A MODIFIED DEPTH OF KNOWLEDGE FRAMEWORK FOR WORD PROBLEMS

Jonathan D. Bostic
Bowling Green State Univ.
Bosticj@bgsu.edu

Timothy Folger
Bowling Green State Univ.
tdfolge@bgsu.edu

Kristin Koskey
Drexel Univ.
kk3436@drexel.edu

Gabriel Matney
Bowling Green State Univ.
gmatney@bgsu.edu

Toni May
Drexel Univ.
Tas365@drexel.edu

Gregory Stone
MetriKs Amérique
gregory@metriks.com

Depth-of-knowledge (DOK) is a means to communicate the cognitive demand of tasks and is often used to categorize assessment items. Webb's (2002) framework has been applied across content areas. The aim of this two-phase iterative study was to modify Webb's DOK framework for word problems. Through work with school partners, this iterative design-research based study provides supportive evidence for a modified DOK framework reflecting levels of complexity in word problems. The resulting modified DOK framework presents an opportunity for mathematics educators to reflect on various aspects of cognitive complexity.

Keywords: assessment, problem solving

Alignment between student learning standards and assessments can be evaluated based on content and cognitive complexity (AERA et al., 2014; Webb, 2007). Webb (1997; 2002) has argued for models that categorize assessment items by the cognitive demand, or depth-of-knowledge (DOK), required to successfully complete the item. Webb (2002) classified DOK into the four levels: (1) Recall, (2) Skills and concepts, (3) Strategic thinking, and (4) Extended thinking (see table 1). Kilpatrick and colleagues (2001) define a mathematical exercise as a task that promotes efficiency with a known procedure, which aligns with level one DOK. This study examined the DOK required to solve mathematical problems (i.e., not exercises). For instance, word problems may require translating from text to a mental model, and later to a mathematical model prior to applying a mathematical strategy (Verschaffel et al., 2000). A multi-step process such as this is likely to extend beyond rote procedure; suggesting such word problems are unlikely to be classified as a level one DOK.

Table 1: Webb's DOK (2002) and Associated Descriptions

Level	Description
Level 1: Recall	"The recall of information such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula" (Webb, 2002, p. 3).
Level 2: Skills and Concepts	"The engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity… These actions imply more than one step" (Webb, 2002, p. 4).
Level 3: Strategic Thinking	"Requires reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3… [Level 3 items may require] developing a logical argument for concepts; explaining phenomena in terms of concepts, and using concepts to solve problems" (Webb, 2002, p. 4).
Level 4: Extended Thinking	"Requires complex reasoning, planning, developing, and thinking most likely over an extended period of time… At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections - relate

ideas *within* the content area or *among* content areas - and have to select one approach among many alternatives on how the situation should be solved" (Webb, 2002, pp. 4-5).

Problem statement

A team of researchers developed Problem-Solving Measures for grades 3-8 (PSM 3-8) that align with the Common Core State Standards for Mathematics (CCSSM; e.g., Bostic & Sondergeld, 2015, Bostic et al., 2020). Each word problem was developed using two synergistic frameworks for problems: Schoenfeld (2011) and Verschaffel et al. (1999). These frameworks suggest that tasks are problems if (a) a task is complex enough such that a strategy is not readily apparent, (b) it is unknown whether the task has a solution, and (c) it can be solved using multiple strategies. These elements were drawn upon to design mathematics word problems aligned to the CCSSM, which ultimately became the sample space for the study. Two sample items are shared in Figure 1. It was hypothesized that items categorized as problems might be classified at a DOK higher than level one DOK. The word problems transcended level one DOK and led to a question: How might DOK better capture unique aspects of lower complexity compared to higher-complexity grades 3-8 DOK word problems? An intended outcome from this work is to develop a modified DOK framework that has potential to inform researchers and practitioners about DOK within the context of mathematics word problems and extend prior scholarship on DOK. A second outcome was to obtain sufficient rater agreement using that modified DOK framework. Two phases framed this study. Phase I entailed developing a modified DOK framework such that DOK level descriptors reflected variation in the cognitive complexity of mathematics word problems. Phase II examined rater reliability applying the modified DOK framework to the grades 3-8 PSMs.

PSM7 (7.NS.2): Rosalinda plans to make marshmallow treats to share with her class. The recipe requires $2\frac{1}{8}$ cups of marshmallows, $3\frac{1}{4}$ cups of rice cereal, and $\frac{1}{4}$ cup of butter. It serves eight people. How many cups of rice cereal will she need if she must make treats for 28 people?

PSM4 (4.OA.3): Josephine sold tickets to the fair. She collected a total of $1,302 from the tickets she sold. $630 came from the adult ticket sales. Each adult ticket costs $18. Each child ticket costs $14. How many child tickets did she sell?

Figure 1: Sample items from PSM4 and PSM7

Phase I

Development of a modified DOK framework was an iterative process with numerous feedback loops, drawing upon a design-based research approach (Middleton et al., 2008). To develop the framework, a team of five panelists were selected based on content expertise and prior experience developing PSM items. The panel consisted of two mathematics education faculty, one doctoral student, and two masters-level students. One of the two faculty member served as an external other throughout the process.

All participants read and reflected on material's related to Webb's DOK (2002). The group conducted a literature search, moved possible reading materials into a shared folder, scanned materials for content overlap and uniqueness, then reflected on materials that were essential and those that were recommended for becoming familiar with DOK. The result was three materials recommended for further review (i.e., Hess, 2006; Petit & Hess, 2006; Webb, 2002). The team read, discussed, and reflected on each reading to better understand Webb's (2002) framework

(see table 1). Once the team reached consensus in operationalizing the modified DOK language and classifying a subset of grade seven PSM items, then they applied it to a larger set of 15 grade 6 PSM items in pairs. One faculty member and one masters student formed a group while the doctoral student and a second masters student formed a second group. Memos were made during this application about limitations in using Webb's DOK framework, word problems where it was easier to classify than others, and potential modifications to Webb's DOK framework. Discussions between two team members were documented. A goal from those discussions was for each team to have agreement on a DOK score for each item. Those scores were then compared between the two pairs of team members. Disagreements were resolved through discussions and each item received a single score. Results from DOK coding were shared with the external expert as a way to check the credibility of the process as rigorous, well documented, and logically applying modifications to Webb's DOK framework.

Next in the process, the team sought areas to revise the emergent DOK framework with the intention of improving the DOK classification process for word problems. Consensus was reached that some word problems had less complexity (suggesting a lower DOK score), while other word problems required students to draw upon multiple developmentally appropriate, content-focused strategies to solve the task (suggesting a higher DOK level). The team began the process of creating a modified DOK framework by revising Webb's DOK (2002) framework. Discussions of memos from earlier in the process generated big ideas, which informed DOK modifications. The team developed a draft revised DOK framework for word problems. The four team members sought feedback from the external other, which was integrated into the revision. The team tested the draft revised DOK framework with four randomly selected items from a pool of grades 6-8 items. Each team member worked individually and memoed about their rationale for each rating. Then, they met with their team partner for agreement and discussion of memos, followed by a whole-group meeting. Revisions were made to the draft modified DOK framework based upon emergent patterns from memos. The team shared DOK scores, memos, and patterns with the external other as part of the process. After further rounds of revisions in this fashion, the team developed a revised DOK framework specifically for mathematics word problems, which was again shared with the external other. The team applied this revised DOK framework to a set of 10 randomly sampled grades 3-8 word problems that were new. The degree of inter-rater reliability agreement (κ) was interpreted using Cohen's guidelines (1960): x< 0.20 (none to slight), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (substantial), and x> 0.81 (almost perfect). There was agreement across nine items and disagreement with one item generating Cohen's $\kappa = .9$. On that one item, pairs of coders had agreement, but the codes differed by one adjacent unit. Discussion ensued and the team ultimately reached agreement. Thus, there was almost perfect agreement during the final iteration across four team members. At this time, the team developed training materials that for future coders and/or re-calibration. Developed materials included a manual, sample items with explanations of coding, as well as a set of items to be used for calibration coding. The modified DOK framework is shared in Table 2.

Table 2: Modified DOK Framework for Word Problems

Level	Description
Level 1: Recall & Reproduction	Respondents are expected to recall or reproduce knowledge and/or skills while problem solving. This typically involves working with facts, terms, details, calculations, and/or simple procedures or formulas. A solution to a level 1 item does not need to be "figured out" or "solved".

Level 2: Skills & Foundational Concepts	A student is expected to decontextualize from a situational context when solving items. A level 2 DOK item requires students to process information regarding the concept being measured.
Levels 3: Strategic Thinking & Reasoning	A level 3 DOK item requires mathematical reasoning and higher order thinking processes. Students are expected to combine multiple skills or heuristics and conceptual understanding to reach a solution. There may be multiple viable solutions.

Phase II

The modified DOK framework for mathematics word problems was then used by three raters (mathematics education graduate students) not familiar with the development process. The training manual and selected readings (e.g., Hess, 2006; Webb, 2002) were shared with them in advance of their training session. A 90-minute training with the faculty member from phase one included discussing readings and sample PSM items. Following coder training and calibration, the raters were asked to work collaboratively on three items and come to consensus using the revised DOK framework. One week later, the faculty member and three raters debriefed about how they reached decisions with the framework, nuances related to each word problem, and concerns they had while applying codes. Following this debriefing, the raters engaged in independent coding. The raters were asked not to disclose ideas with each other related to their independent codes. They received six-word problems and a copy of the revised DOK framework. All three raters agreed on four items, yet there was a discrepancy on two items by one adjacent level (i.e., DOK 2 vs DOK 3). This indicated substantial agreement using Cohen's (1960) guidelines (Cohen's $\kappa = .71$); however, the team wanted to have a stronger understanding of the revised DOK framework. The raters convened with the faculty member to discuss coding results and their memos from applying the modified framework. Emerging from their use of the framework were challenges understanding developmental differences of content knowledge at a specific grade level (e.g., grade 7 vs grade 8) and what counted as two distinct strategies. After coming to agreement and finalizing discussions with the lead faculty member and external other, this team completed independent coding as a confirmation of their ability to reliably code items. They were provided with seven randomly selected grades 3-8 PSM items and asked to (a) code each item using the modified DOK framework and (b) provide a rationale for their code. Results from their independent coding showed that the three raters applied the same codes for five of seven items. One rater differed from the other two raters on two of seven items by an adjacent code. Rater agreement (Cohen's $\kappa = .89$) yielded almost perfect agreement. Coding memos across the three raters communicated further support for differentiating between levels two and three was warranted to help improve agreement in classifying DOK level.

Discussion

The present work serves as a means to reconsider applications of Webb's (2002) DOK framework for word problems. Through a focused, design-research process, we developed a modified DOK framework that helps to magnify potential differences between word problems. Thus, this modified framework extends prior DOK scholarship. Much like a magnifying glass can highlight unique features not seen with the naked eye, we believe this framework may do the same for word problems: highlighting the unique features of word problems. A second intention in this design was to test the potential of others using this framework. An initial use with a small sample demonstrates potential for effectiveness (i.e., scaling up) trials this revised DOK framework with a larger sample of coders and other items. One limitation with this framework is that the word problems reviewed did not lend themselves towards Webb's (2002) level four

DOK. Thus, our revised DOK framework did not reveal changes in this area. However, we believe reviewing tasks that promote modeling with mathematics (CCSSI, 2010) may lend themselves to this level and result in revisions to the framework. Another limitation is that the team has only reviewed word problems associated with a project that designed word problems. A future study might explore use of the revised DOK framework with samples from textbooks used in classroom instruction.

References

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

Bostic, J., & Sondergeld, T. (2015). Measuring sixth-grade students' problem solving: Validating an instrument addressing the mathematics Common Core. *School Science and Mathematics Journal, 115, 281-291*.

Bostic, J., Matney, G., Sondergeld, T., & Stone, G. (2020, March). *Measuring what we intend: A validation argument for the grade 5 problem-solving measure (PSM5). Validation: A Burgeoning Methodology for Mathematics Education Scholarship*. In J. Cribbs & H. Marchionda (Eds.), Proceedings of the 47th Annual Meeting of the Research Council on Mathematics Learning (pp. 59-66). Las Vegas, NV.

Common Core State Standards Initiative. (2010). Common core standards for mathematics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46.

Hess, K. (2006). Exploring cognitive demand in instruction and assessment. National Center for the Improvement of Educational Assessment, Dover NH.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). Adding it up: Helping children learn mathematics. Washington, DC: National Academy Press.

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

Petit, M., & Hess, K. (2006). Applying Webb's depth of knowledge and NAEP levels of complexity in mathematics.

Schoenfeld, A. H. (2011). How we think: A theory of goal-oriented decision making and its education applications. New York, NY: Routledge. https://doi.org/10.4324/9780203843000

Verschaffel, L., De Corte, E., Lasure, S., Van Vaerenbergh, G., Bogaerts, H., & Ratinckx, E. (1999). Learning to solve mathematical application problems: A design experiment with fifth graders. Mathematical Thinking and Learning, 1(3), 195-229. https://doi.org/10.1207/s15327833mtl0103_2

Verschaffel, L., Greer, B., & De Corte, E. (2000). Making sense of word problems. Lisse, Netherlands.

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: University of Wisconsin, Wisconsin Center for Education Research.

Webb, N. L. (2002, March). Depth of knowledge levels for four content areas. Wisconsin Center for Education Research, University of Wisconsin-Madison, Madison, WI.

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