from her work if this student realized she should have been looking at the middle 50% of the data when calculating the interquartile range.

Figure 34

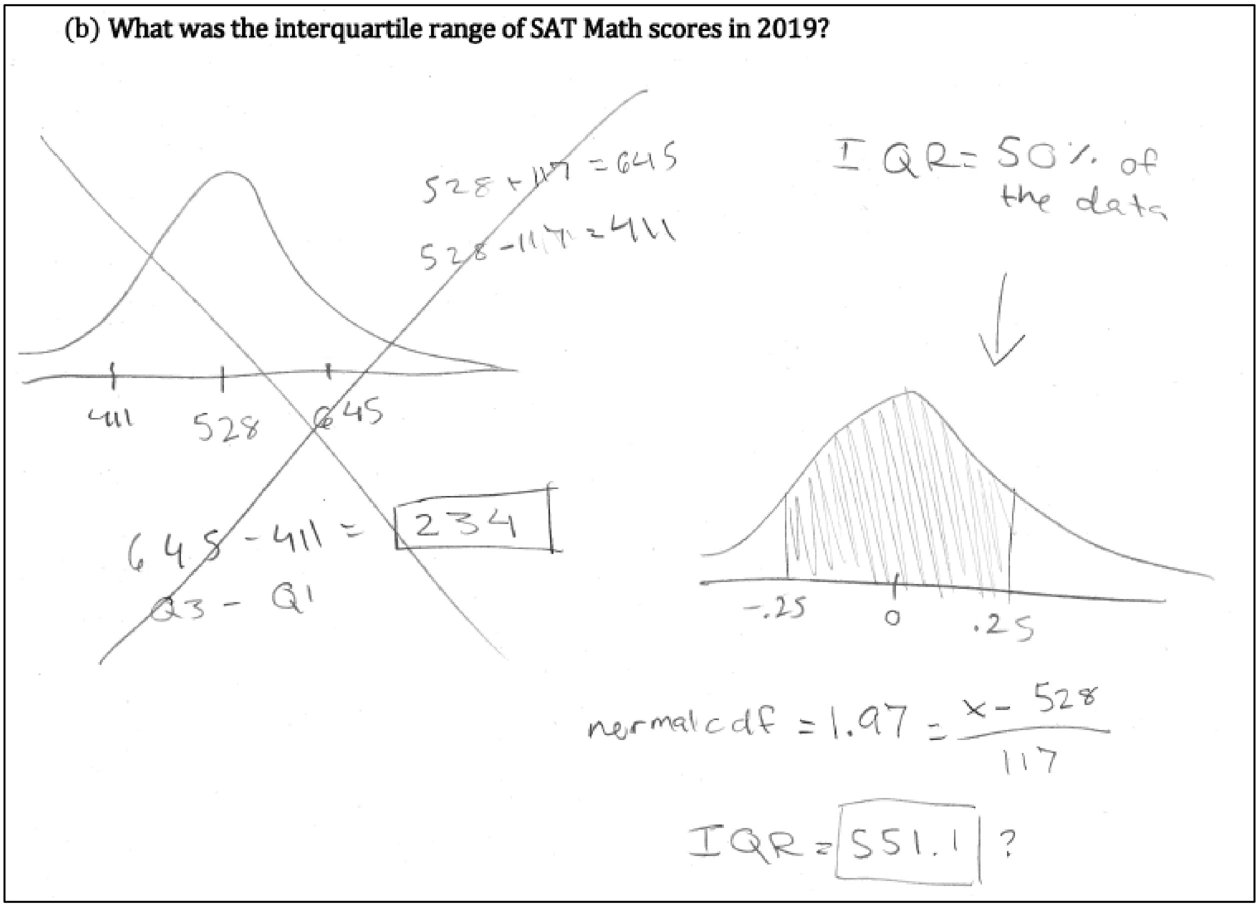
Student H, Written Assessment #2, Part (b)



As shown in Figure 35, Student A seems to have started with the same incorrect reasoning as Student M and Student H, as revealed in her crossed out work on the left side of her paper. However, Student A corrected herself and clearly demonstrated her understanding that the interquartile range is the range of the middle 50% of all data points. That being said, Student A still demonstrated low proficiency on this component by incorrectly labeling the bounds of her shaded region as -0.25 and 0.25.

Figure 35

Student A, Written Assessment #2, Part (b)



Ten students were able to demonstrate high proficiency by obtaining the correct z-scores, and they did so while showing a variety of work. Student D shaded in the middle 50% of the normal curve and then indicated that he was looking for the corresponding z-scores with areas of 0.25 and 0.75 below them, as shown in Figure 36. Student L, on the other hand, drew attention to the symmetry of the normal distribution, showing that 25% of the data is between the first quartile and the mean, and another 25% of the data is between the mean and the third quartile, as shown in Figure 37.

Figure 36

Student D, Written Assessment #2, Part (b)

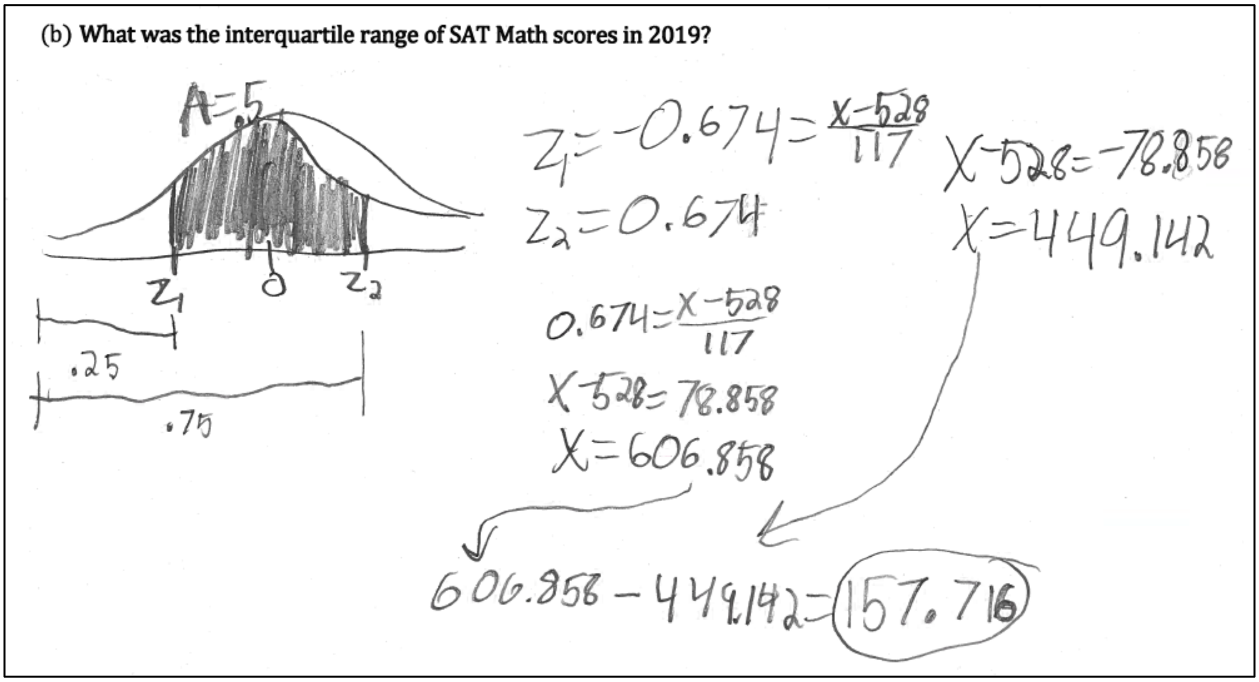
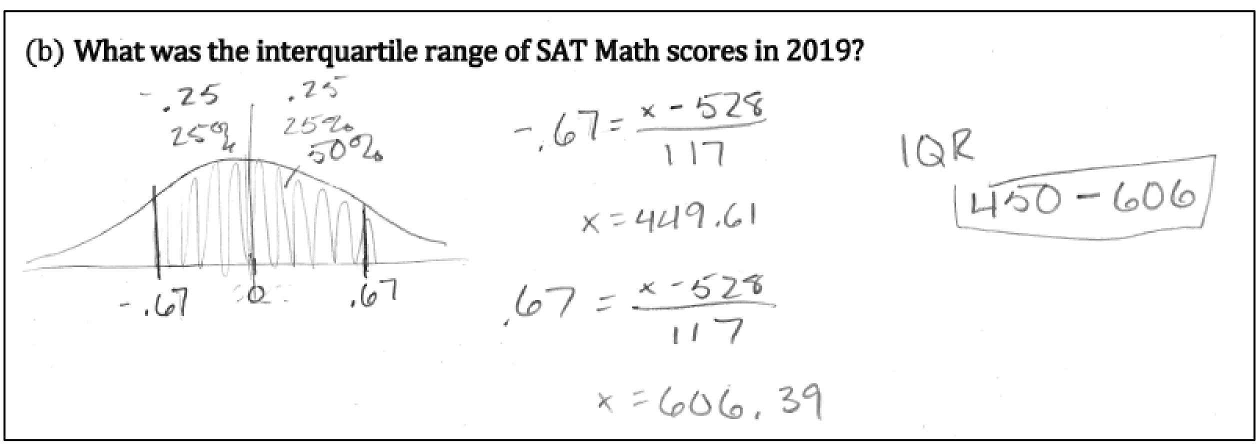


Figure 37

Student A, Written Assessment #2, Part (b)



Component #2: Converts Z-Score

The second component of the analysis was to assess students' proficiency in converting from a standardized z-score to an SAT Math score. Students who did not attempt to set up an equation were classified as demonstrating low proficiency, students who set up an incorrect equation or arrived at an incorrect value were classified as demonstrating medium proficiency, and students who arrived at the correct value were classified as demonstrating high proficiency. Table 13 summarizes the number of students who demonstrated each level of proficiency of the second component for each part of the second written assessment.

Table 13

Results from Component #2 of Written Assessment #2

	Low Proficiency	**Medium Proficiency**	**High Proficiency**
Part (a)	2	0	13
Part (b)	4	1	10

The same two students who demonstrated medium proficiency on the first component of part (a) ended up demonstrating low proficiency on the second component of part (a). Since Student M did not arrive at any sort of z-score (see Figure 28), he was unable to receive credit for converting the z-score into an SAT score. Student K, on the other hand, did arrive at z-score (see Figure 29), and could have received high proficiency on this component had she used this z-score to calculate an SAT score. Based on her annotations, however, it appears as if Student K calculated the area to the right of her incorrect z-score of negative 1.64 using the calculator function "normalcdf." Then, Student K multiplied this value by 1000 to result in her final answer of 949. Student K did have the correct z-score equation template written and then crossed out on her paper, but she does not seem to have attempted to use the equation.

The remaining thirteen students who found the correct z-score were able to perform the algebra to convert that standardized score into an SAT Math score. Some students, such as Student B, showed no work after setting up the equation (see Figure 38), and others, such as

Student N, showed some intermediate steps (see Figure 39), but all thirteen students arrived at a reasonable answer.

Figure 38

Student B, Written Assessment #2, Part (a)

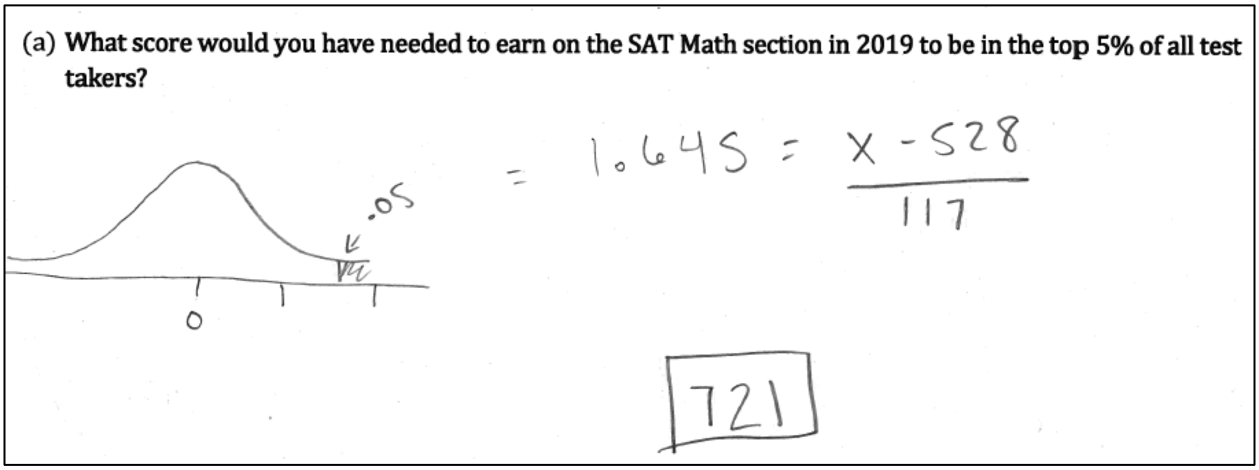
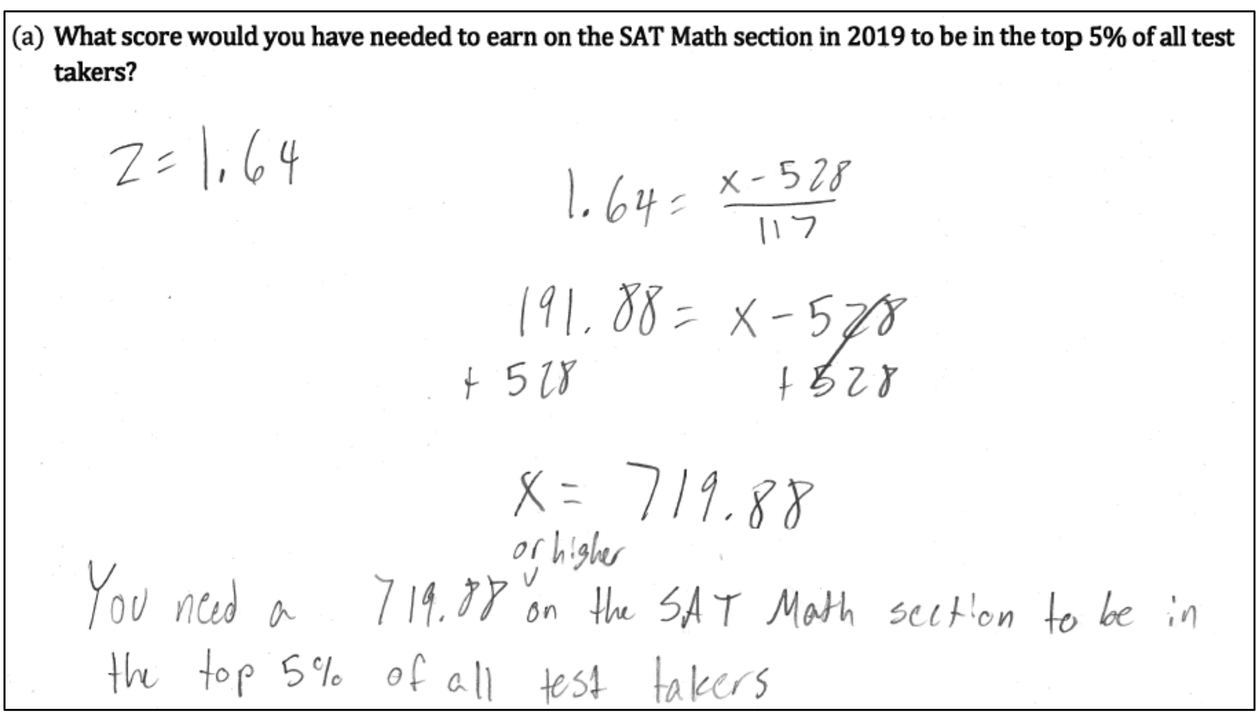


Figure 39

Student N, Written Assessment #2, Part (a)



Similarly as in part (a), the same five students who demonstrated low proficiency on the first component of part (b) demonstrated low or medium proficiency on the second component of part (b). Four of the five students, including Student M and Student H, demonstrated low proficiency as they could not even attempt to calculate a z-score when they had not found a z-score in the first place (see Figure 33 and Figure 34). Student A, however, received medium proficiency on this component, as she at least set up a z-score equation, albeit using an incorrect z-score (see Figure 35).

The same ten students who demonstrated high proficiency in finding the correct z-scores were able to demonstrate high proficiency in converting those z-scores to SAT Math scores and providing the interquartile range. Again, students showed a variety of work. Student J set up the correct equations but did not show her work in solving them (see Figure 40), whereas Student F showed some of his intermediate work (see Figure 41). It is important to note that students could demonstrate high proficiency in this component even if they expressed the interquartile range as a range of values (as Student J did) instead of as the width of the interval (as Student F did).

Figure 40

Student J, Written Assessment #2, Part (b)

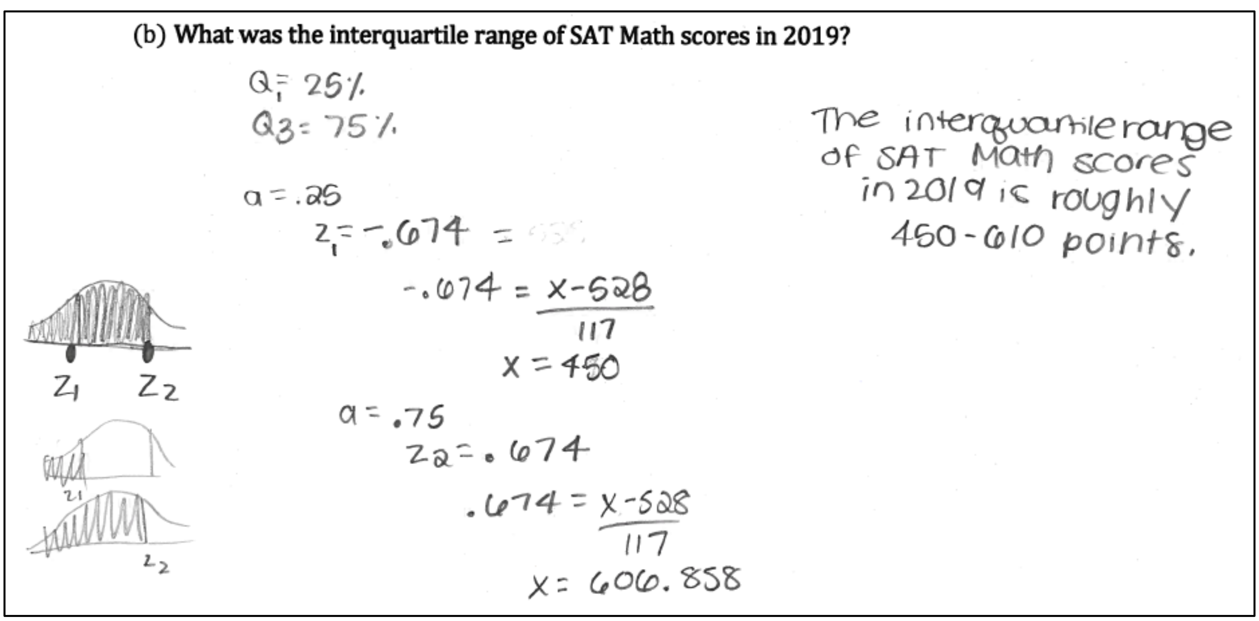
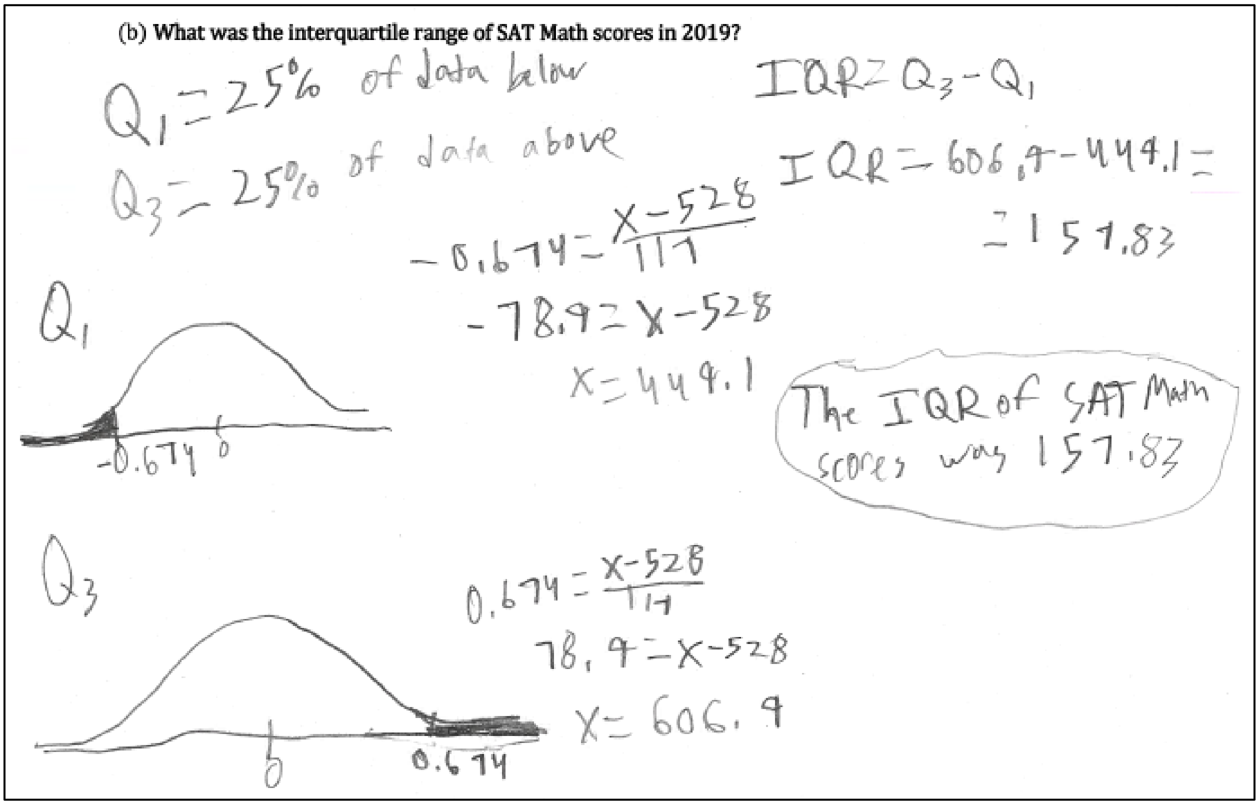


Figure 41

Student F, Written Assessment #2, Part (b)



Comparison of Oral Assessment #2 and Written Assessment #2

Again, after analyzing the oral and written assessments separately, students' oral assessment classifications were compared with their written assessment results. Just as before, each overarching component, obtaining a z-score and converting the z-score, was analyzed separately.

Component #1: Obtains Z-Score

To start, the distribution of students' proficiency in obtaining a correct z-score on each part of the written assessment was compared among the three levels of students' demonstrated knowledge of this component in the oral assessment. Figure 42 and Figure 43 show the distributions for part (a) and part (b), respectively.

Figure 42

Distribution of Component #1 Proficiency Ratings, Written Assessment #2, Part (a)

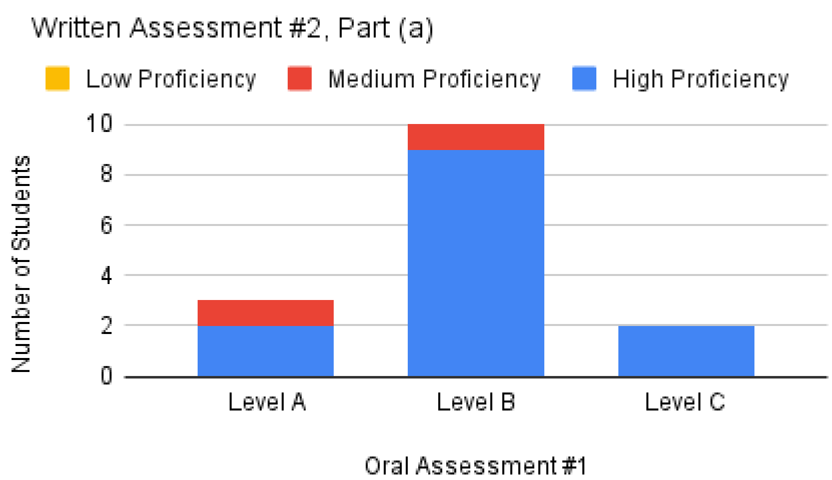
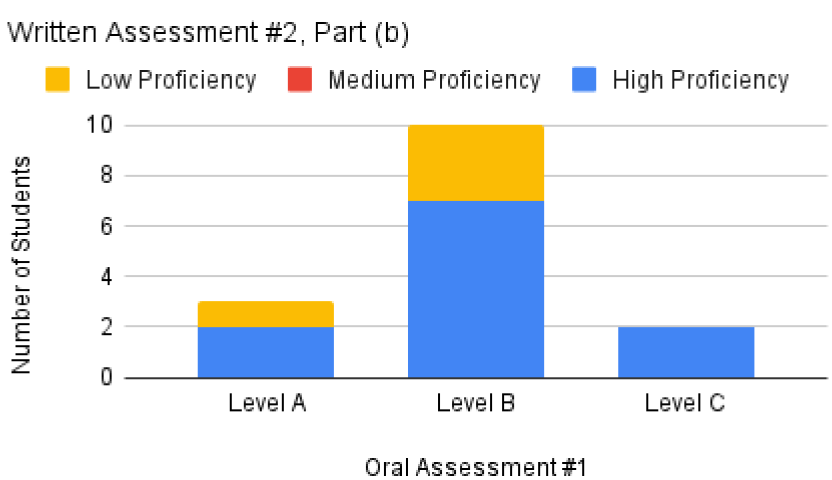


Figure 43

Distribution of Component #1 Proficiency Ratings, Written Assessment #2, Part (b)



It is important to note that the overwhelming majority of students demonstrated high proficiency in obtaining a correct z-score on the written assessment, regardless of level of knowledge demonstrated on the oral assessment. It is also important to note that knowledge levels on this oral assessment were not nearly as uniformly distributed as they were on the first oral assessment. With both of those considerations in mind, it is interesting to note that the proportion of students demonstrating high proficiency on written assessments tends to increase as demonstrated knowledge level increases.

Component #2: Converts Z-Score

Next, the distribution of students' proficiency in converting a standardized z-score to an SAT Math score was compared among the three levels of students' demonstrated knowledge of this component in the oral assessment. Figure 44 and Figure 45 show the distributions for part (a) and part (b), respectively.

Figure 44

Distribution of Component #2 Proficiency Ratings, Written Assessment #2, Part (a)

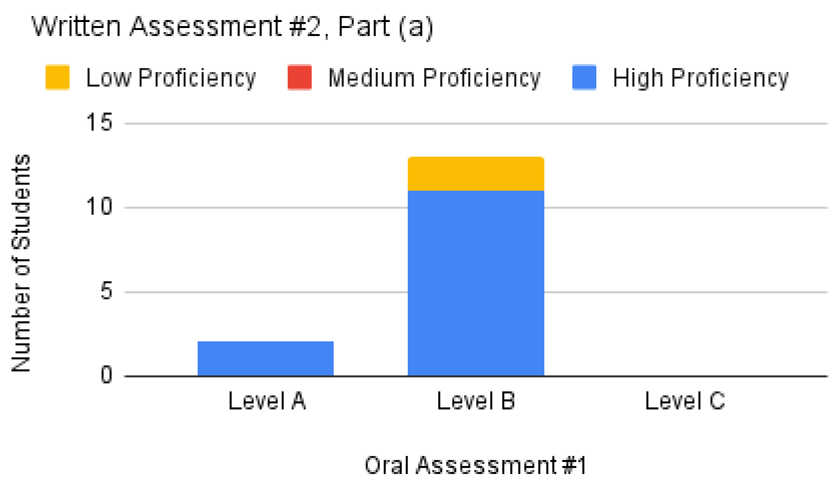
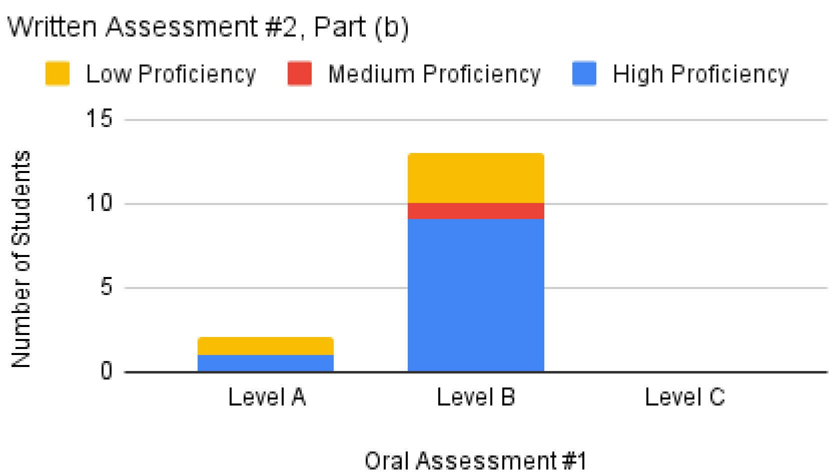


Figure 45

Distribution of Component #2 Proficiency Ratings, Written Assessment #2, Part (b)



As no student demonstrated Level C knowledge on the oral assessment for this component, and as the vast majority of students demonstrated Level B knowledge, it is difficult to make any generalizations about the relationship between knowledge level and proficiency for the component of converting z-scores. It is also important to acknowledge that it was difficult to demonstrate proficiency on the second component without first demonstrating proficiency on the first component; in other words, if a student did not obtain a z-score, it was nearly impossible for them to demonstrate proficiency in converting a z-score.

CHAPTER 5: DISCUSSION

The purpose of this study is to explore the use of oral formative assessment in secondary mathematics classrooms as a link between conceptual understanding and procedural fluency. The rationale for this study is composed of several elements. First, as classroom teachers were navigating the unprecedented challenges of facilitating student learning during the COVID-19 pandemic, the researcher became interested in how to leverage technology to promote authentic learning experiences. In particular, this study was grounded in the mathematical teaching practice of building procedural fluency through conceptual understanding (NCTM, 2014). There is a wide range of literature discussing the connection between procedural knowledge and conceptual knowledge in mathematics education (Baroody et al., 2007; Hiebert & Lefevre, 1986; Nilsson, 2020; Rittle-Johnson et al., 2001; Rittle-Johnson et al., 2015; Star, 2005; Star & Stylianides, 2013). However, despite all this literature, many assessments in mathematics classrooms lack the ability to provide meaningful insight into student thinking (Niemi, 1996). As stated previously, NCTM (2014) suggests that "obtaining evidence about understanding and reasoning requires the use of tasks and methods designed for those purposes" (p. 92). This study seeks to explore oral formative assessments as a tool designed specifically for insight into students' conceptual understanding of mathematics.

This chapter is divided into four sections: limitations, addressing the research questions, implications, and recommendations. To begin, limitations of this study will be acknowledged and discussed. Next, conclusions will be presented with respect to the research questions. The following section will identify and discuss implications for teachers. Finally, the chapter concludes with recommendations for future research.

Limitations

The major limitation to this study is that it is based on a small group of volunteer participants who are all enrolled in the same AP Statistics course. As such, any generalizations based on the findings are limited. Furthermore, the researcher recognizes that this was not a controlled study, and no quantitative analysis of the data was attempted. However, despite this large limitation, the researcher believes that a qualitative analysis of participants' oral and written work provides worthwhile preliminary insight into the use of oral formative assessments in secondary mathematics classrooms.

Another limitation to this study is that while students' conceptual knowledge as demonstrated in video recordings is analyzed and evaluated, student responses may not be entirely reflective of everything they know about the normal distribution. Student responses are likely influenced by the quality of the question presented to them, the classroom culture and norms, and other underlying factors. For example, in some video responses, the researcher noted students saying verbatim expressions that had been used by the teacher in class; it is difficult to assess if students truly understood what they were saying or if they were simply repeating conceptual phrases that they remembered. On the other hand, some students who recorded very brief, surface-level responses may have had a deeper level of conceptual understanding than what they chose to communicate in their videos.

A third limitation is that both the oral and written assessments were developed by the researcher for the purpose of this study, and flaws were uncovered in the process of analyzing the data. In particular, in the second written assessment, it proved difficult for students to demonstrate high proficiency on the second component (converting a z-score) if they had not demonstrated proficiency on the first component (obtaining a z-score). Because these two components were intrinsically related, it was difficult to analyze the two components separately.

Finally, this research is based on only one topic (the normal distribution) in one course (AP Statistics). There is likely variation in the utility and implementation of oral formative assessments among different topics and within different courses. This study looks specifically at the context of the normal distribution and does not seek to serve as a template for analysis of conceptual and procedural knowledge of other topics.

Addressing the Research Questions

This section will discuss the two research questions explored in this study, which were first presented in Chapter 2:

1. What is the nature of students' conceptual understanding of the normal distribution as revealed by oral formative assessments?
2. How does a student's conceptual understanding translate to procedural fluency on written assessments?

Research Question 1

The two oral formative assessments that each student recorded during this study were designed with the intent of eliciting evidence of students' conceptual understanding of the normal distribution. As described previously, on each oral assessment, students were presented with a worked-out solution to a problem and instructed to describe how the solution was reached as clearly as possible. Using Star's (2005) and Baroody et al.'s (2007) frameworks for types and qualities of knowledge, students' recordings were analyzed for evidence of superficial procedural knowledge, deep procedural knowledge, superficial conceptual knowledge, or deep conceptual knowledge.

There was a wide range of knowledge demonstrated by students on these oral formative assessments. On the first oral assessment, it was difficult to detect if some students had any sort of conceptual understanding of a z-score, as their explanations were purely procedural and lacked any context. Other students were able to explain that a z-score represented a number of standard deviations away from the mean, but there were varying levels of clarity in their explanations; some students presented their explanations quickly with no elaboration, begging the question of whether they were simply repeating a phrase heard in class, while other students took more time to explain each step they took and how that led them to the number of standard deviations the data point was away from the mean. The few students who interpreted the calculated z-score in the context of SAT scores revealed a deep conceptual understanding of the process of transforming raw data into z-scores, and what the z-score reveals about a single data point's position with respect to the mean of the distribution.

Again, on the first oral assessment, it was unclear if some students had any sort of conceptual understanding of normal density curves, as their explanations for how to find the shaded area under a normal curve were purely procedural, with step-by-step instructions of how to use a calculator or a standardized normal table to arrive at the desired answer. Some students, on the other hand, demonstrated procedural flexibility in their explanations. A number of students described how subtracting the area on the left side of the normal density curve from one would result in the complementary area on the right side of the normal density curve. This flexible procedural knowledge hinted at a conceptual understanding of density curves. Finally, the students who were able to clearly articulate that the area under a normal density curve represents the proportion of observations in that region demonstrated their deep conceptual knowledge.

On the second oral assessment, students were first evaluated on their demonstrated knowledge about obtaining z-scores corresponding to a certain percentile on a normal distribution curve. It was unclear whether some students had even basic procedural knowledge about how to obtain the z-score, as their explanations lacked any sort of replicability. The majority of students, however, demonstrated deep procedural knowledge, clearly explaining how to use either a calculator or a standardized normal table to find the desired z-score, connecting it to the appropriate area either above or below that z-score. The two students who demonstrated deep conceptual knowledge were able to clearly articulate that the area of interest on the standardized curve represented the proportion of data points in that region. Then, they were able to explain that the z-score they obtained represented the location on the standardized normal curve with a certain proportion of the data above or below it.

Once students had been evaluated on their description of obtaining the z-score, they were then evaluated on the knowledge demonstrated when describing the conversion of the standardized z-score to a contextualized SAT score. In a course like AP Statistics, where algebraic manipulation is not central to the material, it was interesting to note the apparent ease with which students jumped to a procedural description of "solving for X." Two students demonstrated superficial procedural knowledge by describing the algebraic manipulation purely numerically, with no consideration as to what the numbers they were using represented. The remaining students in the study demonstrated superficial conceptual knowledge by simply incorporating the terms "mean" and "standard deviation" into their procedural description. It was interesting to note that no students demonstrated deep conceptual understanding of what it means to convert a z-score to an SAT score, namely that multiplying the z-score by the standard deviation of the SAT score distribution converts the z-score from a standardized score to a number of points that an SAT score is away from the mean SAT score.

As acknowledged previously, it is important to note that these analyses of students' procedural and conceptual knowledge are based purely on the verbal explanations elicited from students, which are not necessarily indicative of all a student may know about the normal distribution. That being said, these oral formative assessments provided insight into students' thinking and understanding that is not easily gained by paper-and-pencil assessments.

Research Question 2

After each oral assessment, students participated in a written assessment, and those assessments were evaluated for procedural proficiency. The next question of interest was whether there was any relationship between a student's conceptual understanding as demonstrated on the oral assessment and their demonstrated procedural proficiency on the written assessment. Again, it is important to note that this was a volunteer study, so generalizations of any observations are very limited. Still, a few interesting trends were found in the comparison of oral and written assessment data.

The first interesting trend is that all students who demonstrated deep conceptual understanding of a topic in their oral formative assessments also demonstrated high proficiency in that topic on all parts of their written assessments. In other words, in this study, the ability to communicate deep conceptual understanding was a clear indicator of future high procedural fluency.

The alternate side of that trend is more nuanced; just because a student demonstrated superficial procedural knowledge on the oral assessment did not necessarily indicate that they would demonstrate low proficiency on the written assessment. For example, three students demonstrated superficial procedural knowledge on both components of the first oral assessment. Two of those students demonstrated a mix of low, medium, and high proficiency on the different parts of their written assessment, but the third student demonstrated high proficiency on each part of the written assessment.

In short, while demonstration of deep conceptual knowledge was a clear predictor of high procedural proficiency in this study, demonstration of superficial procedural knowledge was not a clear predictor of procedural proficiency. Again, this lends support to the important acknowledgement that students may have deeper knowledge than what they demonstrate in an oral formative assessment.

Implications

Based on this study, suggestions can be made for teachers about both the practicality and utility of implementing oral formative assessments in secondary mathematics classrooms. Literature has often suggested that having students verbalize their thoughts about mathematics is beneficial for both teachers and students (Ellemor-Collins & Wright, 2008; Ghousseini et al., 2017; Kazemi, 1998; Long & Ben-Hur, 1991; Wagganer, 2015; Walter, 2018). That being said, implementing oral formative assessment with regularity has proven difficult due to the time-consuming nature of interview-based assessments (Ellemor-Collins & Wright, 2008).

With the advent of recording technologies and student access to personal recording devices, asynchronous recorded responses provide teachers with the unprecedented opportunity to "get inside their students' heads in an ongoing way" (Schoenfeld, 2007, p. 277). In the thirty videos recorded as part of this study, the average video length was under two minutes. While this is not an insignificant amount of time when considered in the context of a classroom of thirty students, it is not a prohibitive amount of time; teachers should be able to implement short oral formative assessment with some regularity.

If secondary teachers do choose to make oral formative assessment a regular part of their assessment process, they will need to work on developing classroom culture, norms, and expectations around students' recordings. With recorded responses, teachers lose the ability to question students, pressing them to elaborate on their responses and support their arguments (Long & Ben-Hur, 1991). Through in-class analysis of sample recorded responses, students can learn over time what it means to truly "articulate and justify their own mathematical ideas" (Carpenter et al., 2003, p. 6). This will help teachers learn as much as possible about what a student's true conceptual understanding is, without taking the time to individually interview each student.

However, the responsibility of high quality recorded responses does not rest solely on the student; the teacher is also responsible for presenting the student with a prompt that elicits conceptual reasoning and justification. These prompts must be developed thoughtfully and carefully, with attention to the conceptual understandings central to the learning objectives at hand. It takes conscious effort to create oral assessment prompts that exhort students to share their thinking processes.

Since recorded response as a form of oral formative assessment has not been widely used in secondary mathematics classrooms, professional development is needed to help teachers implement it effectively. Mathematics teacher educators should support in-service teachers in developing prompts that elicit useful student responses, specifically in writing and testing questions that implore students to reveal their conceptual understanding, not just describe a procedure. Furthermore, support should be provided for teachers as they use these responses to make instructional decisions. As these assessments are mean to be formative, teachers should feel confident in using the information they gain about students' conceptual and procedural knowledge to promote student growth and learning.

Recommendations for Future Research

As stated previously, because this study was based on a small group of volunteer participants in a specific mathematical context, the findings cannot be widely generalized. More research is needed to determine the best way to implement oral formative assessment in secondary mathematics classrooms, and also how both teachers and students can use the data from oral formative assessments to improve student learning. For example, some future research questions might include:

- What types of questions elicit student responses that demonstrate their conceptual understanding?
- How can data from oral formative assessment inform teachers' instructional decision making?
- How do oral formative assessments foster student self-assessment?

REFERENCES

Albert, L. R. (2000). Outside-in – inside-out: Seventh-grade students' mathematical thought processes. *Educational Studies in Mathematics, 41(2)*, 109-141. https://www.jstor.org/stable/3483186

Back, R., Marinila, L., & Wallin, S. (2009). Student justifications in high school mathematics. *Proceedings of the Sixth Conference of European Research in Mathematics Education*, 291-300. Retrieved August 5, 2021 from http://ife.ens-lyon.fr/publications/edition-electronique/cerme6/wg2-12-back.pdf

Baran, E. (2014). A review of research on mobile learning in teacher education. *Journal of Educational Technology & Society* , *17*(4), 17-32. https://www.jstor.org/stable/10.2307/jeductechsoci.17.4.17

Baron, L. M. (2016). Formative assessment at work in the classroom. *The Mathematics Teacher*, *110*(1), 46–52. https://doi.org/10.5951/mathteacher.110.1.0046

Baroody, A. J., Feil, Y., & Johnson, A. (2007). An alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education 38(2)*, 115-131. https://doi.org/10.2307/30034952

Batanero, C., Tauber, L.M., & Sánchez, V. (2004). Students' reasoning about the normal distribution. In D. Ben-Zvi & J. Garfield (Eds), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 257-276). Springer, Dordrecht. https://doi.org/10.1007/1-4020-2278-6_11

Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F.E. Weinert & R.H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Erlbaum.

Carpenter, T. P., Franke, M. L., & Levi, L. (2003). *Thinking mathematically: Integrating arithmetic and algebra in elementary school*. Heinemann.

Carraher, T. N., Carraher, D. W., & Schliemann, A. D. (1987). Written and oral mathematics. *Journal for Research in Mathematics Education*, *18*(2), 83–97. https://doi.org/10.2307/749244

Cole, K. A. (1999). Walking around: Getting more from informal assessment. *Mathematics Teaching in the Middle School*, *4*(4), 224–227. https://doi.org/10.5951/mtms.4.4.0224

College Board. (2020). *AP statistics course and exam description. https://apcentral.collegeboard.org/pdf/ap-statistics-course-and-exam-description.pdf*

Ellemor-Collins, D. L., & Wright, R. J. (2008). Assessing student thinking about arithmetic: Videotaped interviews. *Teaching Children Mathematics*, *15*(2), 106–111. https://doi.org/10.5951/tcm.15.2.0106

Ghousseini, H., Lord, S., & Cardon, A. (2017). Supporting math talk in small groups. *Teaching Children Mathematics*, *23*(7), 422–428. https://doi.org/10.5951/teacchilmath.23.7.0422

Ginsburg, H. P. (1981). The clinical interview in psychological research on mathematical thinking: Aims, rationales, techniques. *For the Learning of Mathematics*, *1*(3), 4-11. https://www.jstor.org/stable/40247721

Ginsburg, H. P. (2009). The challenge of formative assessment in mathematics education: Children's minds, teachers' minds. *Human Development*, *52*(2), 109–128. https://doi.org/10.1159/000202729

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522. https://doi.org/10.1111/1467-9280.00395

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analysis. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 1-27). Erlbaum.

Joyner, J. M., & Bright, G. W. (2016). *Informative assessment: Formative assessment to improve mathematics achievement, middle and high school*. Math Solutions.

Kazemi, E. (1998). Research into practice: Discourse that promotes conceptual understanding. *Teaching Children Mathematics*, *4*(7), 410–414. https://doi.org/10.5951/tcm.4.7.0410

Long, M. J., & Ben-Hur, M. (1991). Informing learning through the clinical interview. *The Arithmetic Teacher*, *38*(6), 44–46. https://doi.org/10.5951/at.38.6.0044

McIntosh, M. E. (1997). Formative Assessment in Mathematics. *The Clearing House*, *71*(2), 92–96. https://www.jstor.org/stable/30192092

Murray, F. B., & Sharp, H. R. (1986). Foreword. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. xi-xiii). Erlbaum.

National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: NCTM.

National Council of Teachers of Mathematics. (2014). *Principles to actions: Ensuring mathematical success for all*. Reston, VA: NCTM.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, D.C.: U.S. Department of Education.

Niemi, D. (1996). Assessing conceptual understanding in mathematics: Representations, problem solutions, justifications, and explanations. *The Journal of Educational Research*, *89*(6), 351–363. https://doi.org/10.1080/00220671.1996.9941339

Nilsson, P. (2020). A framework for investigating qualities of procedural and conceptual knowledge in mathematics - an inferentialist perspective. *Journal for Research in Mathematics Education*, *51*(5), 574–599. https://doi.org/10.5951/jresematheduc-2020-0167

Reinholz, D. L., & Gillingham, D. (2017). Forms of formative assessment: Eliciting and using student thinking. *For the Learning of Mathematics*, *37*(1), 9–11. https://www.jstor.org/stable/44382736

Resnick, L.B., & Ford, W.W. (1981). *The psychology of mathematics for instruction*. Erlbaum.

Reyes, L. H. (1984). Affective variables and mathematics education. *The Elementary School Journal*, *84*(5), 558–581. https://doi.org/10.1086/461384

Rittle-Johnson, B., Schneider, M., & Star, J.R. (2015). Not a one-way street: Bidirectional relations between procedural and conceptual knowledge of mathematics. *Educational Psychology Review*, *27(4)*, 587-597. https://doi.org/10.1007/s10648-015-9302-x

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346

Santos, L., & Semana, S. (2014). Developing mathematics written communication through expository writing supported by assessment strategies. *Educational Studies in Mathematics*, *88*(1), 65–87. https://doi.org/10.1007/s10649-014-9557

Schoenfeld, A.H. (2007). Reflections on an assessment interview: What a close look at student understanding can reveal. In A.L. Schoenfeld (Ed.), *Assessing mathematical proficiency* (pp. 269-277). Cambridge University Press.

Senk, S. L., Beckmann, C. E., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for Research in Mathematics Education*, *28*(2), 187–215. https://doi.org/10.2307/749761

Shermis, M. D., & Di Vesta, F. J. (2011). *Classroom assessment in action*. Rowman & Littlefield Publishers.

Silver, E. A. (1986). Using conceptual and procedural knowledge: A focus on relationships. In J. Hiebert (Ed.), *Conceptual and procedural knowledge: The case of mathematics* (pp. 181-198). Erlbaum.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*(5), 404-411. https://doi.org/10.2307/30034943

Star, J. R., & Stylianides, G. J. (2013). Procedural and conceptual knowledge: Exploring the gap between knowledge type and knowledge quality. *Canadian Journal of Science, Mathematics and Technology Education*, *13*(2), 169–181. https://doi.org/10.1080/14926156.2013.784828

Wagganer, E. L. (2015). Creating math talk communities. *Teaching Children Mathematics*, *22*(4), 248–254. https://doi.org/10.5951/teacchilmath.22.4.0248

Walter, H. A. (2018). Beyond turn and talk: Creating discourse. *Teaching Children Mathematics*, *25*(3), 180–185. https://doi.org/10.5951/teacchilmath.25.3.0180

APPENDICES

Appendix A

Oral Assessments

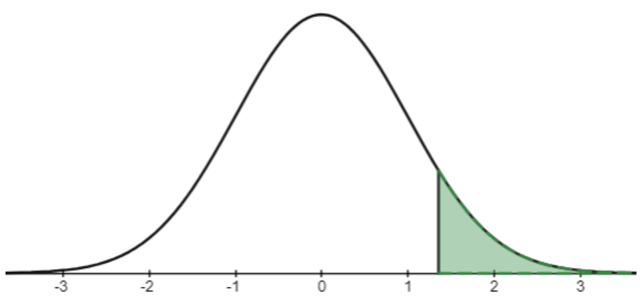
Oral Assessment #1

According to the 2019 SAT Suite of Assessments Annual Report, SAT scores in 2019 followed an approximately normal distribution with a mean score of 1059 and a standard deviation of 210.

According to NC State University Admissions, the average SAT score of admitted freshmen in the fall of 2020 was 1344. **Approximately what percent of all SAT takers scored higher than 1344?**

SOLUTION:

$$z = \frac{1344 - 1059}{210} \approx 1.36$$



8.7%

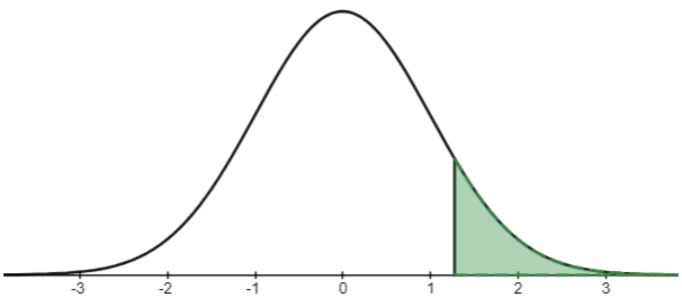
Create a Flipgrid video describing how this solution was reached. Explain each step (how it was done and what it means) as clearly as possible.

Oral Assessment #2

According to the 2019 SAT Suite of Assessments Annual Report, SAT scores in 2019 followed an approximately normal distribution with a mean score of 1059 and a standard deviation of 210.

To be in the top 10% of all test takers, approximately what score do you need to earn?

SOLUTION:



$$1.28 = \frac{x - 1059}{210}$$

$$268.8 = x - 1059$$

$$x = 1327.8$$

To be in the top 10% of all test takers, you must earn a score of approximately 1330 or higher.

Create a Flipgrid video describing how this solution was reached. Explain each step (how it was done and what it means) as clearly as possible.

Appendix B

Written Assessments

Written Assessment #1

According to the 2019 SAT Suite of Assessments Annual Report, SAT Math scores in 2019 followed an approximately normal distribution with a mean score of 528 and a standard deviation of 117.

(a) On your first attempt in 2019, you scored a 630 on the SAT Math section. **Approximately what percent of all SAT takers scored higher than you did?**

(b) On your second attempt in 2019, you scored a 680 on the SAT Math section. **At approximately what percentile did you score?**

(c) **Approximately what percent of all SAT takers in 2019 scored in the 600-700 range on the SAT Math section?**

Written Assessment #2

According to the 2019 SAT Suite of Assessments Annual Report, SAT Math scores in 2019 followed an approximately normal distribution with a mean score of 528 and a standard deviation of 117.

(a) **What score would you have needed to earn on the SAT Math section in 2019 to be in the top 5% of all test takers?**

(b) **What was the interquartile range of SAT Math scores in 2019?**