COGNITIVE DEMAND OF TEACHER-CREATED MATHEMATICS ASSESSMENTS

by

Megan Marie Schmidt

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Mathematics

Boise State University

August 2022

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Megan Marie Schmidt

Thesis Title: Cognitive Demand of Teacher-Created Assessments

Date of Final Oral Examination: 13 June 2022

The following individuals read and discussed the thesis submitted by student Megan Marie Schmidt, and they evaluated her presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Michele Carney, Ph.D.	Chair, Supervisory Committee
Laurie Cavey, Ph.D.	Member, Supervisory Committee
Joe Champion, Ph.D.	Member, Supervisory Committee
Angela Crawford, Ed.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Michele Carney, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

DEDICATION

This thesis is dedicated to Adam Christensen, who was and continues to be my biggest supporter of my goals inside and outside the classroom. His encouraging words motivated me during this process; and his positive attitude is contagious, making my hardest days bright. This is also dedicated to my friends and family who have encouraged me to continue my education. They have been by my side throughout challenges, and I'm grateful to have their support as I close a chapter in my life.

ACKNOWLEDGMENTS

I would like to acknowledge and thank my committee chair, Michele Carney, for her support, advice, and time throughout the completion of this thesis. I would also like to thank her for her mentorship throughout the years I have worked with her. I would also like to thank each member of my committee; Joe Champion, Laurie Cavey, and Angela Crawford. Each member of the committee supported me by providing meaningful feedback on my work to help me become a more intentional writer. I would like to thank each one of them for not only their support in my writing, but also for the influence each of them have had on my professional career. I would like to especially thank Joe Champion for his assistance in the analysis and display of the data collected in this thesis, and for always continuing to teach me something new about statistical analysis. Finally, I would like to thank the teachers and students participating in the study. Without their work, I would not have been able to complete the work this thesis describes.

ABSTRACT

This study analyzed assessments created by middle school mathematics teachers participating in a large scale research project in the Northwestern United States. Assessments were coded using the frameworks of Webb's Depth of Knowledge (DOK) for mathematics content (2002), and Smith and Stein's Levels of Demand (LOD) (1998). Teachers in the study were instructed to create an assessment using 5 common-items provided by the research team and 5 of their own sourced items. Assessment items were coded using each framework and data was collected based on the DOK framework, LOD framework, and grade level. Findings indicate on average, teachers assessments were relatively balanced between procedures and conceptual understanding and balanced between *recollection*, application of a *skill/concept*, and *explanation of thinking*. When looking at data based on grade level, assessments tended to address higher-level thinking as grade levels progressed. However when common-items were not included in the analysis, assessments tended to address the items using procedural thinking, and less frequently required explanation of thinking for solutions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAs	Classroom Assessments
DOK	Depth of Knowledge
LOD	Levels of Demand
ROOT	Researching the Order of Teaching
EAC	Explicit Attention to Concepts
SOS	Student Opportunities to Struggle

CHAPTER ONE: INTRODUCTION

One expectation of teachers is to assess their students regularly in order to meet state and district standards, monitor progress of student learning objectives, and collect and use data to inform instruction. Teachers use a variety of testing formats such as standardized testing, but the most frequently used and teacher-influenced testing format is that of Classroom Assessments (CAs). Classroom Assessments are assessments administered to students in a classroom setting, and are presented as questions to monitor student understanding and performance of the specific content they are studying (McMillan, 2013). CAs range in a continuum of formality and are used for formative and summative purposes. CAs can be administered in a variety of formats too, such as check-ins, exit tickets, verbal questioning, and written exams. While state standards influence both CAs and high-stakes standardized assessments, CAs are not classified as high-stakes assessments (Schneider et al., 2013).

Because formative and summative CAs can measure the breadth and depth of what students know in a particular content area understanding, teachers' choices in writing CAs is important. In this thesis, I analyzed grades 6-8 mathematics teachers' professional decisions in CA design in order to gain insight into the cognitive demand teachers deem appropriate for a summative CA context.

There is no one way a teacher writes summative CAs. Not only is there variety in teachers' processes of writing or selecting items, there is also variety in teachers' attending to the sequencing and progression of the summative CA (Moss, 2013). As a

result, each teacher creates a completely unique CA every time it is administered. The uniqueness serves as a benefit in cases when the teacher is knowledgeable of common CA goals and practices. Research suggests that teachers who are well-versed in CA goals and practices create more effective CAs (Popham, 2009), and also incorporate more variety into CAs (Bailey & Heritage, 2008; Green & Stager, 1986, 1987). However, most teachers possess a limited knowledge of CAs, making it challenging to adequately reach the goals and adhere to the practices of CAs (Stiggins, 2002). The goals and practices important for this study are to elicit student thinking in a way that aligns with state standards for mathematics. The participants in this study did this through their work in the classroom, and the goal is for there to be alignment between information deemed important in the classroom to be reflected on the CA.

During my time as an undergraduate student, I had opportunities to observe and partake in the creation of CAs for mathematics students both independently and with a collaborator. As a graduate research assistant for a large-scale project, I collected and recorded data on end-of-unit CAs administered by teachers involved in the project. In my observations as an undergraduate and graduate student, new noticings arose from my experience with CAs, such as length, difficulty, content, question format, sequencing, etc. My professional experience has led me to want to know more about the processes and understandings of teachers in their creation of summative CAs.

Proposal

The goal of this thesis was to examine aspects of teacher-created CAs that relate to the cognitive demand of assessments. I used two well established frameworks to operationalize cognitive demand: Webb's Depth of Knowledge (DOK) for mathematics

content (2002), and Smith and Stein's Levels of Demand (LOD) (1998). While similar, DOK and LOD combine to provide a fuller picture of cognitive demand of assessment than each might separately. In turn, the descriptions of cognitive demand can help gain insight into teachers' choices while assessment planning.

Context

Participants in the study were a part of the large-scale research project Researching the Order of Teaching (ROOT). ROOT participants received professional development on teaching practices based on Explicit Attention to Concepts (EAC) and Student Opportunities to Struggle (SOS). The selection of these practices was based on research identifying EAC and SOS practices associated with increased conceptual understanding in students (Heibert & Grouws, 2007; Stein et al., 2017). Within the ROOT project, the construct of EAC was conceptualized as focusing on concepts, making concepts explicit and public, and emphasizing connections. Additionally, the construct of SOS was conceptualized as focusing on sense-making, applying sustained mental effort, and engaging with important math (Champion et al., 2021). In order to support teachers in engaging their students with EAC and SOS, teachers participated in modules geared toward targeting their instruction and fine-tuning their understanding of what EAC and SOS looks like in a classroom. The teachers implemented and reflected upon their work with EAC and SOS throughout the three year project. As part of the teachers' work in the project, they designed assessments to assist in evaluating the impact of EAC and SOS practices on student achievement.

CHAPTER TWO: LITERATURE REVIEW

Classroom Assessment

Classroom Assessment is an integral part of pedagogy, and the data collected from CAs holds major implications about teachers and students. The CAs require teachers to account for state standards, local curriculum, classroom tasks, and student's current knowledge of the content; to name a few. Randel & Clark describe CA as more than a product or tool, but a teaching practice of which the purpose is informing students and educators of student learning (2013). Because of the stakes associated with informing student learning, effective assessment practice must be a part of educators' pedagogy, however a variety of barriers cause teachers to possess limited CA knowledge and literacy. Stiggins found two primary barriers were (1) lack of professional development or courses offered to in-service and preservice teachers on educational testing and measurement, and (2) teacher's perspectives on the usefulness of such training (1995). The lack of professional development around CA and teacher perceptions of usefulness of CA training affect the writing of teacher-created assessments.

Teachers' experiences with CAs have been presented in educational research through numerous studies and measures of CA knowledge. When teachers were given CA sources and examples, teachers deemed their own customized CA as better measures for their students (Boothroyd, McMorris, & Pruzek, 1992; Stiggins & Bridgeford, 1985). This finding is not surprising due to the numerous aspects teachers account for when creating their CAs. Teachers' expertise of their students and districts benefits the CAs

they produce, in that they fine-tuned CAs to the students, school, and district involved. However, when teachers lack knowledge and ability of effective practices of CAs, a fine-tuned CA might be lacking in the necessary components required to assess students at appropriate levels of rigor.

One way pedagogical knowledge of effective practices has been addressed in research is through the Assessment Practices Inventory (API) self report created by Zhang & Burry-Stock (1994). The API is a 67-question instrument that considers a teacher's frequency of assessment practices as well as a teacher's assessment skills. In Zhang & Burry-Stock's study they found there was a correlation between the grade level taught and the approach that teachers report to have toward assessment items. While elementary school teachers incorporate performance tasks in their assessments, significant evidence shows middle and high school teachers rely more on objective assessment items. This is cause for concern because as learning objectives become more challenging, there appears to be a decrease in the amount of genuine problem solving a student is exposed to during assessments.

Zhang & Burry-Stock's study also supports the need for measurement training for pre-service or in-service teachers (1994). The findings indicate as little as one course of measurement training for teachers led to a 10% increase in teachers' API scores. Interestingly, years of experience has not been shown to have a significant factor in teachers' API scores. The use of measurement training could benefit teachers in their CA practices and skills, especially for secondary school teachers because of the increased difficulty to assess learning objectives through problem solving.

Depth of Knowledge (DOK)

As educational reform has become increasingly present in schools, frameworks to assist in alignment between state standards and the methods to which standards are assessed have been developed. At the forefront was Webb's Depth of Knowledge framework. Webb's DOK framework specializes based on content; i.e. Reading, Writing, Mathematics, Science, Social Studies. However, across all content specific frameworks there is a common idea of a progression of cognitive thinking in how assessment items are solved. In this study, Webb's DOK framework was chosen based on its specialization of mathematics tasks, and its close relation to Bloom's Taxonomy by eliciting increasing levels of cognitive rigor (Hess et al., 2009). Webb's framework outlines four distinct DOK levels specific to the content areas of Language Arts, Mathematics, Science, and Social Studies. The Mathematics DOK levels are described as 1) *Recall*, 2) *Skill/Concept*, 3) *Strategic Thinking*, and 4) *Extended Thinking*. The basis of this increasing difficulty is established from how an assessment item is structured in order to elicit cognitive reasonings from students. While the differences in cognitive reasoning can often be displayed through the language of an assessment item, an assessment item can imply what a student must do in order to display the reasonings. Figure 2.1 displays each of the four levels of Mathematics DOK and examples of the associated action words a teacher could possibly use to assess for each DOK level. Assessments sometimes 1) use these exact words, and 2) are aligned to the meaning of each word; however, there are cases in assessments when one of these two factors is not present. The more important of the factors is the second one, and connects an assessment item to a particular DOK.

Source: N. Webb (2002)

Figure 2.1 Mathematics DOK Levels and Associated Action Words (Darling-Hammond et al., 2013)

Webb's purpose for DOK levels is to argue they should each be represented in both the curriculum standards and assessments to ensure there is alignment between the two. Based on this alignment with curriculum standards, some have recommended assessments be made up of approximately two-thirds of conceptual knowledge and abilities (DOK 2, 3, and 4) (Darling-Hammond et al., 2013). To achieve this, cohesion between standards and assessment must be achieved. In an analysis of the alignment between both the state standards and state assessments, the complexity of the standards caused the DOK to vary greatly in assessment items (Webb, 2007). Many alignment issues are still yet to be resolved, however one question in particular is, how do specific grade level standards lend themselves to the DOK framework?

Research involving teachers' interpretation of the DOK framework suggests alignment between standards and assessment are unrealistic in practice. Teachers' support that the appropriate use of each of the DOK levels would involve DOK 1, 2, and 3 in CAs; DOK 4 in large-scale class projects; and DOK 3 and 4 in class activities requiring an in-depth understanding (Hess et al., 2009). The exclusion of DOK 4 in CAs could be explained in part from the action words required of DOK 4, such as design, connect, synthesize, etc. which leads to complex CAs. Without major change in CAs, or creative CAs, it is reasonable to state DOK 4 assessment items should not be included with CAs, aligning with teachers' views toward DOK and assessment items. Within the first three levels of DOK, teachers claimed there should be an appropriate blend of each DOK level, one of which aligns with content standards. However, what was found in students' mathematics learning through homework, classwork, quizzes, and exams was that the distribution of DOK was weighted toward lower DOK levels. In a 3rd grade classroom, student work was categorized primarily as DOK 1 (82%) followed by DOK 2 (17%) (Hess et al., 2009). These DOK levels fall below the ⅔ level of conceptual understanding expected based on the mathematics standards.

The lack of conceptual understanding evidenced by student work is partially a product of high-stake state assessments. Teachers' instruction and CAs sometimes are focused on expectations of students put in place by high-stakes state assessments. Studies taking place before states made transitions toward higher levels of conceptual understanding show a disparity in levels of DOK required from students on state achievement assessments. In a study conducted from 17 states, selected based on their higher standards and more ambitious state assessments, fewer than 2% of mathematics

items were categorized as achieving higher levels of DOK (DOK 3 and 4) (Yuan & Le, 2012). This study also found DOK was greatly affected by question format, which primarily was multiple choice. The format of multiple choice is conducive to large scale assessments, but a question arises as to what is the highest level of DOK that can be achieved while using a multiple choice format. Yuan & Le coded items by having two content-specific coders analyze an item based on the DOK framework. The codes were based on the increasing difficulty in the cognitive levels of the framework, which was often described based on how in-depth the process would be for each assessment item.

While Yuan & Le's study analyzed the specific assessment items individually (2012), another study analyzed the state assessments as a complete unit (Polikoff et al., 2011). Polikoff & colleagues analyzed alignment of state standards with state assessments, and defined their cognitive demand framework where one basis for their framework was Webb's DOK framework. Assessments were analyzed independently by multiple coders and were recorded in a particular cognitive demand cell. In the study conducted by Polikoff & colleagues, it was found that from 19 state assessments 80% of mathematics assessments only used lower level cognitive demand (DOK 1 and 2) in the assessment. The study also analyzed individual items and found the state assessments for mathematics only contained 7% of assessment items requiring high levels of cognitive demand (DOK 3 and 4).

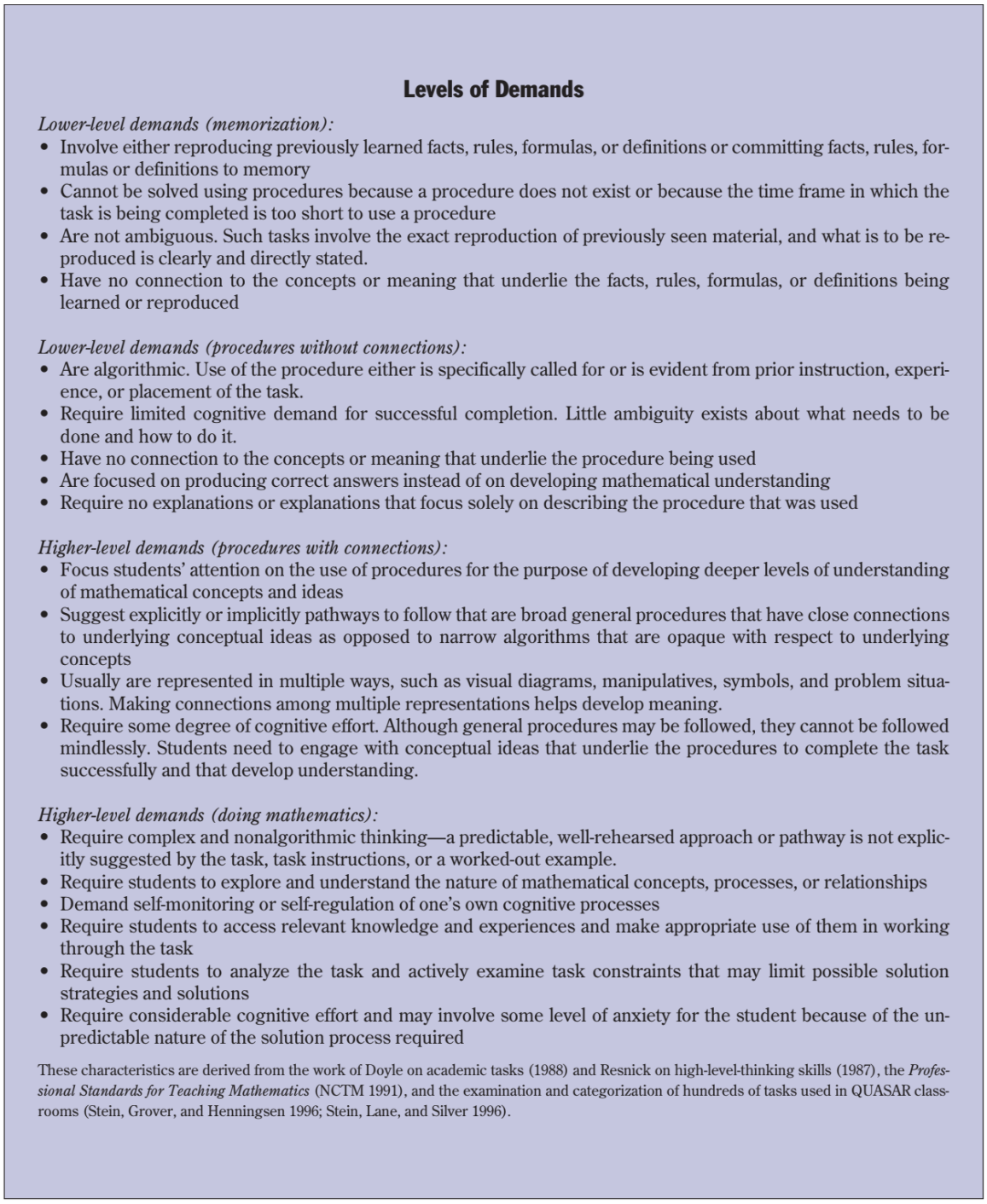
State assessments have undergone some reform since both Yuan & Le and Polikoff & colleagues' research were conducted. Several large-scale educational assessments have been thoroughly researched and vetted by multiple organizations. One leading organization is the Smarter Balanced Assessment Consortium (SBAC), where

one purpose of SBAC is providing assessments aligned with standards. SBAC developed a blueprint describing how students should be assessed according to state standards, and through their assessment resources, an appropriate distribution of DOK can be achieved. Within the blueprint SBAC presents information on the four claims of 1) Concepts and Procedures, 2) Problem Solving, 3) Communicating Reasoning, and 4) Modeling and Data Analysis; scoring format CAT (machine-scored) and PT (hand-scored); and the DOK level of assessment items are addressed and quantified in detail (SBAC, 2019). Interestingly, overlap between lower and higher levels of the claims, scoring format, and DOK are available. For the claim of Concepts and Procedures a DOK 3 question can be achieved at the 3rd, 4th, and 11th grade levels. Overlap is also present between the machine scored items and high DOK levels. Displayed at each grade level, DOK 3 and 4 are possible to be presented in a machine-scored question format. Sources such as SBAC are examples of ways alignment between standards and assessments in a high-stakes testing environment can be achieved. With these reforms toward state assessments, it is expected for CAs to also experience a shift.

Levels of Demand (LOD)

The other framework used in this analysis is Smith and Stein's Levels of Demand framework, which was established to assist teachers in the creation of mathematically rich tasks. Smith and Stein's LOD framework is a part of a larger body of work from Stein's Implementing Standards-Based Mathematics Instruction (2000). In Stein (2000) the overarching pedagogical goal of establishing meaningful connections in mathematics concepts motivates the usefulness of the LOD framework. When students engage in tasks at a higher difficulty according to the framework, students are presented with ways to

establish conceptual understandings of mathematics. The LOD is related to how tasks are presented in a classroom setting before a CA is administered, however the intention in using this framework is analyzing alignment between teachers assessment items and how teachers tend to present concepts in class activities. The reason why Smith and Stein's LOD framework was chosen was because the framework focuses on mathematics classroom tasks and showed alignment between CAs and the work done in class. The framework consists of four distinct levels, 1) *Memorization*, 2) *Procedures without Connections*, 3) *Procedures with Connections*, and 4) *Doing Mathematics* (1998). Smith and Stein described the characteristics of each level in Figure 2.2. The lower LOD are 1) *Memorization* and 2) *Procedures without Connections*, while the higher LOD are 3) *Procedures with Connections* and 4) *Doing Mathematics*.



Levels of Demands

Lower-level demands (memorization):

- Involve either reproducing previously learned facts, rules, formulas, or definitions or committing facts, rules, formulas or definitions to memory
- Cannot be solved using procedures because a procedure does not exist or because the time frame in which the task is being completed is too short to use a procedure
- Are not ambiguous. Such tasks involve the exact reproduction of previously seen material, and what is to be reproduced is clearly and directly stated.
- Have no connection to the concepts or meaning that underlie the facts, rules, formulas, or definitions being learned or reproduced

Lower-level demands (procedures without connections):

- Are algorithmic. Use of the procedure either is specifically called for or is evident from prior instruction, experience, or placement of the task.
- Require limited cognitive demand for successful completion. Little ambiguity exists about what needs to be done and how to do it.
- Have no connection to the concepts or meaning that underlie the procedure being used
- Are focused on producing correct answers instead of on developing mathematical understanding
- Require no explanations or explanations that focus solely on describing the procedure that was used

Higher-level demands (procedures with connections):

- Focus students' attention on the use of procedures for the purpose of developing deeper levels of understanding of mathematical concepts and ideas
- Suggest explicitly or implicitly pathways to follow that are broad general procedures that have close connections to underlying conceptual ideas as opposed to narrow algorithms that are opaque with respect to underlying concepts
- Usually are represented in multiple ways, such as visual diagrams, manipulatives, symbols, and problem situations. Making connections among multiple representations helps develop meaning.
- Require some degree of cognitive effort. Although general procedures may be followed, they cannot be followed mindlessly. Students need to engage with conceptual ideas that underlie the procedures to complete the task successfully and that develop understanding.

Higher-level demands (doing mathematics):

- Require complex and nonalgorithmic thinking—a predictable, well-rehearsed approach or pathway is not explicitly suggested by the task, task instructions, or a worked-out example.
- Require students to explore and understand the nature of mathematical concepts, processes, or relationships
- Demand self-monitoring or self-regulation of one's own cognitive processes
- Require students to access relevant knowledge and experiences and make appropriate use of them in working through the task
- Require students to analyze the task and actively examine task constraints that may limit possible solution strategies and solutions
- Require considerable cognitive effort and may involve some level of anxiety for the student because of the unpredictable nature of the solution process required

These characteristics are derived from the work of Doyle on academic tasks (1988) and Resnick on high-level-thinking skills (1987), the *Professional Standards for Teaching Mathematics* (NCTM 1991), and the examination and categorization of hundreds of tasks used in QUASAR classrooms (Stein, Grover, and Henningsen 1996; Stein, Lane, and Silver 1996).

Figure 2.2 Characteristics of Levels of Demands (Smith & Stein, 1998)

The commonalities between how each level is described is related to the students' required level of cognitive effort and the ambiguity of the task at hand. Together, the idea of mathematical resilience in problem solving is evident in the framework and makes it

useful to teachers to develop mathematical tasks. To develop these tasks, teachers often rely on the resources of their curricular materials. A study using the LOD framework analyzed the cognitive demand of probability questions found in textbooks for middle grade mathematics. In Jones & Tarr's study multiple coders used Smith and Stein's LOD framework shown in Figure 2.2 to code mathematics tasks as one of the four LOD categories. This was done by interpreting a task based on the complexity of mathematics needed for a student to complete the problem. The results of this research found the majority (84.5%) of textbooks questions were considered to have low LOD for the topic of probability (Jones & Tarr, 2007). However the study found textbooks written in alignment with standards-based learning was the exception in the textbooks analyzed, more frequently asking questions achieving a higher LOD. These more current standards-based textbooks had a higher frequency of high LOD questions, however the overall distribution of each LOD had experienced little change across all textbooks used in the study.

Jones & Tarr's research also made tangible distinctions between lower and higher LOD through examples found in textbooks. A lower LOD has a narrow approach in the solution strategy, and in order to correctly answer the question, a student was aware of the procedures required, or of a formula applied. An example of the lower LOD questions found by Jones & Tarr is displayed in Figure 2.3. The mathematics asked of students in Figure 2.3 only required an algorithmic approach to the problem, leaving little ambiguity in the solution, and did not connect to conceptual ideas in mathematics.

3. Tell how to find the probability of two dependent events.

In a bag there are 5 red marbles, 2 yellow marbles, and 1 blue marble. Once a marble is selected, it is not replaced. Find the probability of each outcome.

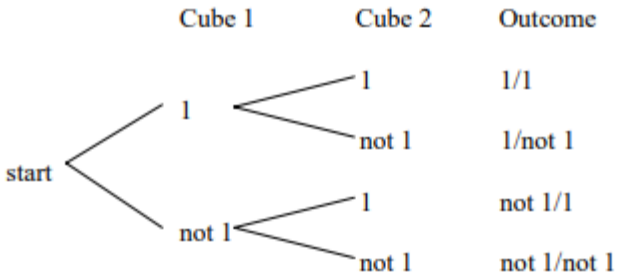
10. a red marble and then a yellow marble
11. a blue marble and then a yellow marble
12. a red marble and then a blue marble
13. any color marble except yellow and then a yellow marble
14. a red marble three times in a row

From *Mathematics: Applications and Connections, Course 3* © 1998, Collins, Dristas, Frey-Mason, Howard, McClain, Molina, et al. Published by Glencoe/McGraw-Hill. Used by permission.

Figure 2.3 Lower LOD (Jones & Tarr, 2007)

Examples of higher LOD are displayed in Figure 2.4. Two tasks are presented as examples in which a student would be required to elicit high LOD. The first task in Figure 2.4 required students to analyze the process and determine the legitimacy of a solution. Students reason about and make connections between a representation showing probability and then determine the errors or lack thereof in the theoretical student's explanation. The solution a student gives to this prompt could vary in the depth of understanding because the student's explanation would be based on the knowledge they possess, rather than a lack of knowledge. The second task in Figure 2.4 has a student develop a method to theorize the probability of an event happening. The task structured the work required of the student but left the development of the mathematical ideas up to the student. This leads to genuine mathematical thinking from the student.

19. Tricia wants to determine the probability of getting two 1s when two number cubes are rolled. She made a counting tree and used it to list the possible outcomes.



She says that, since there are four possible outcomes, the probability of getting 1 on both number cubes is $\frac{1}{4}$. Is Tricia right? Why or why not?

Tawanda's Toys is having a contest! Any customer who spends at least $10 receives a scratch-off game card. Each card has five gold spots that reveal the names of video games when they are scratched. Exactly two spots match on each card. A customer may scratch off only two spots on a card; if the spots match, the customer wins the video game under those spots.

Problem 6.1

If you play this game once, what is your probability of winning? To answer this question, do the following two things:

A. Create a way to simulate Tawanda's contest, and find the experimental probability of winning.

B. Analyze the different ways you can scratch off two spots, and find the theoretical probability of winning a prize with one game card.

Problem 6.1 Follow-Up

1. a. If you play Tawanda's scratch-off game 100 times, how many video games would you expect to win?
   b. How much money would you have to spend to play the game 100 times?
2. Tawanda wants to be sure she will not lose money on her contest. The video games she gives as prizes cost her about $15 each. Will Tawanda lose money on this contest? Why or why not?
3. Suppose you play Tawanda's game 20 times and never win. Would you conclude that the game is unfair? For example, would you think that there were not two matching spots on every card? Why or why not?

Figure 2.4 Higher LOD (Jones & Tarr, 2007)

Pilot Study

Using the aspects of DOK, LOD, and item difficulty, a pilot study was conducted with the teachers in the ROOT project during the fall teaching studies in the second year of the project. The study was based on findings from Gore & Gitlin (2004) in considering the differences between approaches in work around assessment of researchers and teachers. To reduce differences in assessments in the study, we provided teachers with high-quality resources to assist them with assessing the results of their teaching studies. However, we also wanted to honor their knowledge and expertise in developing assessments. The teachers were asked to prepare an end-of-unit CA consisting of 5-7 content specific assessment items and 1-2 modeling and problem solving assessment items administered as a pre, post, and optional mid assessment. Teachers were provided sample items from SBAC Smarter Content Explorer for grade level 6-8 mathematics. The item bank for each grade level allowed teachers to choose SBAC-sourced items focused on 3 - 4 mathematics topics and items focused on modeling/problem solving for their assessments. Teachers were encouraged to select items from these assessments but were welcome to prepare their own items either as a supplement or to fully source their assessments. They were also permitted to use Idaho Standards Achievement Test (ISAT) Interim Assessment Blocks (IABs) or focused IABs in place of preparing an assessment.

The research questions considered in the pilot study were the following:

- What item sources did teachers use?
- Was availability of items for particular math topics a potential factor in item sourcing?
- For the teacher-sourced items, what was the cognitive demand of assessment items?

Teachers' assessments were sourced from (a) solely SBAC items, (b) SBAC and teacher-sourced items, (c) solely teacher-sourced items, and (d) IAB items. The results of the item sources preferences used by the course taught are displayed in Figure 2.5. Solely teacher-sourced item assessments were favored, making up 49% of the assessments, which are the blue bars. This was followed by 26% of assessments composed of a mixture of SBAC and teacher-sourced items, which are the orange bars. The SBAC items were incorporated into the assessments of 41% of teachers, which are both the light gray and orange bars, making it less favored than teacher-sourced items.

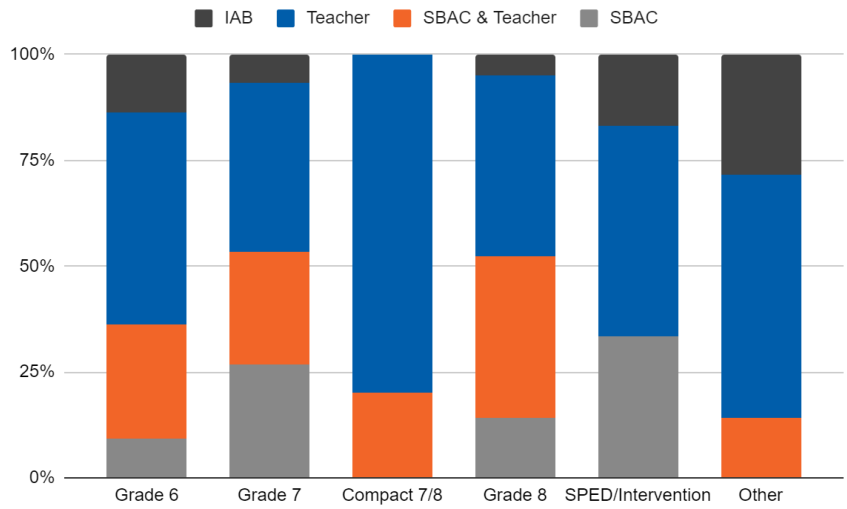


Figure 2.5 Pilot Study Assessment Preferences by Courses

The availability of items for particular math topics was analyzed. From the solely teacher-sourced items, 65% of teachers taught a mathematical topic which was not provided in the item bank from the research team. The remaining 35% of teachers taught a mathematical topic where items were available in the item bank, but preferred to use their own teacher-sourced items. Each item of the teachers' assessments were documented by the research team. SBAC items were given a unique item code describing

the grade level and mathematical content. Teacher-created assessment items were given an item code based on what the teacher reported as the appropriate grade level and mathematical topic of the assessment.

Finally, the cognitive demand of the individual teacher-sourced assessment items was measured using the DOK and LOD framework. A rubric was developed from the language of DOK and LOD describing each level of each framework and was used to assign codes for each assessment item. The rubric can be found in Appendix A. Within each grade level a simple random sample of 30 assessment items was collected and coded using the DOK framework and LOD framework. The results of the cognitive demand of all assessment items are displayed in Figure 2.6. From the mosaic plots, the LOD of teacher-sourced items shows items tended to address mathematical concepts through the use of procedures by the two larger areas of *procedures without connections* and *procedures with connections* in Figure 2.6. Each makes up 46% of the assessments analyzed. There was minimal variability in the LOD framework measured by the outcomes of DOK levels, with a range of variability between 0.51 - 0.61. The DOK framework tended to require an approach that aligned with the lower levels of the DOK framework, represented by the light gray and orange sections. The levels of *recall* and *skill/concept* made up 41% and 31% of assessment items respectively. The DOK framework measured by outcomes of LOD had more variability with a range of 0.24-0.49. Further analysis was conducted for each grade level, producing similar results to that seen in Figure 2.6. For all grade levels, there was a statistically significant association between the two frameworks, measured via Cramer's V and was a strong association of 0.67.

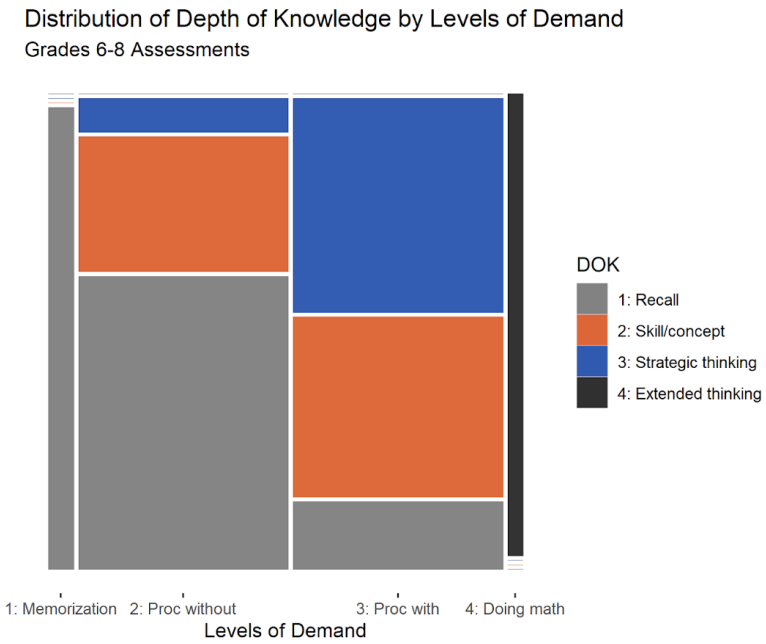


Figure 2.6 Pilot Study Distribution of DOK by LOD - Grades 6-8

The results of the pilot study show assessment items' LOD tended to address *procedures without connections* and *procedures with connections*, and DOK tended to address *recall* and *skill/concept*. This is described as an overall mid-level cognitive demand for assessment items. While there was little variability within the LOD framework, the DOK framework had more variability. The variability of the DOK framework meant teachers had greater range in how students were asked to respond to assessment items, but the most common type of response solicited was one that required *recollection* or the use of a *skill/concept* learned. From these observations of the most common categories of LOD and DOK , a question was raised as to whether teachers' assessments as a whole would result in a similar distribution seen in Figure 2.6, or if teachers' assessments are more variable.

CHAPTER THREE: METHODOLOGY

Theoretical Framework

The theoretical framework of this study is based on two pre-existing frameworks, Webb's DOK framework and Smith and Stein's LOD. The DOK framework was used to measure the level of difficulty experienced by the student from the assessment items (Webb, 2002). This was identified through the cognitive process elicited by an assessment item. This was often seen in the phrasing of assessment items (e.g. calculate, compare), however the intended reasoning of an assessment item was the factor determining the DOK level selected.. The LOD framework was used to measure the depth of mathematical thinking required of the students from the assessment items. This framework focused on the complexity of the mathematical thinking and conceptual understanding required of a student by an assessment item.

Research Questions

This study examined the following research questions in order to gain insight on teacher practices of CAs.

- How do assessments compare to each other in terms of LOD and DOK when items are chosen for a summative CA?
- How much variability exists within and between assessments in terms of LOD and DOK?

The research questions measured information on how teacher CA practices compare to one another and the variability between assessments. These measurements

were also analyzed based on grade level, determining if there are differences in assessments as students progress through 6th, 7th, and 8th grade.

Study Parameters

Participants

The participants in this study were grades 6-8 mathematics teachers in the northwestern United States (N = 79). The participants in the study all taught in public schools, with nearly all of them working in brick-and-mortar schools and one teacher working in a virtual public charter school. Teachers' mathematics instruction often spanned multiple grades (29 taught grade 6, 31 taught grade 7, and 36 taught grade 8) and courses (36 taught one course, 31 taught two courses, and 12 taught three or more courses). During the first year of the ROOT project, demographics of the schools and teachers were recorded. The teachers were from 34 schools within 22 school districts and worked in a variety of school settings, both in terms of students' socio-economic status (mean eligibility for federal free or reduced school lunch was 58%, SD = 21%) and locale type (31% rural, 69% suburban or small city). Teacher demographics showed variability in mathematics teaching experience (mean = 9.8 years, SD = 7.4, Range = 1 to 32), and they primarily self-identified as female (77%) and white (96%). The majority, 57%, of teachers' highest achieved degree was a bachelor's degree, though 40% held a master's degree, and 2% held an Ed.S.

Context

During year two of the project, the ROOT research team conducted individual crossover teaching studies based on teachers' sequencing of two EAC and/or SOS instructional practices, or individual comparison teaching studies based on teachers'

sequencing an EAC or SOS instructional practice and the teachers' regular instructional practices. For the crossover and comparison studies, teachers started with a pre assessment administered to all of their students. This was followed by the teachers implementing both of the strategies they had selected by dividing their students into two groups. This was most commonly done by assigning a class period an instructional strategy. The mid assessment was administered at the halfway point in the study, and following the mid assessment the instructional strategies were switched between the groups of students. To close the study a post assessment was administered. The teachers completed 30 - 45 days using the instructional practices. The administered assessment consisted of 5 common assessment items for the appropriate grade level and 5 content specific items. The common assessments were provided by the ROOT research team, and the content specific items were provided by the teacher. The data collected for the assessments was completed in the spring semester of the second of three years in which the teachers participated in the ROOT project.

Data Collection

Participants in the study were asked to administer a 10 question assessment as a pre, mid, and post assessment for a unit in their respective mathematics content. The 10 question assessment was composed of 5 common assessment items for each grade level, provided by the researcher team, and 5 topic-specific assessment items, provided by the teacher. The majority of teachers assessed 5 teacher-created assessment items, but some assessed more or less than 5. A record was kept of the assessment items administered by each teacher, and the student performance data on the pre, mid, and post assessments.

The record of assessment items consisted of 47 unique assessments and 272 assessment items for middle school mathematics content that were administered by 69 teachers.

For the 5 common assessment items, the research team selected items representing a spread based on percent correct from the pre assessments. The pool of potential common-items were considered from assessment items used during the pilot study (described in chapter 2). Item difficulty was measured based on the percent correct from the pre-assessment of the pilot studies. For each grade level, the targeted item difficulty was items near the 10th, 30th, 50th, 70th, and 90th percentiles. Additional criteria considered for the common-items was (a) variety in mathematical content assessed, and (b) reasonable (10 - 30) percent gains between pre and post-assessments.

Coding

All teacher-created assessment items and common-items were coded based on the DOK and LOD frameworks. A rubric was developed based on the language from DOK and LOD which can be found in Appendix A (which is also the same rubric used in the pilot study). For the coding process, assessment items were read and interpreted for each framework. When coding for DOK, items were considered based on the type of solution the assessment item explicitly asked for, or implied for. For example, when assessment items asked students to select all that apply (as long as selections required a conceptual understanding), the item was coded as DOK 2 because of the implications of categorizing solutions. When coding for LOD, items were analyzed for the mathematical process a student goes through in order to successfully complete the problem. For example, when a student was asked to solve a multi-step, single variable equation the item was coded as LOD 2 because the student is applying an algorithmic process that does not connect to

the concepts behind performing inverse operations to solve for unknown variables. Oftentimes, context made an assessment item be coded as a LOD 3 item, however if the context of the question did allow ambiguity in the mental process, it would not be coded as an LOD 3 item. One of the four levels of the DOK and LOD framework were recorded for each item. Based on student performance data, item difficulty was recorded from the percent correct from the pre assessment. The use of the item difficulty was to verify the categorical coding of DOK and LOD was within reason of how students performed on the assessment items.

Common-items provided by the research team were coded based on their LOD and DOK. Table 3.1 displays each item's grade level, item ID, DOK code, and LOD code. For DOK the common-items consisted of 13% (2) DOK 1, 60% (9) DOK 2, and 27% (4) DOK 3. For LOD the common-items consisted of 40% (6) LOD 2, and 60% (9) LOD 3.

Table 3.1 Common Item Coding Results

Grade	Item ID	DOK Code	LOD Code
6	G6RPR24	1	2
	G6RPR36	3	3
	G6NS11	2	3
	G6In3	2	3
	G6RPR6	1	2
7	G7EE12	2	2
	G7EE17	2	3
	G7EE16	2	3
	G7RPR26	3	2
	G7NS2	2	2
8	G8SE23	2	3
	G8LR20	3	2
	G8PSM8	2	3
	G8EE11	2	3
	G8LR7	3	3

Data Analysis

In the analysis of the data, some levels of the DOK and LOD framework were not included in the analysis. Across all assessment items, DOK 4 and LOD 4 were not used.

This is based on the nature of the rubric descriptions for each code, which can be found in Appendix A. DOK 4) *Extended Thinking* was described by the rubric with language indicating an experimental approach toward the problem. None of the assessment items required students to take an approach to the problem requiring DOK 4. The same situation was apparent for LOD 4 as well. The description of LOD 4) *Doing Mathematics* required a task where students' mathematical process required self-regulation and openness in the answer to the question. All assessments provided by the teachers were traditional CAs that did not approach assessment items in this way, although the teachers in the study provided classroom instruction in this way. The results of the pilot study predicted both codes would have low frequencies, therefore making both levels obsolete for the study. Additionally, evidence previously described in Chapter 2 explained teachers' perspectives of the appropriateness of DOK levels in CAs (Hess et al., 2009). The results of frequency of DOK by LOD are reported in Table 3.2. The most common assessment item for LOD was LOD 2) *procedures without connections* with a proportion of 56.8% (281) of items. The most common assessment item for DOK was DOK 2) *skill/concept* with a proportion of 48.3% (239) of items.

Table 3.2 Coding Results of DOK by LOD

	LOD				
DOK	1	2	3	4	Total
1	1.6% (8)	29.3% (145)	0	0	30.9% (153)
2	0.2% (1)	17.8% (88)	30.3% (150)	0	48.3% (239)
3	0	9.7% (48)	11.1% (55)	0	20.8% (103)
4	0	0	0	0	0
Total	1.8% (9)	56.8% (281)	41.4% (205)	0	100% (495)

After all assessment items were recorded, items were then matched to each unique assessment that was administered in the teaching studies. The frequencies of the levels of DOK and LOD were measured by the percent of questions at the levels of DOK 1, 2, and 3 as well as LOD 1, 2, and 3 for each assessment.

CHAPTER FOUR: FINDINGS

Analysis by Assessments

Assessment Distribution by LOD

To answer the question of how assessments compare to each other in terms of a) LOD and b) DOK when selecting items for a summative CA, data was analyzed from the assessments teachers provided for the study. Figure 4.1 displays the distribution of each category of LOD for all assessments using Figure 4.1a) all items that make up an assessment and Figure 4.1b) only the teacher-created items that make up an assessment. The nature of the study required a report on both the assessments made up of all items and assessments made up of only teacher-created items. Measuring the assessment data under the two perspectives of all items and teacher-created items only shows information based on the variability of how teachers made choices for their assessment design. The displays in Figure 4.1 show the categories of LOD 1, LOD 2, and LOD 3 on each axis of the trivariate plot (LOD 4 was not included because no assessment items received an LOD 4 code). Each point in the display represents an assessment from the study, and the colors of red, green, and blue represent grade 6, 7, and 8 assessments respectively. Each point in the plot describes what percentage of each level of the particular framework is represented by the items in an assessment. For example, a point lying in the center of Figure 4.1 would represent that assessment items were 33.3% LOD 1, 33.3% LOD 2, and 33.3% LOD3. In both plots of Figure 4.1 the majority of items lie on the axis that LOD 2 and LOD 3 share. This location indicates the assessments were composed of primarily

LOD 2 and LOD 3 items addressing primarily *procedures without connections* items, but also addressing *procedures with connections* items respectively.

When examining the difference between assessments made up of all items versus only teacher-created items the teacher-created item assessments are shifted closer to the vertex of LOD 2. This shift indicates the majority of teacher-created items assess students at a LOD 2 level, and less frequently addressed items at a LOD 3 level. Therefore indicating the teacher-created items were less conceptual. Because the points in the distribution of Figure 4.1b are more spread out compared to Figure 4.1a, this provides evidence of a greater difference in the proportions of assessments' LOD distribution when the common-items were removed from the analysis. That is, there is more variability in assessments with only teacher-created items. In Figure 4.1b assessments lie on the axis that share LOD 2/3 as well as LOD 1/2. For each group of assessments at a particular grade level there is a larger spread when common-items were removed. Because of this spread, common-items became an anchor point, causing each assessment to have an increased frequency of LOD 3 items, and often meant assessments were closer to a 50 - 50 split between LOD 2 (*procedures without connections*) and LOD 3 (*procedures with connections*). But when the common-items were removed the split moved closer to 70 - 30 between *procedures without connections* and *procedures with connections*.

Figure 4.1a LOD Distribution of Assessments by Grade Level

All Items

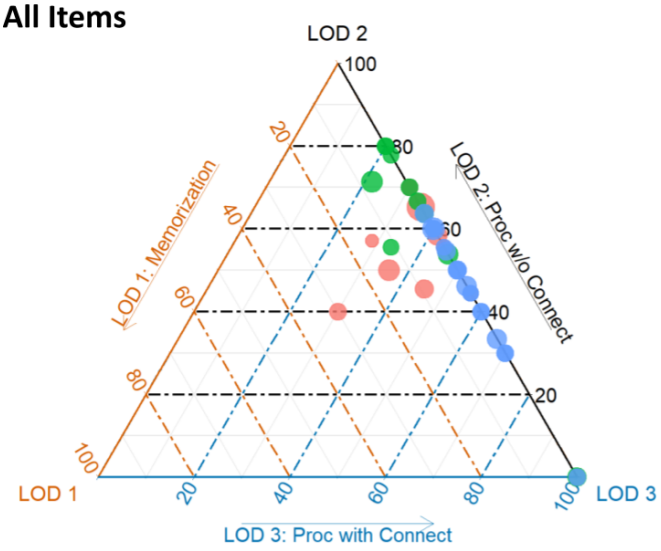
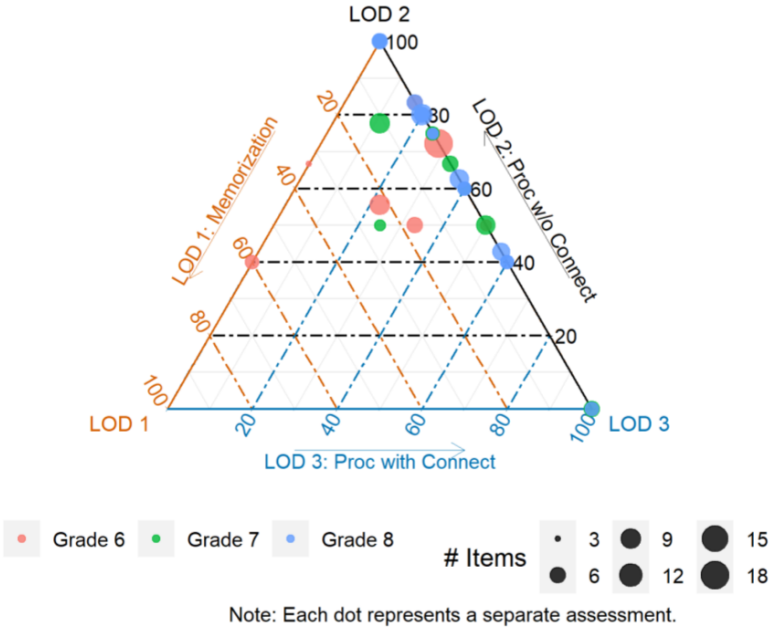


Figure 4.1b LOD Distribution of Assessments by Grade Level

Teacher-Created Items Only



Note: Each dot represents a separate assessment.

Figure 4.1 LOD Distribution of Assessments by Grade Level

Figure 4.1 also displays trends within grade levels. For assessments including all items, grade 8 assessments reported the highest frequencies of LOD 3 items (52.2%), grade 7 assessments reported lowest frequencies of LOD 3 (33.3%), and grade 6 reported a mid-level of LOD 3 items (36.8%). For assessments with only teacher-created items grade 8 again reported the highest frequency of LOD 3 items (28.1%), grade 7 assessments reported a mid-level frequency of LOD 3 items (26.8%), and grade 6 assessments reported the lowest frequency of LOD 3 items (17.7%). Table 4.1 displays the average of the LOD 1, LOD 2, and LOD 3 item distributions for assessments including all items (and assessment including only teacher-created items) by grade level.

Table 4.1 Average LOD by Grade Level for All Items (and Teacher-Created Items)

Grade Level	LOD 1: Memorization	LOD 2: Procedures w/o Connections	LOD 3: Procedures w/ Connections
Grade 6	4.0% (7.3%)	59.2% (75.0%)	36.8% (17.7%)
Grade 7	1.4% (2.8%)	65.3% (70.4%)	33.3% (26.8%)
Grade 8	0.0% (0.0%)	47.8% (71.9%)	52.2% (28.1%)
All grades	1.8% (3.4%)	56.8% (72.6%)	41.4% (24.0%)

Assessment Comparison by LOD 3 and Grade Level

The percentage of LOD 3 items for each assessment was computed to make comparisons between individual assessments and across the assessments for each grade. Figure 4.2 presents the frequency of LOD 3 items in each assessment for Figure 4.2a) all items that make up assessments and Figure 4.2b) only teacher-created items. In Figure 4.2, the percentage of LOD 3 assessment items are recorded for each assessment. Each row represents a grade level, and is also indicated by the colors of red, green, and blue for

grade 6, 7, and 8 respectively. Each point below the horizontal line represents an assessment, and the density curve above shows a generalization of the distribution of the LOD 3 items at each grade level. Figure 4.2a indicates a uniform distribution for grade 8 assessments and a multimodal distribution for grade 7 and grade 6 assessments both of which are slightly skewed, with grade 7 skewed right, and grade 6 skewed left. For Figure 4.2a assessments for grade 8 had a median value of nearly 50% of items assessing at the LOD 3 level, while grade 7 and grade 6 had a median of nearly 33% of items assessing at the LOD 3 level (the median is indicated by the black dot). However, in Figure 4.2b the common-items are removed and this dramatically affects each grade level's LOD 3 distribution.

One of the more prominent observations is the evidence of assessments with 0% LOD 3 items for each grade level. There are three assessments each in grades 7 and 8 with no LOD 3 items in the teacher-created items, and 7 in grade 6. Additionally, each grade level experiences a large decrease in the percentage of LOD 3 items. The median percentage of LOD 3 items for each grade level becomes slightly less than 20% for grade 8, slightly more for grade 7, and nearly 10% for grade 6 when the common-items are removed. The most dramatic decrease in grade level medians is from grade 8 assessments, while the smallest decrease in grade level medians is for grade 7 assessments. Despite the changes between all items and teacher-created items only, grade 6 assessments resulted in the lowest median amount of LOD 3 frequencies in assessments for both Figure 4.2a and Figure 4.2b. Despite frequencies of LOD 3 assessments being lower between Figure 4.2a and Figure 4.2b, evidence of assessments with at least 50% of LOD 3 items is displayed in all grade levels, with one each in grades 6 and 7, and three in

grade 8. This indicates that some teacher-created item assessments did obtain a nearly 50-50 split without the use of common-items.

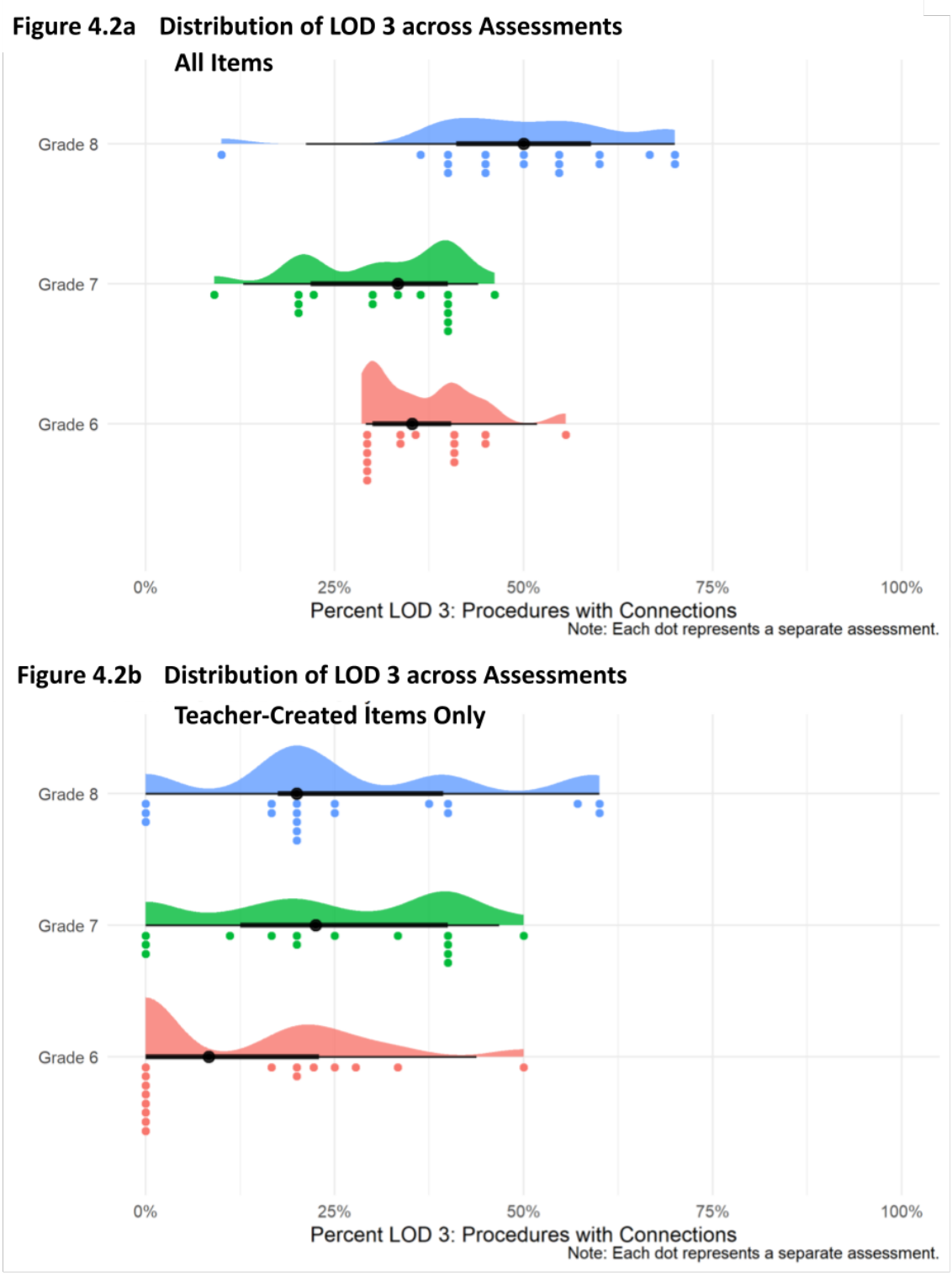


Figure 4.2 Distribution of LOD 3 across Assessments

Assessment Variability by LOD

Using the percentage of LOD classifications of assessment items, the variability of LOD within assessments was analyzed using mean absolute deviations. For assessments that included common-items, the mean of the percentage of the assessment that consisted of each LOD was computed. The standard deviation was computed to understand the variability within assessments' percent distribution of LOD. The sample size was 47 assessments. The mean percentage of items by LOD within an assessment is necessarily 33.3% because the total percentage is 100% and there are three LOD. The average of the mean absolute deviation within assessments was 21.0%. The mean value of 33.3% with the high variability of 21.0% means the LOD between assessments could have large differences.

For assessments consisting of only teacher-created items, the same measures of center and variability were examined. The sample size was 46 (because one assessment was made entirely of common-items). The mean percentage of LOD within assessments is again 33.3%, but the variability for teacher-created items only is 26.6%, which is higher than the variability from assessments that included all items.

Using the percentage of LOD classifications of assessment items, the variability in LOD of items between assessments can be analyzed using mean absolute deviations. For assessments including all items, the mean percentage of items for that LOD was computed. The mean absolute deviation was computed for each LOD to understand the variability between assessments. The following are the results of this analysis; N = 47, LOD 1 M = 1.8%, LOD 1 MAD = 3.2%; LOD 2 M = 56.8%, LOD 2 MAD = 9.5%; LOD 3 M = 41.4%, LOD 3 MAD = 9.4%. These results indicate LOD 2 items often made up

more than half of an assessment, and the relatively small MAD indicates the variability in LOD 2 items was relatively minimal. LOD 3 items often made up less than half of the assessment and again, the relatively small MAD indicates variability in LOD 3 items was relatively minimal. Overall, there was little variability in LOD across assessments.

For the analysis that included only teacher-created items, the same measures of variability were calculated. The following are the results of this analysis; N = 46, LOD 1 M = 3.4%, LOD 1 MAD = 6.4%; LOD 2 M = 73.2%, LOD 2 MAD = 16.5%; LOD 3 M = 23.4%, LOD 3 MAD = 15.3%. From these results, LOD 2 made up the majority of the items on the assessments and the variability increased when compared to the analysis that included the common-items. LOD 3 items often made up less than a quarter of the assessment items and there was quite a bit of variability around that mean. With common items removed, variability in LOD and frequency of LOD 3 both decreased.

Across all assessments, the range was 0% - 30% LOD 1 items, 20% - 80% LOD 2 items, and 20% - 70% LOD 3 items. For only teacher-created assessments, the range was 0% - 60% LOD 1 items, 40% - 100% LOD 2 items, and 0% - 60% LOD 3 items. The large differences of ranges indicates the use of common items provided some assessments with more conceptually focused assessment items.

Assessment Distribution by DOK

Figure 4.3 displays the distribution of the categories of DOK for all assessments by both Figure 4.3a) all items that make up an assessment, and Figure 4.3b) teacher-created items only that make up an assessment. The displays show the categories of DOK 1, DOK 2, and DOK 3 on each axis of the trivariate plot (DOK 4 was not included because no assessment items received a DOK 4 code). Each point in the display

represents an assessment from the study, and the colors of red, green, and blue represent a grade 6, 7, and 8 assessment respectively. The trivariate display of Figure 4.3a represents DOK being relatively evenly distributed across an assessment's items because the points lie near the middle of the plot. There is a cluster of points near the DOK 2 vertex of the plot, meaning that the type of question that is more dominant in assessments is one asking about a *skill/concept*. In Figure 4.3b the distribution of DOK is slightly more spread out, implying greater variability when the common assessment items are removed. Figure 4.3b also shows some assessments having variety in the DOK level, and others shift closer to the DOK 1 vertex, meaning assessments are based more upon *recall* based items. The inclusion of common-items made the majority of assessments items addressed DOK 2 or DOK 3 more frequently, causing assessments to approach items based on *skill/concept* or on *strategic thinking*.

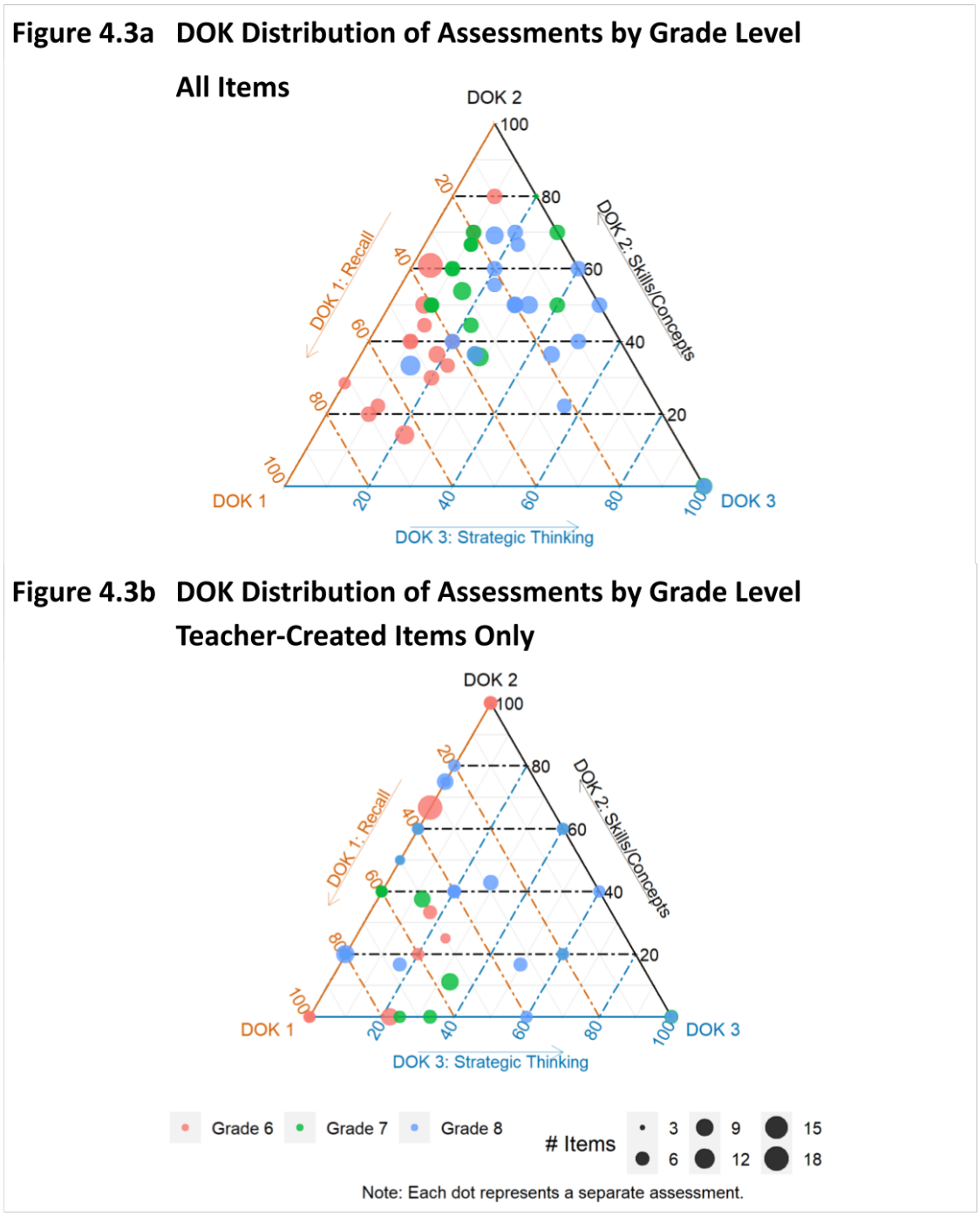


Figure 4.3 DOK Distribution of Assessments by Grade Level

Figure 4.3a displays grade level clusterings of DOK frequencies from the red, green, and blue points. Grade 6 assessments, represented by the red points, tended to cluster near DOK 1 and DOK 2. Grade 7 assessments, represented by the green points,

tended to cluster near DOK 2. Grade 8 assessments, represented by the blue points, tended to cluster near DOK 2 and DOK 3. In Figure 4.3b, the distribution of assessments when the common-items were removed and only teacher-created items were analyzed resulted in a large change from Figure 4.3a. In Figure 4.3b all grade level assessments shifted toward the DOK 1 vertex, and the clusterings by grade level became less prominent. The average of each level of the DOK framework by grade level assessments with all items (and only teacher-created items) and across all assessments can be displayed in Table 4.2.

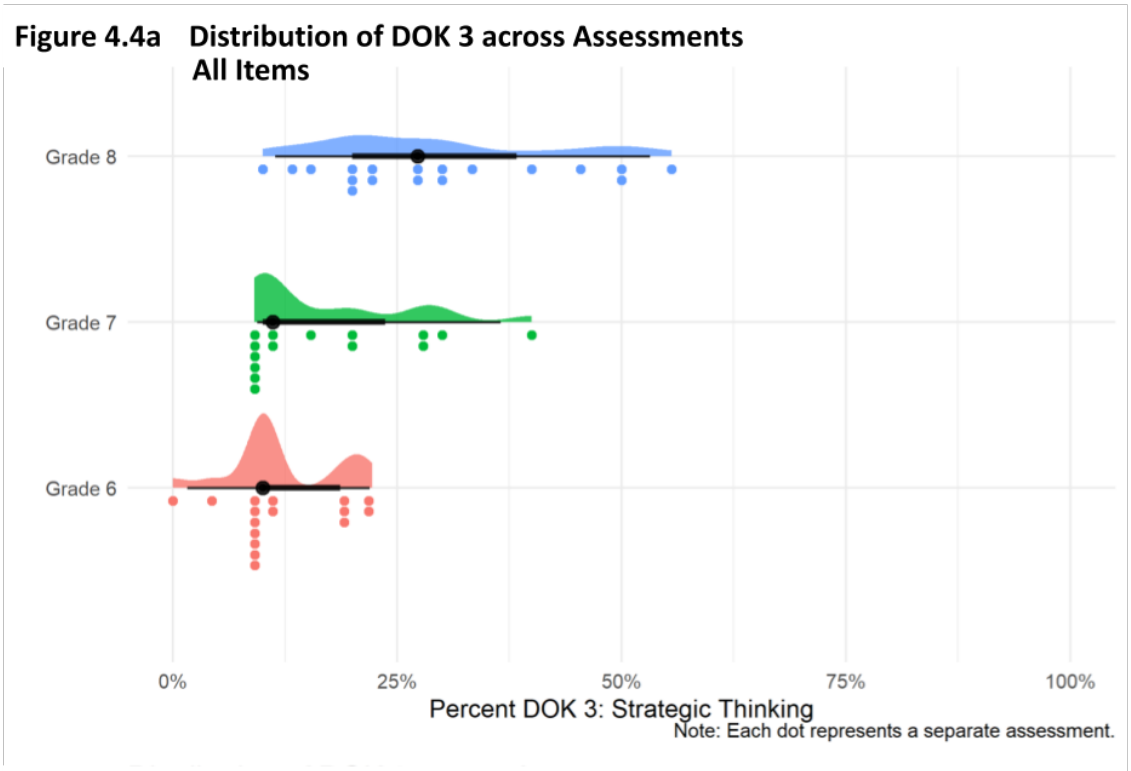
Table 4.2 Average DOK by Grade Level for All Items (and Teacher-Created Items)

Grade	DOK 1: Recall	DOK 2: Skill/Concept	DOK 3: Strategic Thinking
Grade 6	45.4% (50.0%)	42.5% (43.8%)	12.1% (6.3%)
Grade 7	26.2% (52.1%)	54.6% (29.6%)	19.2% (18.3%)
Grade 8	20.6% (38.5%)	48.9% (39.6%)	30.6% (21.9%)
All grades	30.9% (46.4%)	48.3% (38.4%)	20.8% (15.2%)

Assessment Comparison by DOK 3 and Grade Level

The distribution of DOK 3 across grade levels resulted in lower frequencies for each grade level than LOD 3. Figure 4.4a displays a uniform distribution for grade 8 assessments with a median frequency near 25%, a right skewed distribution for grade 7 assessments with a median near 12.5%, and a multimodal distribution for grade 6 assessments with a median near 12.5% (the median is indicated by the black dot). However, in Figure 4.4b, we see each median value decrease, and each distribution changed to a left skewed distribution. This is caused by the assessments that were coded

with 0 DOK 3 items in the assessment. For teacher-only assessments for grade 7 and 8, 7 assessments were coded with no DOK 3 items; and for grade 6, 11 assessments were coded with no DOK 3 items. For Figure 4.4b the median value of grade 8 assessments is near 18%, grade 7 assessments is near 7%, and grade 6 assessments is at 0%. While the grade level ranking still stayed the same between Figure 4.4a and Figure 4.4b, it is evident that common-items provided an increased higher-level thinking based on DOK levels for all grade level assessments.



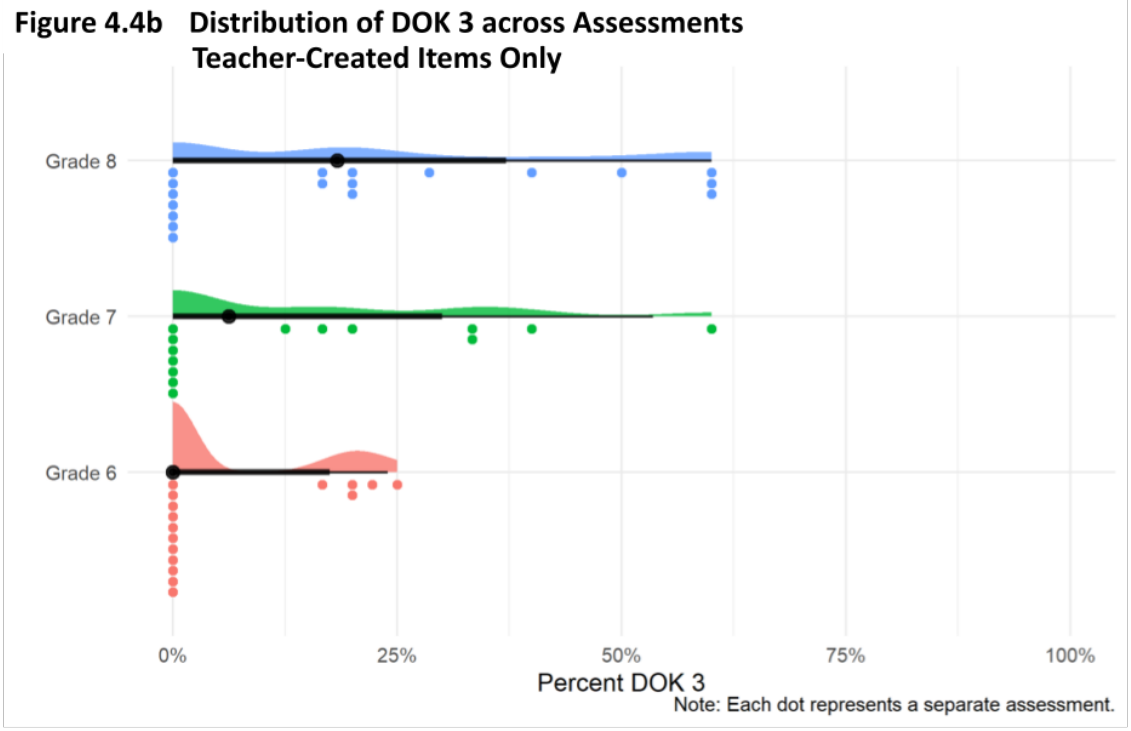


Figure 4.4 Distribution of DOK 3 across Assessments

Assessment Variability by DOK

Using the percentage of items at each DOK classification, the variability of DOK within assessments was analyzed through mean absolute deviations. For assessments including all items, the mean was computed to summarize the data for the assessment distribution of DOK variables. Mean absolute deviation was computed to understand the variability within assessments' distribution of DOK. The sample consisted of 47 assessments. The mean percentage of items by DOK within an assessment is necessarily 33.3% because the total percentage is 100% and there are three DOK. The mean value was 33.3% and the mean absolute deviation was 10.0%. The comparatively small variability indicates teachers represented a variety of DOK in their assessments and there was balance of the types of thinking elicited in the assessments.

For the analysis that included teacher-created items only mean and mean absolute deviations were calculated. The sample was 46 (because one assessment only included the common assessment items). Again, the mean was 33.3%, and the mean absolute deviation was 12.0%. The variability had similar results between all items and teacher-created items only, meaning that even without the common-items, most assessments are balanced between DOK. The data from Table 4.2 indicates the slight change in variability with the lower frequency of DOK 3 for teacher-created items only.

Using the percentage of DOK classifications of assessment items, the variability of DOK between assessments was also analyzed using mean absolute deviation. For the analysis that included common items, the mean was computed to summarize the data for the levels of DOK. Mean absolute deviation was computed to understand the variability between assessments' using the levels of DOK variables. The results were N = 47, DOK

1 M = 30.9%, DOK 1 MAD = 15.3%; DOK 2 M = 48.3%, DOK 2 MAD = 13.0%; DOK 3 M = 20.8%, DOK 3 MAD = 10.3%. This means DOK 2 was most frequently assessed, with the least variability. DOK 3 items were more rare, also with low variability. For most assessments DOK 3 was addressed around one third of the time, but usually did not exceed this. For DOK 1 items, on average it was addressed one third of the time, however it had a mid-level variability meaning DOK 1 could be addressed quite frequently or infrequently for by assessments.

For assessments including teacher-created items only, the mean absolute deviation was calculated. The results were N = 46, DOK 1 M = 46.7%, DOK 1 MAD = 20.6%; DOK 2 M = 37.9%, DOK 2 MAD = 20.4%; DOK 3 M = 15.3%, DOK 3 MAD = 16.9%. This means the majority of items were DOK 1, making up nearly half of assessment items, although DOK 1 had a somewhat large variability. This is similar to results from the analysis of all items, indicating the differences come from the frequency of DOK 1 assessed for only teacher-created items. Similarly, although DOK 2 was no longer the most frequently assessed DOK level for teacher-created items only, it still makes up a large proportion of the assessments' items. DOK 2 also had considerable variability, meaning assessments have a variety in the number of DOK 2 items assessed. For DOK 3 the mean tells us there are a low number of assessment items addressing DOK 3, and has a moderate variability. This low average and moderate variability together mean that some assessments addressed DOK 3 at a level that is evenly distributed, but the majority of assessments addressed DOK 3 less frequently.

The range of frequencies of DOK levels was also recorded. Across all assessments, the range was 0% - 71.4% DOK 1 items, 14.3% - 80% DOK 2 items, and

0% - 55.6% DOK 3 items. For only teacher-created assessments, the range was 0% - 100% DOK 1 items, 0% - 100% DOK 2 items, and 0% - 60% DOK 3 items. Because the ranges become wider with teacher-created items only, the use of common-items may have helped ensure a variety of cognitive demand requested of students during the assessment.

CHAPTER FIVE: SUMMARY AND CONCLUSION

Assessments in this study were able to obtain LOD and DOK distributions that would be expected from results of previous studies using these constructs (Yuan & Le, 2012; Hess, et al., 2009; Jones & Tarr, 2007) however, assessments were not reaching the goal of being representative of the state standards to which they were meant to be aligned. There were large differences between how assessments were composed when all items were included and when only teacher-created items were included. One standout of this is based on the number of high level items, which for this study are LOD 3 items and DOK 3 items. There were many assessments that did not include either of these types of items when assessments were solely based upon teacher-created items only. This indicates teacher-created assessments may not address higher-level thinking to a high degree, but rather may focus on less conceptual ideas. Based on the teacher's training in EAC and SOS for the project they were participating in, their instruction was focused on helping students gain conceptual understanding. Upon reflection of these results, the assessments designed may not have been well-aligned to the instructional practices the teachers were implementing. A caveat of this potential misalignment is based upon the lack of discussion of the assessments with teachers one-on-one. The teachers' perceptions of common-items could have influenced their choices in assessment design, causing the teachers to assess at lower levels with the intention of providing balance in the assessment. The answers to the questions of teachers' perceptions is beyond the scope of

this thesis and the research questions addressed, but motivates further research about teacher beliefs and practices of assessments.

When analyzing assessments including all items, LOD 3 was addressed by roughly half of the items on assessments. This distribution means that teachers valued both a procedural display of knowledge and a procedural with connections display of knowledge when it came to CAs. This result is not surprising in the context of a classroom because of the fast pace of a classroom and the amount of content a teacher presents in a year. There is also a factor of vertical alignment between grade levels and necessary skills needed for students to progress. Although the distribution is expected, the retention of procedural skills is cause for concern. This relates to intentionality of CAs, which at times is different from a teacher's beliefs about student learning. A teacher might feel pressured to check off boxes for students in order for them to progress, therefore making a procedural approach more common in assessments. As learning shifts more and more toward the emphasis of conceptual ideas, assessments must also shift in order to reflect how a student is learning in the classroom.

Distributions of DOK were relatively similar between the levels. However, DOK 3 had the lowest frequency. Again, this was not unexpected based on pressures of time and expectations of performance a school faces. DOK 3 focused on explanations of thinking. The teachers in this study had students explain mathematical concepts in their classroom instruction, however this expectation was less prominent in the CA format. Assessments created by SBAC outline a distribution of DOK levels aligned with the standards. Because one claim of SBAC is Concepts and Procedures, sections of the SBAC assessments items address DOK 1 and 2 only, however the other three claims of

Problem Solving, Communicating Reasoning, and Modeling and Data Analysis address assessment items using DOK 1, 2, 3, and 4. The use of all four levels of DOK requires students to engage with assessment items in a way that requires deeper explanation of thinking. One issue in providing more questions that require an explanation of thinking is related to the number of questions that are algorithmic or procedural. An explanation of a procedure is not explaining a mathematical concept. Because of this, without more examples of questions that are conceptual, there cannot be more items where it is necessary to explain one's thinking. This means there could be a relationship between the distribution of LOD items and the frequency of DOK 3 items. This relationship should be further explored in future research of these two frameworks.

Trends between grade levels showed grade 8 assessments had higher overall LOD and DOK levels, although grade 8 assessments were more variable in both DOK and LOD. This suggests some grade 8 teachers may be assessing at levels that are focused on concepts and require explanations of thinking, while others are not. While there was a lack of consistency for grade 8 assessments, assessments at the grade 6 and 7 levels displayed teachers' choices were more consistent with one another. However, with this consistency, assessments on average were assessed at lower levels of DOK and LOD than grade 8. Grade 6 and 7 assessed more consistently at procedural based questions (LOD 2) and tended to assess at items that required more of a *recall* or *skill/concept* based approach (DOK 1 and 2). Further analysis about mathematical content at each grade level could help answer questions about the differences between grade level assessments.

Certain limitations of the study still leave questions of teachers' task selection open to interpretation. The use of common-items could be a reason why teachers chose to

assess students at lower cognitive levels to provide a balance of questions they deemed reasonable. In future studies using these constructs, an interview process with teachers could address that limitation. Additionally, sampling more assessments at each grade level could help clarify the observed differences, as could collecting multiple assessments from each teacher to examine relationships between content and cognitive demand within teachers and across grade levels.

The way in which teachers assess their students helps teachers know what students are able to do. In this study, data indicated reasonable proportions of procedural and conceptual understanding for assessment items, and moderate variation of required knowledge. However, when those results were analyzed without the included common items, assessments tended to become more procedural and required thinking based on *recall* and *skills/concepts*. The lack of higher-level cognitive demand through establishing connections and thinking strategically suggests further questions about assessment design. Future study of teachers' beliefs and practices of assessments could help explain how researchers and practitioners can improve alignment between the cognitive demand of standards taught in the classroom and the cognitive demand of assessments.

REFERENCES

Bailey, A. L., & Heritage, M. (Eds.). (2008). *Formative assessment for literacy, grades K-6: Building reading and academic language skills across the curriculum*. Corwin press.

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. M. (1992). What do teachers know about testing and how did they find out. In *annual meeting of the National Council on Measurement in Education, San Francisco*.

Champion, J., Crawford, A., & Carney, M. (2021) Articulating Effective Middle Grades Instructional Practices in a Teacher-Researcher Alliance [Brief Report]. Psychology of Mathematics Education - North America (PME-NA) 42 Conference 2020, Mazatlán, Mexico and Virtual.

Darling-Hammond, L., Herman, J., Pellegrino, J., Abedi, J., Aber, J. L., Baker, E., ... & Steele, C. M. (2013). Criteria for high-quality assessment. Stanford Center for Opportunity Policy in Education, 2, 171-192.

Gore, J. M., & Gitlin, A. D. (2004). [Re]Visioning the academic-teacher divide: power and knowledge in the educational community. Teachers and Teaching: Theory and Practice, 19(1), 35-58. https://doi.org/10.1080/1354060032000017091 8

Green, K. E., & Stager, S. F. (1986). Measuring attitudes of teachers toward testing. *Measurement and Evaluation in Counseling and development*, *19*(3), 141-150.

Green, K. E., & Stager, S. F. (1987). Differences in Teacher Test and Item Use with Subject, Grade Level Taught, and Measurement Coursework. *Teacher Education & Practice*, *4*(1), 55-61.

Hess, K. K., Jones, B. S., Carlock, D., & Walkup, J. R. (2009). Cognitive Rigor: Blending the Strengths of Bloom's Taxonomy and Webb's Depth of Knowledge to Enhance Classroom-Level Processes. Online Submission.

Hiebert, J., & Grouws, D. A. (2007). The effects of classroom mathematics teaching on students' learning. *Second handbook of research on mathematics teaching and learning*, *1*(1), 371-404.

Jones, D. L., & Tarr, J. E. (2007). AN EXAMINATION OF THE LEVELS OF COGNITIVE DEMAND REQUIRED BY PROBABILITY TASKS IN MIDDLE GRADES MATHEMATICS TEXTBOOKS. Statistics Education Research Journal, 6(2).

McMillan J. H. (2013). Research on Classroom Summative Assessment In McMillan, J. H. SAGE handbook of research on classroom assessment. SAGE Publication, Inc., https://dx.doi.org/10.4135/9781452218649

Moss, C. (2013). Research on Classroom Summative Assessment In McMillan, J. H. SAGE handbook of research on classroom assessment. SAGE Publication, Inc., https://dx.doi.org/10.4135/9781452218649

Polikoff, M.S., Porter, A.C., & Smithson, J. (2011). How well aligned are state assessments of student achievement with state content standards? American Educational Research Journal, 48(4), 965–995.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental?. *Theory into practice*, *48*(1), 4-11.

Randel, B. & Clark, T. (2013). Measuring Classroom Assessment Practices In McMillan, J. H. SAGE handbook of research on classroom assessment. SAGE Publications, Inc., https://dx.doi.org/10.4135/9781452218649

Schneider, M., Egan K., & Julian M. (2013). Research on Classroom Summative Assessment In McMillan, J. H. SAGE handbook of research on classroom assessment. SAGE Publication, Inc., https://dx.doi.org/10.4135/9781452218649

Smarter Balanced Assessment Consortium (SBAC). (2019). Mathematics Summative Assessment Blueprint. Test Development. Smarter Content Explorer, https://contentexplorer.smarterbalanced.org/test-development

Smith, M. S., & Stein, M. K. (1998). Reflections on practice: Selecting and creating mathematical tasks: From research to practice. Mathematics teaching in the middle school, 3(5), 344-350.

Stein, M. K. (2000). Implementing standards-based mathematics instruction: A casebook for professional development. Teachers College Press.

Stein, M. K., Correnti, R., Moore, D., Russell, J. L., & Kelly, K. (2017). Using theory and measurement to sharpen conceptualizations of mathematics teaching in the common core era. *AERA Open*, *3*(1), 2332858416680566.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. Phi Delta Kappan, 77(3), 238.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758-765.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. *Journal of educational measurement*, *22*(4), 271-286.

Webb, N. L. (2002). Depth-of-knowledge levels for four content areas. Language Arts, 28(March).

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. Applied measurement in education, 20(1), 7-25. https://doi.org/10.1080/08957340709336728

Yuan, K., & Le, V. (2012). Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests. Santa Monica, CA: RAND Corporation.

Zhang, Z., & Burry-Stock, J. A. (1994). Assessment practices inventory. *Tuscaloosa, AL: The University of Alabama*.

APPENDIX

Coding Rubric – DOK

DOK Level	Description
Recall (level 1)	- recalling facts, definitions, terms, procedures - performing and algorithm, applying a formula - "identify", "recall", "recognize", "use", and "measure"
Skill/concept (level 2)	- mental processing beyond habitual response - students make decisions on how to approach problem - degree of complexity of task determines if it is a level 2 or 3 - noticing and describing patterns, explaining the purpose and use of experimental procedures, carry out experimental procedures - "classify", "organize", "estimate", "make observations", "collect and display data", "compare data"
Strategic thinking (level 3)	- reasoning, planning, using evidence at a higher level than 1 and 2 - explanations of thinking - drawing conclusions from observations, citing evidence and developing a logical argument for concepts, explaining phenomena in terms of concepts, and using concepts to solve problems - "explain"
Extended thinking (level 4)	- complex reasoning , planning, developing, and thinking (most likely over time) - make several connections, relate ideas with the content area or other content areas - select an approach among many options - developing and proving conjectures, designing and conducting experiments, making connection between and finding and related concepts, combining and synthesizing ideas

Coding Rubric - LOD

LOD Level	Description
Memorization (level 1)	- reproducing or having memorized facts, rules, formulas, and definitions - not ambiguous - cannot be completed with a procedure - no connections to concepts
Procedures without connections (level 2)	- algorithmic - little ambiguity, little cognitive demand - no connections to concepts or meaning that underlie a procedure - focused on producing a correct answer, and not about the mathematical understanding - no explanations, or the explanation is the procedure
Procedures with connections (level 3)	- procedures with a purpose of developing a deeper level of understand a mathematical concepts or idea - ambiguity in the pathways because the procedures are broad and have close connections to the underlying conceptual ideas - usually multiple representations and promotes making connections between representations - cognitive effort required
Doing mathematics (level 4)	- complex and non algorithmic thinking - explore and understand the nature of the mathematical processes, concepts or relationships - students are self-regulating and self-monitoring their cognitive process - access relevant knowledge and appropriately use them - analyze task itself and its restraints - high cognitive effort and certain level of anxiety for the student because of the unpredictability of the solution