RTD Approach to Using Norman Webb's Depth of Knowledge (DOK) Typology of Cognitive Complexity

By: Marjorie Wine
Alexander Hoffman

ALEDEV CONSULTING

The RTD (Rigorous Test Development) project is an attempt to build a professionalized content development practice that focuses on individual item quality, particularly by leaning into the importance of validity throughout the content development process. It assumes that content development professionals develop professional judgment that can be raised, honed and calibrated by providing frameworks and clarifying expectations in ways that account for the constraints and demands of typical practice within test development, today. RTD is a conscious and deliberate attempt to respond to the disparity in status, training and shared knowledgebases between psychometrically oriented professionals and content development professionals.

Table of Contents

ALEDEV CONSULTING

This approach to classifying the cognitive complexity of standards and of items is grounded in three different constructs.

First, and most importantly, it is grounded in Norman Webb's (2002) work on cognitive complexity, *Depth of Knowledge*. This is the most commonly used typology for cognitive complexity in the assessment field. This approach attempts to take Webb's work very seriously, and to recognize the criticisms of assessments that DOK was designed highlight.

Second, this approach also grounded in *Rigorous Test Development* (Wine & Hoffman, 20XX). The heart of RTD can be found in the central tenet *valid items elicit evidence of the targeted cognition for the range of typical test takers*. RTD is clear that all forms of alignment and meaningful evaluation of items must focus of the *cognition of test takers*, rather than merely the wording of items (Wine & Hoffman, 2020). That is, items are designed to prompt and measure test taker cognition. Because it is the cognition that is being reported upon, item performance should be evaluated by considering the cognition that items actually prompt.

Third, this approach takes seriously the idea that pervades cognitive psychology that learning and increasing proficiency qualitatively changes how information is processed (Ericsson, 2014; Moors & De Houwer, 2006; Salomon and Perkins, 1989). Learning and experience increase automaticity, reduce the conscious attention required to respond to an item and thereby reduce the cognitive load. (See appendix for a fuller discussion.)

This approach recognizes that – because much of what becoming proficient with a skill is reducing the need to consciously deliberate through the process of applying the skill – less proficient students are engaged in *more* cognitively demanding and complex cognition when they engage with a task than more proficient test takers. In other words, the top *performing* test takers are not necessarily engaging in cognition with the highest DOK levels when they engage with assessments. The relationship between test taker performance and cognitive complexity is quite complex.

DOK Basics

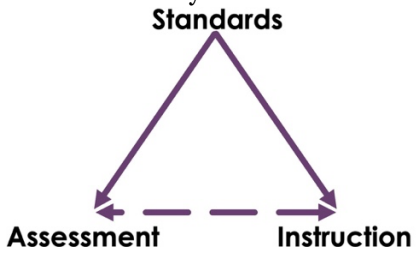
Webb's Depth of Knowledge typology has four levels. These levels can be used to describe standards, classroom activities and assessments, and standardized assessments

Level	Name	Description
DOK 1	Recall	Recitation or recognition of facts, basic reading comprehension, rote use of algorithms or procedures. Includes recitation or identification of explanations learned previously.

Level	Name	Description
DOK 2	Skill/ Concept	Some degree of inference and analysis, basic decision making, performance of work *without* strategic planning, selection of the correct simple tool or procedure and its application.
DOK 3	Strategic Thinking	Explanation of decisions, thinking process and/or work performed. Strategic planning or the application of multi-part reasoning to determine a course of action. Citing evidence to support reasoning.
DOK 4	Extended Thinking	Thinking that is extended across multiple contexts or concerns in ways that connect those contexts or concerns. Arriving at generalizations based upon a range of information or ideas. Analysis that includes multiple factors or issues and account for those issues in the final product.

Webb's development of his DOK typology and its subsequent widespread use are both grounded in an agenda: *education should be about more than just memorization, recall and rote application of simple skills* (Barber, 2018). Every system for recognizing cognitive complexity presupposes that cognitively simple tasks are an important part of education and that cognitively more complex tasks are vitally important as well. This agenda lies at the heart of the next generation standards for math and literacy (The Common Core State Standards) and for science (Next Generation Science Standards) that followed Webb's original work on DOK by a decade or more.

Therefore, it incredibly important to be diligent about applying the same understandings of DOK to different parts of the assessment triangle. Though there are significant differences between estimating the DOK levels of the cognition described in state learning standards and the DOK levels required in students tasks – be they in classroom activities and assessments or part of standardized assessments – fulfilling the purpose of DOK requires careful consistency between the two.



Common DOK Misconceptions

There are a number of common misconceptions about DOK and cognitive complexity.

First, DOK is *not* a synonym for difficulty. In the context of standardized assessment, difficulty is empirically derived by examining the share of test takers who response to an item successfully and DOK is *not* derived empirically. Even when viewed more philosophically, there is a difference between complexity and difficulty. Tasks can be difficult without being complex. For example, memorizing vocabulary words for later recitation of definitions can be difficult even though it is cognitively simple, whereas determining an unknown word's meaning from context clues is more cognitively complex even though it may be easier. Yes, difficulty often *does* tend to increase with cognitive complexity, but it is not *necessarily* the case. Furthermore, more difficult tasks are not always more cognitively complex than easy tasks.

Second, cognitive complexity is *not* determined by the number of steps in a task. More cognitively complex tasks often *do* have many steps, but a large cognitively simple task may instead be composed of mere repetition of a simple task. More of the same does not make for a qualitatively different cognition, even though it may increase difficulty. Moreover, the number of steps in a task is also function of how the skills are stored – a function of automaticity.

Similarly, the scale of scope of a task does *not* determine cognitive complexity. Longer writing is not necessarily more cognitive complex. Arithmetic with more digits is not more complex than arithmetic with fewer digits; once you get beyond two- or three-digit numbers and regrouping, it is just more of the same. A larger canvas may have space for more complex work, but it does not require it and a lot of complexity can fit on rather small canvas.

Fourth, tasks that take more time are *not* necessarily more complex. Again, more time may allow for greater complexity, but a repetitive task may take a great deal of time without being cognitively burdensome.

Fifth, cognitive complexity is *not* determined by the context in which the skill is applied. Spelling words correctly is important, but it is cognitively simple regardless of whether it is done in the context of a single sentence, a single paragraph, a short essay or a long novel. It is the nature of the application of the skill in question that determines DOK level, and not its context.

Last, one cannot determine the cognitive complexity of a standard or a task merely by examining verbs. There is a particular "DOK Wheel" that lists 9-23 verbs for each DOK level. Despite the fact that this document credits Norman Webb and is even distributed by some state departments of education, Webb himself has specifically disclaimed it as a distortion of the concept of cognitive complexity (Walkup, 2014), even in the most vociferous terms, "Although it references my work and uses DOK, the wheel itself misrepresents my work" (personal communication, October 18, 2011). Among other reasons, a single verb can suggest different cognition in different contexts (e.g., *define* a word vs. *define* an interpersonal relationship). It is the nature of the cognition that is in question, and not a question of what words have prompted or the main verb used to describe it.

Extending DOK

Webb's DOK is very robust, and this approach to thinking about classifying items and standards by cognitive complexity is merely a lateral extension of it. The addition of the RTD lens and recognition of the importance of automaticity draws attention to certain ideas that DOK has already long included and/or implied.

The first, as stated above, is that cognitive complexity is a feature of a cognitive path, rather than a feature of an item. This does not mean that items cannot be classified for cognitive complexity – after all, that is one of the two goals of this approach. Rather, it means that doing so requires considering the cognitive paths that test takers might follow when responding to the item. Webb is clearly addressing cognition with his typology. RTD emphasizes this fact, that it *must* be about the prompted cognition.

Second, also as stated above, RTD is built upon the idea that test takers vary in how they respond to items (Wine & Hoffman, 2020). That is, they may have different reactions (e.g., consider sensitivity issues), different prior experience (e.g., consider bias issues) and even different solution paths. This test taker variation – especially their different solution paths – can result in different levels of cognitive complexity in response to the same item.

Third, cognitive complexity is always about the process – the cognitive path – and is never about the result. A complex result does not necessary require a complex process and a complex process can result in a remarkably simple result. One must not be distracted the by complexity of a final product when considering the complexity of the cognition that led to it. In fact, a complex process may lead to an incorrect result or even no result at all. Consider the old expression *paralysis by analysis.* An enormously complex process and end up being entirely ineffectual. Its lack of efficacy does not impact its cognitive complexity because cognitive complexity is not a moral judgement of worth or merit.

Fourth, not only is there variation between different test takers, but there is also variation of a single test taker over time (see Appendix). Processes that may require deliberate planning and care when first learned become automatic and unconscious with the kind of experience that leads to mastery. Webb writes that "rote response[s]", use of "well-known algorithms," "set procedures" and "clearly defined" processes are DOK 1 cognition, but these are not so much traits of the skills themselves as *they are products of proficiency with those skills* (see Appendix). Watching a true master of any skill is most humbling when one realizes how little attention they put into producing results that would be difficult and attention-consuming labor for oneself. That is, it is not necessarily the ceiling on what they can do that is most impressive, but rather how much they can do with distracted ease.

Fifth, there is a particular challenge with classifying *items*. Because cognitive complexity is a feature of cognitive paths and not of items, the highest goal of CPDs (content development professionals) is to write items that elicit unambiguous evidence – be it of the contents of the standard or of cognitive complexity. Inferences made from test results should be based on the strongest possible evidence, and not on mindreading or

wishful thinking. Therefore, when test takers can successfully respond to an items via paths of a range of cognitive complexity, CDPs should acknowledge that reality.

DOK Ceilings: Classroom Activities vs. On-Demand Standardized Assessments

There is no question that classroom activities (and homework) differ in many ways from the tasks available for large scale, on-demand, standards-based standardized assessment (i.e., standardized assessment). Classroom and homework activities often allow for more time for a single extended task – be it ten minutes or two weeks --, whereas standardized assessment rarely is able to do that. Differences in available scoring mechanisms (i.e., attentive teacher evaluation vs. objective/consistent standardized scoring) can also influence what kinds of tasks are available for use. These issues are not intrinsic to standardized assessment, but rather are artifacts of the resources (e.g., time, money) that are available to standardized assessments.

Thus, decisions made about the resources dedicated to standardized assessment (among other issues) determine the range of cognitive complexity that a standardized assessment may include. The Advanced Placement exams, with their time for student writing and/or showing/explaining their work, demonstrate that standardized assessment *can* address DOK 3 quite well. However, scoring, testing time and other limitations can also prevent the higher DOK levels from being included.

Furthermore, classroom teachers – with the range of types of evidence and information they can gather on students – can be more *sure* of the complexity of a particular student's cognition than test developers can. There simply is a ceiling on how confident anyone can be of the complexity of the unseen cognition that leads to a particular product, especially when a test takers may be particularly well prepared for a type of task. On-demand standardized tests often can only report a *range* of cognitive complexity for the cognition that an item prompted, due to the nature of the standard, the limitations of the assessment resources and variation in test takers' preparation.

Understanding DOK: Layered Example #1

The DOK typology for cognitive complexity can be used for any task, not just those that appear on standardized assessment. It can be used for activities performed by anyone, in any context. Even by adults. Even by *teachers*. Consider the following common task: a teacher asks for a show of hands of their students in order to determine whether to move on to the next step of the lesson. This example is full of different knowledge, skills and abilities (KSAs) that operate at different levels of cognitive complexity. The DOK of each must be recognized at the level of the relevant question.

- Regardless of the complexity of the greater context, counting the hands remains a DOK 1 application. It is a rote exercise, even if it includes repeating the question and reminding students to keep their hands up. Calculating how much time remains in the period is similarly a DOK 1 task.

- Evaluating which raised hands to take seriously is likely a DOK 2 exercise. It involved applying prior knowledge of each student, but not in a rote fashion. It likely requires combining multiple kinds of information in the moment, and so is not a rote application of skill. However, this quick sort of analysis of each student/raised hand is a quick determination when done by a proficient teacher.
- The decision of whether to move forward is a strategic decision at DOK 3. That is, the teacher is considering tradeoffs and plans, weighing goals and possibilities. It is done very quickly, but the weighing, planning and consideration is strategic in nature. It is strategic because it is prospectively reflective about the best use of time, looking forward.

This example shows that some KSAs simply remain at lower DOK levels, even when they are applied in in the context of a more complex task. Changing the context of the application without changing the cognition of the application does not alter the DOK level. However, if one *does* change the cognition, the DOK level can change.

DOK 4 is *much* less common. If the teacher realizes that this approach to balancing time and instruction is not working, they might embark on a DOK 4 project. They might experiment with different way to weigh the raised hands in a class. Perhaps trying to ignore certain types of students to focus on others to see if that works better. Perhaps writing down the number of raised hands to make sure they actually stop and count instead of going by an impression of the hands they see. Perhaps insisting that everyone be quiet during counting. This multiweek effort to find an implementation of this approach that works for them and their students incorporates all of the above tasks at all of their DOK levels, and taken together constitutes a DOK 4 task – even though many of the applied KSAs remain at lower DOK levels.

This example points to the great challenge of cognitive complexity and assessment – be it standardized assessment or classroom assessment. *Was that really DOK 3, or was it just DOK 2?* The fact that the teacher made a decision about whether to move forward or stay on that part of the lesson does not tell the external observer whether they engaged in the conscious, deliberate and strategic decision-making of DOK 3. They could have just looked for a certain number of hands by rote or even an impression of a bunch of hands. The external observer would need *evidence* of the thinking process, as students are asked in include in their essays, when they explain their work and when they write up their results.

Cognitive complexity is always about the process – the cognitive path – and is never about the result. The teacher could make the right decision, the wrong decision or even be frozen in uncertainty from a DOK 3 process. A similar range of results could follow from processes at other DOK levels, as well. In order for *items* to rise to the level of DOK 3, they must provide evidence that test takers utilized a DOK 3 process.

In fact, a requirement on test takers to provide such evidence can raise the DOK level appropriately for the most proficient test takers. One can imagine a masterful teachers who uses this approach in their classroom for a variety of reasons. They might

read the room so quickly and automatically that *for them* this is a DOK 2 task, as it does not require the kind of conscious deliberation and strategic thinking of a DOK 3 task. However, when they later explain to their student teacher how they evaluated the class and made their decision – becoming more conscious of their reasoning as they unpack it for their student teacher – they move into DOK 3.

Classifying the DOK Level(s) of a Standard

Classifying the DOK level(s) of a standard begins with a *close reading* of the standard. This means that the plain language of the standard is taken seriously. Inconvenient words or phrases may not be ignored and new words may not be inserted to alter its meaning. Close reading begins with trying to understand what the precise language of the standard describes, on the surface.

DOK is about the complexity of cognition, rather than the complexity of an external task, process or result. Like every other typology for cognitive complexity, it only focuses on particular facets of this broader construct. Determining the DOK level of a standard requires thinking through the cognitive step that a person would have to take to do what the standard describes. We recognize that standards may be written in a way to describes a process that students should be able to use, a product that they should be able to create or even directly describes thing that should be *understood*, *recognized* or *appreciated*. (We refer to these as *process standards, product standards* and *cognition standards,* respectively.) Regardless of the type of standard, classifying its DOK level(s) requires unpacking the *cognition* required to do what the standard describes more deeply than the standard itself does.

Once the cognitive path is unpacked, it is then examined though the particular lens of DOK. Like all cognitive complexity typologies, DOK highlights particular dimensions of complexity and ignores others. DOK levels are determined by focusing on those dimensions, as explained below.

Note that a single standard may include elements at a variety of DOK levels. Therefore, every DOK level should be considered when classifying a standard.

Step 1: DOK 4 (Extended Thinking)

The easiest DOK level to recognize is DOK 4 because of the rarity of standards that focus on the extended thinking that is at the heart of this level. There are multiple CCSS-L standards that explicitly call for this kind of work and none in CSSS-M. Many of NGSS's standards include at least some DOK 4 work.

As DOK 4 is so easy to recognize, it is the easiest level to rule out.

Step 2: DOK 3 (Strategic Thinking)

While DOK 3 is not quite as easy to recognize or eliminate as DOK 4, there are two telltale signs that indicate that a standard includes DOK 3 components. The first is that it calls for an *explanation of reasoning* as part of the final work product. The second is that calls for some amount of conscious and deliberate *planning of future work*.

Both of these signs are types of metacognition (i.e., thinking about thinking). In one, the student is expected to reflect prospectively about the work to come and how to accomplish it. In the other, the student is expected to reflect retrospectively about the work that *has been* done and to explain it.

Note that not all explanations meet this requirement. Repeating an answer and the explanation of the answer that students *were taught to repeat* is a form of recall (i.e., DOK 1) and not of DOK 3. DOK 3 calls on students to explain *their own* thinking, not merely to repeat someone else's explanation of a phenomenon, event or idea.

Step 3: Easy-to-Recognize DOK 1 (Recall)

Many standards include elements that are quite easily recognized and classified as DOK 1 because their cognition constitutes a particular type of memorization/recall or rote application.

- Reading standards that rely on recalling or recognizing what is explicitly stated in a text
- Rote application of regularized procedures or rules in any content area.

Any skill that applied with such automaticity is operating at the DOK 1 level. Standards that call on more or less automatic responses address the DOK 1 level. These skills can be *incredibly* important – even foundational to a content area. The fact that these standards include DOK 1 components does *not* make them unimportant.

Step 4: Differentiating DOK 2 Standards

Unquestionably, the most difficult part of classifying standards by DOK level is this final step. It can be challenging to recognize whether cognition is DOK 1 or DOK 2, or whether it is DOK 2 or DOK 3.

DOK 2 is about applying skills, and is *not* about recall or rote application. The line between rote application and true skill application can be hard to recognize. In fact – and don't tell anyone – one sign of true mastery with a skill can be when it turns into an automatic or rote application (DOK 1) instead of the conscious and deliberate decision-making of DOK 2. This can be seen in the difference between having to *figure out* whether a word is spelled correctly versus *quickly recognizing* that a word is misspelled. Though both are types of vocabulary skills, knowing what a word means is at DOK 1, but figuring out what a word means from context clues is at DOK 2. Understanding what is stated explicitly in a text is DOK 1 cognition and making inferences from a text is DOK 2 cognition – but *how* explicit or *how much* inference is needed for it to really be inference can be a difficult question sometimes.

That conscious decision making and/or conscious application of a skill or understanding/use of concept is at the heart of DOK 2 cognition. But it becomes DOK 3 cognition when the conscious and deliberate cognition is about what skills or path to follow later (i.e., "strategic thinking"). It becomes DOK 3 when the decision(s) have to be justified or explained (see above).

Managing the Standards DOK Classification Process

There is no doubt that most the work – both in terms of effort and in term of time – comes in Step 4. Once the easy classifications have been recognized, differentiating DOK 2 from the levels above and below it is the difficult task.

DOK classification requires consistency in order to arrive at a defensible final product. Because few organizations will expect the entire project to be completed by a single person, it calls on up-front efforts for the team to calibrate to a common approach for resolving questions about those DOK 1/DOK 2 and DOK 2/DOK 3 lines. It is not enough simply to declare that one should lean towards the lower levels or the upper levels or even try to include both. This is because different people will have different ideas about whether a standards might straddle (or even be close to) one of the lines.

Therefore, DOK classification project should include intentional construction of examples from each relevant content area and generalizable explanations for how the classifications decisions are to be made *for this project.* The basic distinctions explained above (i.e., in the DOK Basics table and in the steps of DOK classification) are *not* enough to provide the consistency that is needed. DOK ambiguities appear differently in the different content area and even across different grade bands. If team members cannot refer to shared reference documents, decisions around those DOK 2 lines can easily be so inconsistent the project loses both credibility and usefulness.

There are additional complexities in classifying *items* by DOK (see below) that do not apply to classifying standards. This is because standards describe levels of proficiency and/or mastery that test takers may not have (yet) achieved. That variable impact of proficiency on cognitive complexity that item classification must address is not an issue for classification of standards.

Understanding DOK: Layered Example #2

DOK classification is another layered task that can viewed through the lens of DOK.

DOK 1. Much of Step 1 and 2 are DOK 1 tasks. It is very simple straightforward recognition of when a standard calls for an extended term layered project and when a standards calls for rote application or recall. These classifications usually result from a surface reading of what is explicit in the standards. Similarly, classifications of standards that explicitly call for explanations or planning is often a DOK 1 task for members of a classification team.

DOK 2. Decisions around those difficult lines are DOK 2 tasks. They require conscious application in moment of judgment to make decisions. However, they do not call on planning or other metacognition. Classification itself, regardless of how difficult the decision, is going to be DOK 1 or DOK 2.

DOK 3. Developing the guidelines and examples that a team would use to calibrate together are DOK 3 tasks. They requires examining the thinking that goes into such decisions and explaining it clearly. Explaining reasoning is DOK 3. Planning a structure for those documents is also a DOK 3 task.

DOK 4. Planning and managing the classification project is likely a DOK 4 task. There are elements of time and personnel management, obviously. There is also extending thinking about how inclusive consideration should be and other issues about the overall strategy for addressing decisions around those various classification lines that need to take into account the various audiences of and intended uses for the final product/report.

The cognition requires for different aspects of a classification project exist at different DOK levels. Those taking part in the project may stay with DOK 1 and DOK 2 cognition because repetition of the same tasks/cognition – even for an extended period of time – do not raise the cognitive complexity of the work. Their calibration work, however, would likely be DOK 3 cognition because of the need to interrogate – and even explain – their thinking as the team comes to the shared understanding of the difficult classification lines. The leaders and planners of the project, however, will likely additionally be engaged in DOK level 4 cognition.

Classifying the DOK Level(s) of Test Items

Classifying test *items* by DOK level is quite different than classifying test standards by DOK level. As shown below, they are parallel processes with many similarities, but they differ at a fundamental level.

Standards should be read and understood as originally intended, leaning on the plain and clear meaning. Items cannot be read the same way. Standards generally attempt to describe learning or curricular goals – essentially cognition. Standards are, therefore, just one step removed from actual cognition. Items, on the other hand, aim to elicit *evidence* of particular cognition (Wine & Hoffman, 2020). That extra step – which is mediated through the understanding and performance of test takers – means that the intention of the of an item's authors is simply not relevant to the determination of DOK level.

This *evidentiary* lens is absolutely critical to understanding how items function, what they measure and how they should be understood by CDPs. The questions of DOK level classification *of items* is really *What level of complexity (in DOK terms) of test takers' cognition do I have evidence of?* It requires careful consideration of test takers' potential cognitive paths through items – the paths they followed to get from the beginning of the item to the end of their response.

The common misunderstanding that DOK level is merely about context in which a skill is exhibited is easily dismissed by considering the existence of items within a larger test. Test takers will virtually always engage in some amount of strategizing or time management when taking a test. They may decide to skip items that appear too difficult for them, to give up on items that are taking too long, to come back later to check their work versus double checking it immediately. All of these are strategic decisions that make taking a test a DOK 3 task. But that context does *not* make every item DOK 3 nor each skill applied at DOK level 3. Similarly, virtually every item will call on some DOK 1 skills, from simple reading or arithmetic, or even whatever is required to indicate one's answer. But

this testing context does *not* mean that all items should classified as hitting DOK levels 1 and 3.

In fact, the DOK of an item is a matter of the cognitive complexity *of the application of the targeted cognition.* That is (regardless of the complexity of the entire task), *what is the DOK level of the parts of the cognitive path that the aligned standard describes?* When evaluating items for DOK classification, we generally ignore basic reading[1] and time management components of responding to an item. We also set aside other aspects of test takers' cognition so that we can focus on the cognitive of the application of the standard in the task.

This process is further problematized by the realities of what test items actually look like and the item types that are available. For example, there is a difference between asking a student to volunteer an answer that they have constructed and asking them to select an answer from those that have they have been given as possibilities. The skills applied might be quite different. For example, by plugging in each of the offered answer options (i.e., backsolving), a test taker can turn a subtraction problem into an addition problem. A test taker can otherwise try out the various offered answer options to see which one works best, instead of being responsible for developing an answer themself. This shortcut can reduce the cognitive complexity of some tasks, as compared to constructed response formats.

All of this means that when items are examined for DOK classification, the following guidelines should be followed at every step.

- Consider the cognition prompted by the item, and not the item itself.
- Consider the cognitive path prompted by the *whole* item, and not just what would be prompted by the stem.
- Focus on the part of the cognitive path that is part of aligned standard, and disregard other aspects of the task.
- Be aware of the test taker profile(s) whose cognitive complexity you are classifying (see below).

Step 1: DOK 4 (Extended Thinking)

This step is even easier when classifying items than when classifying standards. On-demand standardized assessment simply does not provide enough time for true DOK 4 cognition. Such assessments quite rarely (as of 2022) allow even a single hour for a single task. The kind of extended projects that embody DOK level 4 are simply not a part of today's on-demand standardized assessments.

Step 2: DOK 3 (Strategic Thinking)

Even DOK level 3 cognition is difficult to achieve in today's on-demand standardized assessments. If items require test takers to provide reasoning and

[1] When the aligned standard is actually about some elements of basic reading skills, we *do* focus on that part of test takers' cognitive paths.

justification for their work – as in many constructed response – it may be DOK 3. Only items aligned to DOK 3 *standards* may be DOK 3 items. Items that really do require test takers to offer conscious explanations and justifications for their thinking process are certainly DOK level 3.

One difficult aspect of determining whether an item prompts level 3 cognition is that the strategic planning that that can call for a DOK 3 classification is often not directly visible in a final work product. A struggling writer may poorly plan an essay and then poorly execute on that plan, but nonetheless that planning process brings it to DOK 3. On the other hand, a more proficient and/or experienced writer might simply write their response off the top of their head without ever stopping to plan their essay. Though the quality of the that latter essay may be higher, the quality of the response does not make it DOK 3 cognition in the absence of *planning*.

The second difficult aspect of determining whether an item prompts level 3 cognition lies in the difference between developing one's own justification/explanation for an answer and merely identifying the correct reason from a provided list. Recognizing and *selecting* a reason for something is different than *constructing* and offering a reason oneself. Identifying a reason is different than explaining own's on thought process.

Last, if an item requires test takers to use the aligned standard to explain how they know something or why they took the approach they did, it should be classified as DOK level 3. Merely showing the work they did to get there is *not* DOK 3 – and may just be DOK 1 – when it is neither reflective prospectively nor retrospectively.

Step 3: Easy-to-Recognize DOK 1 (Recall)

Items that require to test takers to apply the targeted cognition in a rote fashion are classified as DOK 1. This includes most basic skills, regardless of the content area. Automatic and unconscious application of more advanced skills are also DOK 1. Basic reading and recall of explicit details and ideas constitute DOK 1 at most grade levels. Basic math skills, whether applied unconsciously (e.g., memorized multiplication tables) or by rote (e.g., memorized procedures like using the Pythagorean theorem) are DOK 1. Offering or recognizing basic statement of scientific facts are usually DOK 1, as well.

Step 4: Differentiating DOK 2 Standards

Unquestionably, the most difficult part of classifying item by DOK level is this final step. It can be challenging to recognize whether an item prompts cognition at DOK 1 or DOK 2, or whether it is DOK 2 or DOK 3.

DOK 2 is about deliberate application of skills, and is *not* about recall or rote application. The line between rote application and true skill application can be hard to recognize. In fact (as indicated above), one sign of true mastery with a skill can be when it turns into an automatic or rote application (DOK 1) instead of the conscious and deliberate decision-making of DOK 2. This can be seen in the difference of having to figure out whether a word is spelled correct vs. quickly recognizing that a word is misspelled.

When considering whether an item prompts DOK 1 or DOK 2 cognition, one must consider how *test takers* would work through the item and *not* how proficient and experienced adults would work through the item. The savvy and shortcuts of adults might not be available to test takers.

In fact, because proficiency can lower the cognitive complexity of a cognitive tasks (i.e., by raising automaticity), the DOK level can be lower for those who answer the item successfully than for those who are unsuccessful with item. A classification project will have to decide up front whether items can have DOK ranges to account for this or whether another approach is more desired.

Similarly, the line between the conscious application of skill (DOK 2) and conscious and deliberate strategizing of how to combine skills (DOK 3) can be influenced by test takers' proficiency with the skills in question. The uncertainty that can wisely accompany lesser proficiency can prompt cognition at a higher DOK level. Again, greater proficiency may lead to lower cognitive demand and complexity – even for the exact same task.

Managing the Item DOK Classification Process

As with classifying the cognitive complexity of standards, most of the work – both in terms of effort and in term of time – comes in Step 4. Once the easy classifications have been recognized, differentiating DOK 2 cognition from the levels above and below it is the difficult task.

As with classifying standards, classifying items requires a great deal of calibration work (see above). Also like classifying standards, there are important decisions to be made about where lines might be drawn and what to do when near or straddling a line (see above). Unlike standards, items are generally thought to have a single DOK level. However, if the project intends to classify the *cognitive* complexity of the items then it must consider the cognition that the items each prompt. Because test takers vary in their proficiency and because of the inevitable impact of proficiency on cognitive load and cognitive complexity, it simply is a mistake to think that items, generally, each prompt cognition at a fixed single DOK level.

Therefore, any item classification project must decide among a number of options.

- Assume cognition performed proficiently, as described in the standards. This will tend to suggest lower DOK levels.
- Assume cognition of a test taker who is not yet at the proficiency described in the standards, but can still address the item. This will tend suggest higher DOK levels.
- Recognize that items prompt cognition at range of DOK levels. (RTD's refrain that *valid item elicit evidence of the targeted cognition for the range of typical test takers* certainly suggests this path.)

Unfortunately, it simply is not practical to try to assume a "typical" or "average" test taker as such approaches fail to consider *most* test takers. More importantly, they simply beg the question of what level of proficiency the "average" or "typical" test taker possesses and how *that* proficiency is most likely to impact cognitive complexity. This is further complicated

by *quite* problematic questions about what sort(s) of students to consider as "typical" or where an "average" student may be drawn from. That is, it makes virtually every item *more* difficult to classify, perhaps prohibitively so.

Once one recognizes the fact that there's a range of DOK in student cognition, trying to aim for "typical" or "average" just begs the question, even on an individual item level. Such an effort actually makes DOK classification maximally difficult for each item because the effort to figure out a) what a "typical" proficiency with this standard means, cognitively and b) what the impact of that cognitive path is on cognitive complexity becomes absolutely prohibitive -- if even possible.

The Place of Cognitive Complexity Classification in Assessment

Cognitive complexity is a powerful and important idea. Norman Webb's original agenda to highlight particular shortcomings in the assessments that predated even NCLB remains worthy to this day. There is no doubt that *Depth of Knowledge*, in particular, has proven robust and flexible for the standards movement and the assessment industry.

Unfortunately, DOK has often been applied inconsistently from project to project or when classifying standards versus classifying assessment items. At worst, it has been treated as a necessarily-but-ultimately-meaningless step in the assessment development process. And there can be no question that there are forces and decision-makers that end up exacerbating the problematic issues that DOK was designed to highlight. This creates a quite natural resistance to embracing sincere use of DOK even among those are most invested in created absolutely the highest quality assessments.

On the other hand, so many dedicated assessment professionals seek to use DOK (and/or other typologies of cognitive complexity) in their work to help them to develop higher quality assessment that better support meaningful inferences and test uses. It is our belief that this *is* possible, especially when DOK is used to classify test taker cognition – as it clearly was originally intended to do.

Appendix: Grounding in the Literature

The needs for a lateral extension of this typology for cognitive complexity (Bechard, Karvonen, & Erickson, 2021) is grounded in one of the common themes in cognitive psychology – that increases in proficiency do not merely yield improved measurable performance but actually produce qualitative changes in both how information is processed and in the nature of cognition. Depth of Knowledge (Webb, 2002) particularly requires this lateral extension because DOK's levels are often defined in terms of cognition, rather than then descriptions of external tasks. For example, Webb describes DOK Level 1 as "rote response[s]" and "well-known algorithms." Webb names DOK's higher levels by describing cognition (i.e., "Strategic Thinking" and "Extended Thinking"), rather than traits of the external problem, challenge or charge. Hence, the best application of Webb's DOK typology calls for recognizing the *range* of cognitive complexity that an item might elicit.

What to Call DOK's Central Thrust?

As described above, Webb's DOK main focus is the degree of automaticity vs. mindfulness and intentionality. This idea has appeared throughout a broad range of work in cognitive psychology. It has been a part of automaticity theory (Moors & De Houwer, 2006; Stanovich, 1990), ACT* (Adaptive Control of Thought) theory (Anderson 1992, 1996), expertise theory (Ericsson, 2014; Ericsson, Krampe, & Tesch-Römer, 1993), fluency theory (Bianearosa, & Shanley, 2015), schema theory (Anderson & Pearson, 1984; McVee, Dunsmore & Gavelek, 2005; Widmayer, 2004) and no doubt countless others. Even the field of sports psychology considers this issue, and even there it has a variety of names, Singer (2002) points out.

> …conscious vs. nonconscious, controlled vs. automatic, voluntary vs. involuntary, explicit vs. implicit, systematic vs. heuristic, willed vs. nonwilled, aware vs. unaware, internal vs. externally oriented, and intentional vs. unintentional…(p. 359)

We prefer to call this dimension *automaticity* (when not simply referring to it as *wDOK*[2]), knowing that we do not refer to the full automaticity on which automaticity theory usually focuses. Rather, we use Logan's (1985) idea that automaticity is a continuum, rather than a dichotomy. We see Webb's four DOK level's generally residing on the *less* automatic end of the continuum, with the most automatic cognition (e.g., word or letter recognition) simply taken for granted.

(Note that across all these literatures, there are important distinctions between observable performance and level of proficiency, skill or expertise. The latter must be inferred from performance, experience and/or credentials. This distinction is right at home in educational measurement, which distinguishes between raw scores of performance and latent ability. We use the terms *performance* and *proficiency* for these two different constructs.)

[2] *wDOK is meant to refer to Webb's DOK, or DOK as Norman Webb defined it. This stands in contrast with iDOK, meaning DOK as it is generally used and understood in the assessment industry.*

What is Automaticity?

Automaticity is the shift of cognition from "mindful" to "automatic" (Price & Driscoll, 1997, p. 473), including the use of problem solving strategies. Schema theory suggests that this is because "well-structured schemata that are automatically activated during problem solving" (Moreno & Park, 2010, p. 12). ACT* theory suggests that cognitive paths that previously had to be processed as multiple steps can now be processed as fewer steps – or even a single step (Anderson, 1990, 1992).

The point is that automaticity reduces cognitive burden, lower cognitive demands and reduces cognitive complexity. Mindful application requires "greater cognitive capacity usage" and "greater mental effort expenditures" (Salomon and Perkins, 1989, p. 125). Practice leads to "diminished attentional demands" (Moors & De Houwer, 2006, p. 298). "Learners no longer need to concentrate" (Ericsson, 2014, p. 82). Greater proficiency simply leads to less effortful cognition (Anderson, 1982; Fitts & Posner, 1967).

None of this is about general learning or improvement with broader skills. For example, chess playing does *not* improve memory or generic memorization skills. However, it is long established that as players increase in proficiency (as measured by ranking points) the ability to remember chess piece arrangements of increasing complexity improves (Chase & Simon, 1973; de Groot, 1965; Chase & Simon, 1973; Frey & Adesman 1976), but they remain no better than non-chess players at remembering random arrangements of chess pieces (Ericsson, 2014). That is, because their processing of the information of the arrangement of pieces *in chess* qualitatively improves, they are better able to store and retrieve information that is quite complex for lay people.

Automaticity makes skills and procedures "fast, effortless (from a standpoint of allocation of cognitive resources), and unitized (or proceduralized)" (Ackerman, 1987, p.. 4). In other words, with increased proficiency even complex skills and behaviors "eventually become routinized" (Salomon and Perkins. 1989. P. 130). That is, previously much more demanding and cognitively complex tasks become DOK Level 1 tasks. Though Webb calls it "strategic thinking" as poses it as distinct from "rote" application, clearly he means *intentional and deliberate* strategic thinking. In fact, this dynamic is not limited to those who have increased their proficiency (Logan, 1985). Practice can "merely make it less effortful and [more] automatic" even when it fails to "increase the quality of performance" (Ericsson, 2014, pp. R509-R510).

Automaticity Requires a Lateral Extension of DOK

This lateral extension of Webb's Depth of Knowledge structure as being imposed upon it from afar. Rather, DOK – as Webb presented it in 2002! – *is about automaticity.* The very words that Webb used to describe it are the words used in a variety of cognitive psychology branches.

Too often, those making use of DOK miss that DOK is about cognition at least as much as it is about the task in front of a student. While Webb describes kinds of assignments that may call on different levels of cognitive complexity, the range of

automaticity that he describes in each of his five content areas is clear. Anyone who doubts this should revisit his explanations in "Depth-of-knowledge levels for four content areas" (2002).

There is no question that automaticity (e.g., using Webb's language, is "rote" vs. "strategic") is a function of individual's familiarity and perhaps proficiency with the skills being called upon. Therefore, an item's cognitive complexity (i.e., the cognitive complexity that it elicits in test takers) quite often will vary, depending upon the familiarities and perhaps proficiencies of test takers. Hence, the best application of Webb's DOK typology calls for recognizing the *range* of cognitive complexity that an item might elicit.

References

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. *Psychological Bulletin*, 102, 3–27.

Anderson, J. (1982). Acquisition of cognitive skill. *Psychological Review*, 89, 369–406.

Anderson, J. R. (1992). Automaticity and the ACT theory. *The American Journal of Psychology*, 165-180.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, 51(4), 355.

Anderson, R., & Pearson, P. (1984). A schema-theoretic view of basic processes in reading comprehension. In P. D. Pearson (Ed.), *Handbook of Reading Research* (Vol. 1), 255-291. Longman.

Barber, J. (2018). Depth of knowledge and conceptual understanding. *Science Scope*, 41(9), 76-81.

Bechard, S., Karvonen, M., & Erickson, K. (2021). Opportunities and Challenges of Applying Cognitive Process Dimensions to Map-Based Learning and Alternate Assessment. *Frontiers in Education*, 6, 1-23.

Bianearosa, G. & Shanley, L. (2015). What is Fluency? In Cummings, K. D., & Petscher, Y. (eds), *The fluency construct: Curriculum-based measurement concepts and applications*. Springer.

Chase, W. & Simon. H. (1973). Perception in chess. *Cognitive Psychology*, 4. 55-81.

DeGroot, A (1965). *Thought and choice in chess*. Mouton.

Ericsson, K. A. (2004a). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine*, 79(10), S70-S81.

Ericsson, K. A. (2014b). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence*, 45, 81-103..pdf

Ericsson, K. A. (2014c). Expertise. *Current Biology*, 24(11).

Ericsson, K., Krampe, R., and Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychology Review*. 100, 363–406.

Fitts, P., & Posner, M. (1967). *Human performance*. Brooks/Cole.

Frey, P. & Adesman, P. (1976). Recall memory for visually presented chess positions. *Memory & Cognition*, 4(5), 541-547.

John R. Anderson, J.R. (1990). *The Adaptive Character of Thought*. Lawrence Earlbaum Associates, Inc.

Logan, G. (1988). Automaticity, resources, and memory: theoretical controversies and practical implications. *Human Factors*, 30, 583–98.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, 39, 367–386.

McVee, M. B., Dunsmore, K., & Gavelek, J. R. (2005). Schema theory revisited. *Review of Educational Research*, 75(4), 531-566.

Moors, A., & De Houwer, J. (2006). Automaticity: a theoretical and conceptual analysis. *Psychological Bulletin*, 132(2), 297.

Moreno, R. & Park, B. (2010). Cognitive Load Theory- Historical Development and Relation to Other Theories. In Plass, J. L., Moreno, R., & Brünken, R. (Eds.), *Cognitive Load Theory*. Cambridge University Press.

Price, E. & Driscoll, M. (1997). An Inquiry into the Spontaneous Transfer of Problem-Solving Skill. *Contemporary Educational Psychology* 22, 472–494.

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, 24(2), 113-142.

Shiffrin, M & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84, 127–90.

Singer, R. N. (2002). Preperformance state, routines, and automaticity: What does it take to realize expertise in self-paced events?. *Journal of Sport and Exercise Psychology*, 24(4), 359-375.

Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. *Developmental Review*, 10(1), 72-100.

Walkup, R. (2014, December 24). DOK Chart Sabotages Understanding of Depth of Knowledge. *Cognitive Rigor to the Core!*

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

Widmayer, S. A. (2004). *Schema Theory: An Introduction.*

Wine, M. & Hoffman, A. (2020). *Theory of the Item.* AleDev Research.