Reinvigorating Webb's Depth of Knowledge in Three Content Areas

Marjorie Wine
Assistant Director of Test Development
Accessible Teaching, Learning, and Assessment Systems (ATLAS)
University of Kansas

Dr. Alexander M. Hoffman
President
AleDev Research & Consulting

Paper presented at the April, 2023 annual meeting of the National Council of Measurement in Education in Chicago, Illinois.

Abstract

This paper updates Webb's invaluable 2002 *Depth-of-knowledge levels for four content areas,* replacing his Depth of Knowledge construct (wDOK) with the new revised Depth of Knowledge (rDOK). It focuses on practical application of DOK to classifying items for cognitive complexity – particularly in the context of USED's peer review process.

[Note: This paper is a companion paper to their AERA paper, Wine and Hoffman (2023), which dives more deeply into the conceptualization the theoretical background of our revised Depth of Knowledge (rDOK). *This* paper aims to provide a contemporary parallel to Webb's 2002 *Depth-of-knowledge levels for four content areas*, his clearest and most specific explanation of how his Depth of Knowledge construct (wDOK) can be applied to each of the content areas – unquestionably the *best* explanation of wDOK for content development professionals. Like that 2002 effort, this paper is focused on the practice of *applying* a single typology of cognitive complexity in different content areas that function fundamentally differently.]

In 1997, Webb published a "revolutionary article on alignment" (Forte, 2017, p. 6), laying out the importance of aligning assessment with instructional objective.

> Assuring the alignment between expectations and assessments can strengthen an education system in important ways. Teachers give more credence to documents they understand are in agreement, are useful, and will serve to benefit their students. Teachers, already overloaded with responsibilities, are better able to attend to expectations and assessments if they provide a consistent message and have credibility (p. 1).

He laid out six different criteria to examine for alignment, This new approach include six different criteria to consider on alignment, of which Depth of Knowledge was just one of the three he kept in later years and efforts (Webb, 2007).

- Balance of Representation
- Categorical Concurrence
- Depth of Knowledge Consistency
- Dispositional Consonance (eventually abandoned)
- Range of Knowledge Correspondence
- Structure of Knowledge Comparability (eventually abandoned).

Nonetheless, Webb's developing Depth of Knowledge (wDOK) construct has become the dominant definition and operationalization of *cognitive complexity* in the large scale standardized assessment. The United State Department of Education requires that evidence of cognitive complexity be gathered and examined as part of its peer review process (2018), "Documentation of adequate alignment between the State's assessments and the academic content standards the assessments are designed to measure in terms of content (i.e., knowledge and process), balance of content, and cognitive complexity" (2018, p. 47). wDOK is so taken for granted among assessment professionals that the CCSSO's 2022 National Conference on Student Assessment featured a scholar cited in this paper and a panelist assailing USED for requiring wDOK — despite the fact that neither "depth of knowledge" nor "Webb" appear anywhere in its *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process.* Their mistake could have been made by anyone, so dominant is wDOK in content development and alignment studies.

Unfortunately, the work of content development professionals (CDPs) – those with responsibility for the development and refinement of test items – is understudied within the assessment literature. CDPs work is rarely studied and CDPs rarely get significant roles in contributing to scholarly literature. Therefore, their tools may be mentioned, but *use* and *interpretation* of their tools flies under the radar of serious inquiries. This is as true for the use of typologies of cognitive complexity as it for the other tools and techniques used by CDPs.

Webb's 2002 explanation of his wDOK construct, *Depth-of-Knowledge for* Four Content Areas has been the most useful version for CDPs because it offers content-area-specific explanations of *each* of the four wDOK levels for *each* of the content areas (i.e., ELA, Mathematics, Science, Social Studies). In spite of these explanations, wDOK still often misunderstood and/or misapplied (Wine & Hoffman, 2023). This paper follows in the footsteps of Webb's 2002 effort, explaining the application of revised Depth of Knowledge (Wine & Hoffman, 2023) in three of the content areas[1]. Revised Depth of Knowledge (rDOK) addresses many of the sources of confusion around wDOK, simplifying its definition to focus exclusively on the central thrust of wDOK – automaticity vs. deliberation.

rDOK is a product of the Rigorous Test Development Project (RTD), an item-centric and validity-focused effort to increase the quality and reputation of large scale, on-demand, standards-based standardized tests. While acknowledging that test validity is about the uses and inferences made from standardized tests (AERA et al, 2014), RTD recognizes that those tests are being built upon and out of test items and sees that test validity requires *item validity*. That is, test must be built of items that elicit evidence of the targeted cognition for the range of typical test takers. Validity is "the most fundamental consideration in developing tests and evaluating tests" (AERA et al, p. 9) and developing tests for valid purposes require high quality professional attention to item development and refinement.

DOK's Central Central Thrust

Webb's Depth of Knowledge explanations (1999, 2002, 2005, 2007) cite numerous different indicators of the four different wDOK levels, particularly his 2002 version. However, throughout this typology's history, one central thrust has remained constant. This can be seen in the summary of wDOK in Table 1 (below), reprinted from Wine and Hoffman (2022).

Table 1

Summary of Webb's Depth of Knowledge Typology

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.
DOK 2	Skill/ Concept (Tactical Thinking)	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.

[1] We omit social studies because it lacks a consensus set of standards to parallel the *Common Core State Standards for Mathematics*, the *Common Core State Standards for Literacy and ELA*, and the *Next Generation Science Standards*.

wDOK's central thrust of *automaticity vs. deliberation* is apparent in how Webb differentiated his Level 2 from Level 1, a sentence he reused verbatim again and again (1999, 2002, 2005, 2006, 2007[2]) – sometimes multiple times in a single paper.

> A Level 2 assessment item requires students to make some decisions as to how to approach the problem or activity, whereas Level 1 requires students to demonstrate a rote response, perform a well-known algorithm, follow a set procedure (like a recipe), or perform a clearly defined series of steps." (2002, p. 4)

Hence, users of wDOK should not have focused on the type of skills or knowledge in question when classifying items and/or standards for cognitive complexity. Rather, DOK is about the *automaticity* of the response on the part of the test taker or student. Rote responses are low in cognitive complexity, as are "well-known" algorithms (i.e., a reference to test takers' command the of the algorithm, not its fame). Webb's original explanation of Level 3 spoke of "developing a plan" (1999, p. 3), a degree of deliberation that goes beyond the more tactical decision-making of Level 2. His bare 1999 explanation of Level 4 included "time to think" about "non-routine" work (p. 3), an explanation that he further developed through the years – while always stressing the Level 4 cognition requires so much deliberation that it requires more time than is available for an individual task on most large scale standardized assessments.

This is not to say that wDOK is *only* about automaticity vs. deliberation. Many of Webb's explanations include other contributors to cognitive complexity, particularly his 2002 paper. He lists various types of knowledge, skills and/or abilities (KSAs) typical of each level in each content area, though we believe these were meant to be illustrative examples the wDOK construct rather than to define it. Sadly, these have instead complicated the application of wDOK and by suggesting so many different minor determinants/contributors to cognitive complexity that they have both obscured wDOK's central thrust and created avenues to ambiguities in wDOK classification. rDOK maintains that central thrust of *automaticity vs. deliberation*, continuing the rich cognitive psychology tradition of dual process theories (Evans & Stanovich, 2013), simplifying the DOK construct by putting aside other issues.

Multiple Classifications for a Single Standard or Item

rDOK also includes ambiguity in classification, but not by making it unclear which singular level an item or standard should be classified with. Unlike wDOK, rDOK builds on the idea that because *automaticity is a function of proficiency* (Anderson, 1982; Ericsson, 2014; Widmayer, 2004), a single item can elicit different cognitive paths of different complexity by test takers of different levels of proficiency (Wine & Hoffman, 2023). That is, highly proficiency students may apply the targeted cognition with ease and without much thought to a problem that a less proficient test test taker might labor through.

For example, consider two tests takers:

- One who works to verify that they are grabbing the right tool, struggles to remember exactly how to use it and then double checks their work.
- One who immediately recognizes the problem and runs through the algorithm without checking anything.

These different paths to a final response – even if they both come to the same successful final response – are of differing complexities. rDOK – unlike wDOK – embraces this fact. That this complicates the

[2] Google Scholar lists five more uses of this sentence in publications with Webb as an author, to say nothing of the number of times Webb and others have used or quoted this sentences elsewhere.

meaning of cognitive complexity for items and that this is based on the variation in test taker cognition – as opposed to some arbitrary preferred cognitive path by others (Hoffman, 2022) – is a major strength of the rDOK approach. That is, rDOK is more reflective of *test taker cognition*, of the interactions between test takers and items, and thereby better supports meaningful inferences about test taker cognition.

Of course, high levels of proficiency are *not* the only driver of high automaticity or low deliberation. *Any* thoughtless application of skills – whether skilled or unskilled – is rDOK 1 cognition.

rDOK also allows for a single *standard* to be classified at multiple levels of rDOK cognitive complexity. Those in charge of a test (i.e., usually test sponsors, test owners, the clients of test development vendors) must determine the threshold of proficiency at which each standard should be assessed. Consider the archetypal example of the math facts of multiplication tables. Clearly, quick direct recall of each such math fact passes any expected threshold for proficiency. The question would be whether the kind of halting counting through a sequence to figure out a product – something that is too common – constitutes proficient *enough* performance. The halting-but-I-can-figure-it-out cognitive path can sometimes rise to rDOK 2 – and the exact language of standards themselves rarely rule out this level of application. It is a matter for those ultimately responsible for a test to decide whether that kind of less-proficient cognition fits within the boundaries of the standard. If so – as we generally think it should – many standards authentically allow for multiple levels of cognitive complexity in their application.

Similarly, some tasks appear to call for the kind of deliberation that makes for greater rDOK cognitive complexity, but the most proficient students may produce successful answers with more automaticity and therefore lower cognitive complexity (see below).

rDOK classification should recognize the range of cognitive paths that test takers may use to arrive at a successful responses when classifying items. It should also recognize the range of cognitive complexity that is acceptable as proficient with a particular standard.

rDOK for Items in Three Content Areas

Even when the same simplified idea of cognitive complexity is applied, its application works differently in each content area. This is because of the major differences between the content areas – coming as they do from such entirely different disciplines. This can be seen in the different structures of the dominant nationally recognized standards for each of them. A majority of states use each of the Common Core State Standards for Literacy (CCSS-Literacy), Common Core State Standards for Mathematics (CCSS-Mathematics) and/or the Next Generation Science Standards (NGSS) – or base their own standards upon them. Therefore, we use these three sets of standards as the operational definitions of each of these three cognitive academic content areas. (This paper does not address Social Studies because of the absence of such a consensus definition of this content domain.)

CCSS-Mathematics contains two sets of standards, the Mathematics Content Standards and Standards for Mathematical Practice (SMP). The Content Standards are the traditional granularly defined math knowledge, skills and conceptual tools, laid out grade by grade, with different threads and "domains" spanning multiple grades. This set of standards – like math standards that preceded CCSS-Mathematics – dominate large scale math assessments. The Math Practices have a much smaller role in assessment, perhaps because they are not tied to particular grades and CSS does not meticulously trace them through the grades and perhaps because they are not necessarily well suited to the constraints of large scale assessments.

CCSS-Literacy is quite different than CCSS-Mathematics. Unlike the other content areas, Literacy – whether part of English Language Arts (ELA) classes or embedded in other content areas – is built up of two closely related major areas: reading and writing[3]. CCSS-Literacy presents 10 *anchor standards* for each of these two major areas, each not entirely unlike CCSS-Mathematics's Math *Practices*. However, unlike those Math Practices, CCSS-Literacy traces how each of the 20 threads develops through each grade[4] in the K-12 grade span, and there is no parallel to the small skills and subskills of Math's *Content Standards*. These grade-by-grade developments of each of these anchor standards at least as much describe the increasing demands of text – both to be consumed and to be generated -- as they do student cognition. That is, later grade level texts demand more of readers/writers than earlier grade levels, and the standards reflect the development of grade appropriate texts through the K-12 grade span, with the same anchor standards applied in each grade.

NGSS's underpinnings contain both the science content area's own traditional content standards (i.e., *Disciplinary Core Information;* DCIs) and *Science and Engineering Practices* (SEPs) – both drawn from NRC's earlier A Framework for K-12 Science Education (2012). However, NGSS's standards – its *Performance Expectations* – are a select set (for each grade) of combinations of a single DCI and a single SEP[5]. This elevates the importance of NGSS's SEPs in a way that CCSS-Mathematics does *not* with its own practices. This reflects the fact the K-12 science education focus on the products of science (i.e., the DCIs) *and* the process of generating science knowledge (i.e., the work of *doing* science like a practicing scientist; the SEPs) in a way that K-12 math education does *not* focus on the generation of new mathematical knowledge (i.e., the work of practicing mathematicians).

Table 2

Automaticity in the Content Areas

CCSS-Mathematics	CCSS-Literacy Reading	CCSS-Literacy Writing	NGSS Science
Automaticity is the goal, a sign of true mastery with the tools that mathematics provides. The most skilled and proficient students can respond with automatic recognition of problems, automatic selection of tools and automatic use of tools	The most automatic skills in CCSS-Literacy are not found in the *Reading Information* or *Reading Literature* standards which are the focus of assessments. Instead, those Language standards are usually taken for granted. The automatic or deliberative application of the RI and RL standards depends on the relative proficiencies of each test taker vis-à-vis increasing demanding texts.	The type and qualities of texts that may be generated with automaticity generally increases across the grades. Expectations regarding the *contents* of writing (i.e., particularly *idea development*) continually increases (i.e., calling for deliberation), even as other aspects of writing become more automatic.	Automaticity with DCI knowledge is a sign of mastery of that part of science. However, the scientific *process* values deliberation, doubt and even requires replication. Thus, science has two different relationships to automaticity

[3] CCSS-LITERA also contains Speaking and Listening standards, but those standards *rarely* appear on large scale standardized assessment – not withstanding New York's Regents exams' longstanding *Listening* sections.

[4] Strictly speaking, grades 9-10 are built into a single grade band, as are grades 11-12.

[5] PEs also are linked to one of NRC's *Crosscutting Concepts*, but while these serve to connect various PEs (e.g., as when writing curriculum), they are not necessary to understand individual standards.

These three (or four) content areas have different relationships to the kind of automaticity that is at the heart of DOK, as shown in Table 2. This makes rDOK classification different for each content area.

There are four important principles that must be kept in mind when classifying items for cognitive complexity, regardless of the content area.

- The *only* part of the task and the cognitive paths the item prompts to be considered when classifying items for cognitive complexity is the *application of the targeted cognition* portion of those paths. Additional KSAs (Riconscente, Mislevy & Corrigan, 2016) may call on other levels of cognitive complexity, but the question at hand is whether the item is aligned to expectations *of this standard* for cognitive complexity. Therefore, classification is based on the uses of the KSAs in the aligned standard and *their* cognitive complexity, not other KSAs. (See note on this in ELA section below.)
- Cognitive complexity is *not* item difficulty. Item difficulty is measured empirically as the share of test takers that respond to an item successfully. Though everyone – including ourselves – sometimes slips into this confusion, one must remember that these are distinct concepts. Tasks with low cognitive complexity can be quite difficult (e.g., *what is the capital of the 41st state to enter the union?*). rDOK cognitive complexity is about the automaticity-deliberation continuum in cognition, and not difficulty. Certainly, there are many construct/standard irrelevant ways increase the difficulty of an item, and those should be taken as aligning an item's cognitive complexity to a standard's.
- There is no singular preferred cognitive path to a successful response that can be used to determine a singular cognitive complexity for an item (Hoffman, 2022). Because items can prompt different cognitive paths by different test takers, and the rDOK complexity of those paths can vary by test taker proficiency, items must be classified against each rDOK level and they *are not mutually exclusive.* A single item may have as many as three potential rDOK levels. This recognition that different solution paths can have different levels of cognitive complexity *does* stand in contrast with traditional consideration of cognitive complexity. This is the greatest difference between contemporary use of Depth of Knowledge and this rDOK approach.
- The modality of an item can greatly influence its cognitive complexity. Constructed response items usually function quite differently than selected response versions of quite similar – at least on the surface – questions. Recognition of a correct answer may be much more automatic than having to generate it oneself. Certainly, the quality of the distractors of multiple choice items has enormous impact on cognitive complexity, as facially implausible distractors can be dismissed without deliberation. Selected response items that lack facially plausible distractors are quite likely to be classified at rDOK 1 and are more likely to *only* be classified at rDOK 1. Selected response items that lack a singular and definitively correct answer option (i.e., often directing "select the best answer," and perhaps in fact meaning select the least bad answer option) rarely include successful rDOK 1 cognitive paths.

Of course, rDOK levels of items are not determined simply by the nature of the aligned standard or the appearance of particular words in an item. Rather, they are always determined by the level of automaticity or deliberation of the cognitive paths it prompts in test takers.

rDOK Mathematics Items

Mathematics items on large scale assessment reliably tend to each focus on a single mathematics problem, each usually aimed a single content standard. Because of the fine granular nature of mathematics standards, the role of the targeted cognition in a given solution is generally clear – particularly with Content Standards. While individual items do not have to assess an entire standard, the fine grain size of mathematics Content Standards makes that far more likely, contributing to the ease of identifying the targeted cognition in a solution path. However, the common mathematics goal of automaticity as the most desired level of mastery of the KSAs in most mathematics standards does *not* mean that those KSAs cannot be applied with more deliberate levels of cognitive complexity.

It can be much more difficult to pinpoint use of the Standards for Mathematical Practice as part of a cognitive or solution path. Nonetheless, they actually can support the automaticity of high proficiency when they are properly ingrained, and they can also prompt test takers pause and deliberate for a moment – even when properly ingrained.

Math rDOK 1 – Recall

This level of automaticity is the goal of most math instruction. It includes recall of declarative knowledge such as math facts, terms and names. It also includes recall of procedural knowledge, such as standard algorithms for solving quickly recognized problems. Some test takers use the memorized KSAs in the aligned standard in a rote fashion, without having to make any significant decisions along the way – because their lessons and practice have made a complete solution path available to them in advance. Similarly, direct application of definitions is also rDOK 1 cognition. That is, when memorized fact, procedures and/or definition are applied with practiced and facile ease, it is rDOK 1 cognition.

Virtually any sort of math problem on an on-demand assessment can be completed in an rDOK 1 fashion, provided that the test taker has had sufficient practice with similar problems and quickly recognizes the applicability of those solutions paths to this item. Experts in curriculum should be able to recognize when an item presents such an opportunity to test takers, as opposed to the novelty the forestalls rDOK 1 cognition.

Items that are amenable to backsolving usually allow rDOK solutions paths, if they even are accepted as being aligned with the contents of aligned standard[6]. Selected response items that lack plausible distractors are *quite* likely to be only rDOK 1 items. Selected response items that lack a singularly and definitively correct answer option (i.e., often saying "select the best answer," and perhaps meaning *select the least bad answer option*) rarely include successful rDOK 1 cognitive paths.

Math rDOK 2 – Skill/Concept (Tactical Thinking)

rDOK 2 cognitive paths require some test takers to consciously make a substantive decision – or multiple decisions – in order to solve the problem. This may take the form of needing to figure out which prepared tool or algorithm to use, because it was not immediately apparent.

[6] While backsolving is usually a shortcut around the intended application of the targeted cognition, it *is* a path for test takers to solve the problem in front of them. Therefore, it is consistent with that basic goal of mathematics of solving the problem in front of you. Backsolving-compatible problems are accepted on assessments commonly enough that we must find a way to rate their cognitive complexity and their motivation is almost invariably to reduce the complexity of a problem and/or its soluton path.

This may involve some sort of translating or transforming the problem in order put it into an appropriately recognizable form. So long as these steps are part of the targeted cognition, they create DOK 2 item.

Prepared approaches that require decision making along the way may be rDOK 2 cognition, if the decisions have the potential to lead to an incorrect response. Cognitive paths that require test takers to decide or figure out what to do next at many points along the path remain rDOK 2, regardless of how many such decisions the test taker must make – so long as the decisions are not made *before* reaching that point in problem (i.e., the *planning* of rDOK 3).

Interpreting results and/or drawing a conclusion is usually rDOK 2 cognition, when it comes to the kind of well-bounded conclusions that one sees on large scale math assessment.

Items that more proficient students can solve with rDOK 1 of levels of automaticity by using prepared approaches may rise to rDOK 2 for some test takers if the solution path is something they can work out even when they do not recognize the problem or if they have not internalized the tool sufficiently to have the *rote* response of more proficient test takers.

Math rDOK 3 – Strategic Thinking

rDOK 3 solution paths arise when problems are sufficiently novel and complex to some test takers that they must *plan* a solution path for themselves before working their way through that solution (i.e., they did not come to the task with a suitable prepared approach). This differs from lower rDOK levels because in these rDOK 3 cases no approach appears sure to reach a solution, up front.

Alternatively, the test taker might dive in, but have to take stock and – at times – return back to an earlier step and take a different fork in the path. That is, they must evaluate where things went wrong and go to back to an earlier point to make better progress towards the solution. This kind of reflection on their progress can make for a DOK 3 item, so long as the aligned standard includes *this* kind of cognition. This important part of Math Practice #1 (*i.e., Make sense of problems and persevere in solving them)* can be applied when responding to all but the simplest of problems, but it is *not* a part of every content standard.

Items that require test takers to explicitly explain their reasoning – not just show their work – are also rDOK 3, so long as that is part of the aligned standard(s).

Math rDOK 4 – Extended Thinking

K-12 mathematics instruction often positions mathematics as a set of tools to be used for a variety of *other* purposes. Certainly the Standards for Mathematical Practice are mindsets or habits of mind that can be used in a broad array of contexts. Therefore, none of CCSS-Mathematics standards themselves *require* wDOK 4's extended planning and coordination of multi-phase projects that synthesize multiple components and considerations – not even apart from large scale assessment – although some of the SMPs *could* drive and coordinate the kind of extending thinking of larger projects. For example, critiquing the reasoning of others (i.e., part of SMP #3) can be applied in contexts that require sufficient research to recognize problems in an argument and/or build a counter-argument that it would be rDOK 4 cognition, but that is not likely to be part of a mathematics classroom[7].

[7] The broad applicability of the CCSS-Mathematics' Standards for Mathematical Practice to other contexts – even non-quantitative contexts – is incredibly important to us. In fact, we have viewed *these* sorts of tools in mathematics as its most important for over three decades. However, the parts of mathematics that drive rDOK 4 cognition are not

rDOK Science Items

One of the great challenges of assessing the Next Generation Science Standards is how much is packed into each standard (i.e., each Performance Expectation). Even the DCI component of a single NGSS standard can be quite broad, to say nothing of the difficulty of assessing an SEP within the confines of an on-demand large scale assessment. As the assessment industry continues to try to figure out how to address these standards, a common approach is to assess a single standard across multiple items is the scenario set. That is, while Mathematics and ELA often use a shared stimulus for multiple items (i.e., usually with each item addressing a *different* standard), Science assessment often uses the multiple items of a single scenario set to address different components/aspects of a *single* NGSS standard. Like CCSS-Mathematics' Content Standards, relatively traditional items can get at aspects of the *DCI component* of an NGSS standard. The challenge for assessment developers has been figuring out how to address the SEP side of NGSS standards within the context of such assessments.

Science rDOK 1 – Recall

Science items focused on the DCI component of an NGSS standard often function very much like Mathematics items, from the perspective of rDOK. That is, the instructional goal is rDOK 1 levels of automaticity in the application of knowledge, algorithms and/or approaches. Therefore, some test takers may utilize the DCI knowledge or SEP skill(s) to *immediately recognize* (i.e., without deliberate consideration) true or accurate statements, including statements of fact and/or explanations. *This is a function of the automaticity of the recognition, rather than the complexity of the statement's contents.* More procedural tasks can also be classified at rDOK 1 if some test takers can use a well-practiced procedure to arrive at a successful response, without needing to pause or deliberate along the way.

Selected response items that lack plausible distractors are *quite* likely to be only rDOK 1 items. Selected response items that lack a singularly and definitively correct answer option (i.e., often saying "select the best answer," and perhaps meaning *select the least bad answer option*) rarely include successful rDOK 1 cognitive paths.

Science rDOK 2 – Skill/Concept (Tactical Thinking)

Items can be classified as rDOK 2 when some test takers must *work out* their response more deliberately (i.e., without the more rote cognition of rDOK 1). This can include identifying the accuracy of statements and/or explanations, figuring out how exactly to use particular aligned KSAs for *this* problem, or even which practiced conceptual or physical tool to use. Cognitive paths which include decision making along the way (i.e., tactical decision making) are also rDOK 2 cognition.

Certain kinds of deliberate scientific reasoning are also rDOK 2 cognition. This includes translating and/or transforming a problem into something to which the test taker can apply those rote tools (e.g., a word problem or a less conventional form of an equation). It also includes straightforward acts of interpreting a result and/or drawing a conclusion using the aligned KSAs in a conscious fashion.

---

truly a part of K-12 math standards, and therefore not eligible to be *the targeted cognition* for large scale k-12 assessment.

Science rDOK 3 – Strategic Thinking

rDOK3 science items require using KSAs from the aligned standard for the prospective and/or retrospective reflection of rDOK 3. That is, some test takers will use those KSAs to develop a plan – not just grab a previously developed plan – for how to reach a successful response.

One of the SEPs is explicitly about planning (i.e., *Plan and carry out an investigation*), but rDOK 3 planning is *not* limited to PEs that include this SEP. Any time a test taker develops a plan for their solution path before executing it they are engaged in the kind of prospective reflection of rDOK 3 cognition. When this kind of planning can be seen in the *aligned* standards' KSAs, this kind of solution path classifies as item as rDOK3. Of course, making use of *predetermined* plans is *not* developing a new plan.

Similarly, items that call on test takers to reflect on – usually to explain – how or why they came to a particular response are also rDOK 3 items, subject to that being part of the aligned standard. However, recognizing and/or offering a *previously* learned explanation for a phenomenon is *not* rDOK 3 cognition, and may simply be rDOK 1 cognition. It is the *development* of the explanation (i.e., reflection) that makes it rDOK 3. As the disciplines of science are quite focused on explanations, ideally science assessments would contain *many* of rDOK 3 items. However, the modality of most selected response items usually calls on test takers to *recognize* and/or *identify* explanations, rather than *develop* them. Thus, the exact same stem can lead to rDOK 1 or 2, or rDOK 3 cognition, depending on the modality of its item.

Science rDOK 4 – Extended Thinking

Science has a reverence for the most deliberate and careful thinking, almost diametrically opposed to the automaticity goals of K-12 mathematics and CCSS-Literacy's Language standards. The scientific method is *intentionally deliberate*. The modern scientific process adds additional levels of deliberation with its emphasis on replication of previous work as the ultimate contributor to credibility. NGSS's eight SEPs lay out a full scientific process. Quickly and accurately arriving at an answer is *not* how science builds knowledge, and science is *very much* about building knowledge. There are debates about exactly how much K-12 science ought to or does emphasize *doing* science, but there is never a question that this should be an important part of K-12 science curricula.

There is no question that NGSS contains many standards that suggest rDOK 4 tasks. Some of the individual SEPs call for sufficiently complex work and cognition for this level of cognitive complexity. Unfortunately, those kinds of extended projects are not practical within the confines of on-demand large scale assessment.

rDOK ELA (English Language Arts) Items

Classifying ELA reading items for cognitive complexity requires careful examination of the text in question for its standard-specific demands on readers. This must be related to the presumed reading proficiency of the range of typical takers of this test. Texts that are relatively more demanding with a particular standard will generally require greater levels of deliberation (i.e., rDOK levels) in the application of that targeted cognition than texts that are relatively less demanding with that particular standard. Hence, *very* similar items on the same ELA test can elicit quite different levels of cognitive complexity when they are based on texts with different levels of standard-specific demands. That is, cognitive complexity is often determined by the *interaction* between a text's demands and a test taker's level of proficiencies.

Assessment runs into a unique problem with assessing reading of text. Authentic reading generally happens *internally*, without observable indicators to evaluation. Therefore, assessments have to add something in order to have something to observe. Working out a problem and responding with a solution *is* the core of K-12 mathematics. Answering questions about a text is *not* the core of reading a text. It requires *additional* reading/making sense of the question (i.e., the full item) – generally after having read the text. Thus, both the task's difficulty and overall cognitive complexity can stem from other issues and characteristics in the item rather than simply the targeted cognition applying to the authentic reading. This is unique to reading[8]. This part of the cognitive path (i.e., interpreting the meaning of the *item*) should *not* be evaluated for cognitive complexity for reading items, though interpreting the meaning of the *text* quite often is the part of the path that should be considered.

Writing tasks epitomize the impossibility of determining the cognitive complexity of a cognitive path merely by examining the final product. One cannot tell whether the product was produced through the facile ease of a highly proficient and skilled drafter of clean copy or was instead produced by the careful planning, execution and revision of a more deliberate writer. One cannot tell whether a highly proficient writer produced it with ease or less proficient student produced it with diligence. Luckily, rDOK classification does not require examination of any particular test takers' work product. Instead, it is about examining the charge of the item and being open to the potential for those different paths. The quality of the final product does not indicate the tasks' cognitive complexity; the automaticity and/or deliberation of the writing process is what determines rDOK cognitive complexity.

Writing also presents a unique assessment challenge. Authentic writing is about the *integration* an array of skills – found both in the CCSS-Literacy Language standards (e.g., grammar, spelling, punctuation, vocabulary) and the CCSS-Literacy Writing standards (e.g., idea selection, idea development, organization). The practice of *construct isolation* in assessment – one driver of the use of large numbers of faster items (e.g., multiple choice items) – is incompatible with this fundamental aspect of writing. Though rDOK does *not* focus on the breadth of skills brought to bear – undoubtably a contributor to a task's complexity – we acknowledge this creates particular challenges for assessing writing. In practice, writing is often assessed with larger constructed response tasks that are simultaneously aligned with multiple ELA standards. This requires very careful teasing out of each different set of KSAs and the range of automaticity-deliberation with which they may be applied when generating a successful response[9].

In assessment practice, construct isolation drives the creation of items that align to a single standard – either a *reading* standard or a *writing* standard – even though CCSS was designed to integrate reading and writing, almost as much as writing itself is the integration of a range of skills. Selected response items (e.g., multiple choice items) tend towards rDOK 1 or rDOK 2 level, and support this kind of construct isolation that allows for the separation of reading and writing. However, the kinds of *CCSS* tasks that call on rDOK 3 cognitive complexity often bring

[8] "Make sense of a problem" is literally the beginning of CCSS-Mathematics' first Standard for Mathematical Practice. It *is* a fundamental aspect of mathematics. "Make sense of the question" is *not* part of the any CCSS-Literacy standard.
[9] When an item is aligned to many different standards, there may be such a multiplicity of paths to a successful response that success is possible without use/consideration of a particular standard. It may even be possible that most of the aligned standards have a successful path that allows that standard to be ignored.

together reading and writing, as the authors of CCSS (e.g., Susan Pimentel) have always wanted. As discussed below, at the rDOK 3 level, CCSS reading and writing becomes difficult to disentangle.

ELA rDOK 1 – Recall

rDOK 1 reading items can be answered fluidly by some test takers, almost as though by rote. This includes surface-level reading comprehension of texts easily within the reading level of a test taker. Straightforward inferences that do not require deliberation or studying the text are also rDOK 1 reading applications for such test takers. Generally, items whose correct answers are immediately obvious to some readers are rDOK 1 items for those readers. Even if a test taker conscientiously decides to go back and *confirm* their answer, the fact that they accurately understood and retained the answer without deliberate effort makes this rDOK1 cognition. With selected response items, this recognition of the successful response should be evaluated after test takers after read *all* of the answer options, not merely after completing the stem.

rDOK 1 writing items similarly can be successfully responded to though with fluid ease by some test takers. That is, they can be produced as – or nearly as – quickly as the test taker can physically or electronically produce text. This is *not* necessarily a function of text length, as a rough draft or brain storm paragraph – or even far longer – may simply flow out of a writer as fluidly as if by rote. As explained above, apparent polish of the writing does not necessarily indicate that it was not produced through rDOK 1 cognition.

Selected response items that lack plausible distractors are *quite* likely to be only rDOK 1 items. Selected response items that lack a singularly and definitively correct answer option (i.e., often saying "select the best answer," and perhaps meaning *select the least bad answer option*) rarely include successful rDOK 1 cognitive paths.

ELA rDOK 2 – Skill/Concept (Tactical Thinking)

rDOK 2 reading items require some test takers to stop and think through a question and/or a text to come to a successful response. That is, the answer is *not* immediately obvious and must be come to through some level of conscious deliberation. It may entail revisiting the text, but this is *not* a requirement for rDOK 2 cognition. Simply having to work through the task consciously is sufficient. That is, either the item gets at an idea or understanding that some test takers did not get automatically when they read through the text or it gets at something some test takers stopped to consider – or figure out – when originally reading text.

rDOK 2 writing items similarly demand more deliberation from some test takers than the easy fluidity of rDOK 1 processes. This may occur when figuring out what point to make *or* figuring out how to make it, and may just take the form a pause in the middle of writing to figure out what to write next (i.e., a tactical decision). Going back to fix or alter previous words and/or sentences can be rDOK 2 cognition when it is done in the flow of production – as opposed to the more separate process of a later review for substantive alteration that is found with rDOK 3 writing processes. Because line/copyediting is not about substantive chances to contents or presentation (i.e., examining the *thinking* contained in the writing), such reviews still are a part of rDOK 2 writing.

ELA rDOK 3 – Strategic Thinking

The rDOK 3 reading cognition of *developing a plan* quite rarely occurs with literary passages. On the other hand, informational texts may require developing a plan for making sense of them, at least by some test takers. *Developing* such a plan is what makes for rDOK 3 cognition,

not following a plan. Therefore, using an assigned graphic organizer or previously prepared plan for how to attack a particular type text is *not* rDOK 3 cognitive complexity. On the other hand, selecting the appropriate kind of graphic organizer from array of possibilities in light of the text in question *does* constitute sufficient planning for rDOK 3.

Any sort of text can lead to the retrospective reflection (on the reader's *own* thinking) that is the other hallmark of rDOK 3 cognition. That is, when a reader seriously and substantively challenges their own understanding of what they have read, revisiting their thinking and its basis in one or more texts, they are engaged in rDOK 3 reading. This is distinct from the minor revisiting of text for confirmation or minor investigation, and instead is about the reader reflecting on their own thinking even more than reflecting on the text. For example, one could return to a children's text as a proficient adult reader and rethink the values and assumptions that one had in decades earlier, simultaneously interrogating both the text and oneself as a reader.

rDOK 3 reading tasks simply do not occur on large scale assessment very often, in part because of the time that such reflection requires simply runs into the testing time constraints in large scale assessment. However, it is not unheard of. For example, the California Bar Exam includes a number of 60-minute essays, each based on a legal scenario described in just a few hundred words. This assessment allows – and may be best taken with – rDOK 3 reading, as test takers often *should* spend time interrogating their own understanding of the subtleties of what they have read.

This bar exam example is similar to the kind of writing that CCSS-Literacy so strongly prefers: writing about reading. It also demonstrates how rDOK 3 *reading* is often part of rDOK 3 *writing*. That is, high demands for careful and thoughtful (and perhaps accurate) writing often call for prospective reflection on what is to be written – the planning of contents and how to present them. This gets to ideas, idea development, organization and potential voice and audience – and certainly purpose. When writing text this demanding is based another text, the retrospective reflection on reading merges with the prospective reflection on writing. This is not unusual *generally*, but *is* unusual in the context large scale assessment. Sometimes, this rDOK 3 cognition is truly just about revising understanding the earlier text, and sometimes it is about planning the text to be produced. That is, writing about text does not *necessarily* mean there is both rDOK 3 reading and rDOK 3 writing – and it might not even include either.

Other rDOK 3 writing simply requires planning out a piece before it is drafted and/or engaging in a substantive revision process of earlier drafts. That prospective planning is quite feasible for any assessment with a real writing component, even if the plan is little more than a quick 'n dirty five line outline for an essay. This kind of planning does *not* require rDOK 3 reading. It simply requires planning out the piece before executing on that plan.

ELA rDOK 4 – Extended Thinking

As with the other content areas, rDOK 4 cognition is not found on large scale assessments. Even multi-day bar exams do not devote enough time to any one task for that. rDOK 4 is about extended projects that connect multiple ideas, usually multiple texts and require multiple sessions. This level of deliberation and reflection simply requires stepping away from the work and returning to it with renewed eyes.

rDOK 4 reading almost always brings together multiple texts, though that may also occur with lower levels of cognitive complexity. At this level, the texts must present a sufficient array and/or depth of ideas and perhaps even variety of presentations to require this level of

deliberation. This is not any typical type of reading, and it is hard to imagine that it would not be connected to some written product. Certainly library research and scholarly work can entail rDOK 4 reading (e.g., literature reviews). As K-12 research papers, theses and dissertations are forms of assessment, rDOK 4 reading is found in assessment – just not large scale standardized assessment.

rDOK 4 writing is often connected to rDOK 4 reading, though it need not be. It calls for some sort of significant pre-drafting planning and thinking, and a truly significant revision processes – likely multiple rounds and quite often with feedback from others. This degree of deliberation and care through the process of writing often includes consideration of many dimensions of effective writing (e.g., voice, style, tone, organization, flow, idea, allusions/ references, evidence, idea development, length and/or vocabulary in light of intended audience and intended purpose or effect).

rDOK Social Studies Items

Because there is no broadly accepted national set of standards for Social Studies, we do not have the kind of construct definition or domain model for the kinds of cognition that Social Studies assessments would target. More specifically, we do not have a sufficiently credible singular source for the kinds of higher order thinking skills that K-12 Social Studies is aimed towards.

Clearly, items that target the declarative knowledge of names and dates would elicit rDOK 1 cognition, but we know that every dedicated and thoughtful Social Studies teacher sees that knowledge as serving more sophisticated thinking skills, in addition to its own value. When[10] national – even if voluntary – social standards emerge, we will update this document to address rDOK's application to K-12 Social Studies.

Classifying Standards for rDOK Cognitive Complexity

Because rDOK is intended for use in large scale assessment development, it is intended to be used to compare i) the cognitive complexity expected in standards to ii) the cognitive complexity of the application of the targeted cognition by test takers when responding to items. This requires classifying standards for rDOK cognitive complexity. Because test development requires so many items to be produced for each tested standard, item review is an ongoing process that is repeated every cycle, whereas standards classification for cognitive complexity only needs to be done once.

Standards classification differs from item classification in that standards classification is more about the aspirations of the standards and educators, whereas item classification is more about messiness in the variety of test taker cognitive paths. Standards classification does not have to address all the complexities that a realized item evokes. It can just focus on the educational objectives of the standard.

As stated above, standards classification requires deliberate decisions about the level of proficiency that a standard assumes. Is it sufficient for a student/test taker to only have enough mastery of the KSAs in the standard to eventually stumble to a successful response? If so, the standards allows for higher rDOK levels with their greater deliberation. Or, does a standard require sufficient command of the KSAs in the standard that students/test takers should be able to apply

[10] Though the education standards movement has long floundered on the challenges of devising broadly acceptable social studies standards, we remain hopeful that it will someday succeed.

them with great automaticity? If so, the standard only allows lower rDOK levels – perhaps only rDOK 1.

Some standards clearly call only for rDOK 1 cognition. These include standards that describe declarative knowledge that test takers should know – the kinds of things that cannot be figured out and must simply be memorized. Any KSA that students should perform with high levels of automaticity (e.g., understanding words and sentences when read), even if not a memorization/recall task, also makes for rDOK 1 cognition.

Standards that include rDOK 2 cognition call for or allow for decision making in how to apply KSAs that are *not* automatic or merely the rote application of predefined procedures or knowledge. These standards may *also* allow for rDOK 1 cognition, as rDOK classification is not mutually exclusive.

Standards that call for rDOK 3 cognition call for prospective or retrospective reflection on work, such as planning a work process in advance or reviewing the logic of previous work. Such standards might ask students to come to an answer *and* to explain it. Note that this is *not* the same thing as providing an explanation for some phenomenon or occurrence, as that may notbe a reflection on the student's *own* thinking – and may entail simply recalling a previously learned explanation (i.e., perhaps simply rDOK 1). Standards that do not describe such prospective or retrospective reflection do *not* call for rDOK 3 cognition, even if many tasks that make use of the KSAs in the standard may benefit from such reflection.

ELA standards are the least likely to call on a singular level of cognitive complexity, as cognitive complexity with ELA tasks is generally more driven by the standards-specific demands of a text (i.e., usually related to the relative grade level of the text), rather than the nature of the particular standard in question.

Cognitive Paths Not to Consider

While this approach to classifying cognitive complexity requires examining all the cognitive paths that test takers might take in response to an item, it does not *really* require considering *all* the paths the test takers might take.

First, only *successful* paths should be considered. Test takers may make any number or sort of mistakes that lead them sufficiently astray that their cognitive path takes them far from a successful response – or even from the targeted cognition! This requires the leadership of a project (e.g., leadership from the test owner and the test developer, perhaps in consultation with educators or other curriculum specialists) to decide about what counts a *successful* response for multi-point items (i.e., gaining all the points, half the points, a single point, or some other threshold)[11].

Second, although selected response items are quite common in large scale assessment and sheer guessing *can* lead to a successful response (i.e., usually 25% of the time), sheer guessing does *not* even attempt to engage with the contents of the item and has no chance of making use of the targeted cognition. Therefore, this strategy is not relevant to classifying cognitive complexity. However, when guessing is just a part of more complex strategy that includes more thoughtfully ruling out some answer options, that cognitive path *should* be considered – or at least the application of the targeted cognition in that path.

[11] The question of how large a group is need to meet the "some test takers" threshold is another question that project leaders must determine before lower level team members do their evaluations of cognitive complexity.

Third, cognitive paths that do *not* make use of the targeted cognition should not evaluated for cognitive complexity – as it is *the use of the targeted cognition* that should be the focus of any alignment question – even if they would produce a successful response. However, if prominent or likely paths to a successful response to *not* rely on the targeted cognition, the item is poorly aligned with the content standard's substance, and this should be noted and corrected. Items that are amenable to backsolving pose a particular challenge to alignment for precisely this reason.

Last, adoption of an ineffective strategy that requires substantially stepping back in the process to adopt a more appropriate strategy does not make the application of some targeted cognition more deliberate. A second bite of the apple is simply a second bite, not a more complex bite. For example, some students are taught to review items *before* reading the stimulus, so they know what to look for. This does *not* count as returning to a text. This use of a *predetermined* strategy is not rDOK 3, because the strategy is not developed by the test taker for the item. Of course, the use of the this kind of approach is not likely to be part of the KSAs in the aligned standard. Therefore, these additional steps are *rarely* part of determining the rDOK cognitive complexity of the item.

Final Discussion

This approach to classifying the cognitive complexity of items (and of standards) relies upon deep and *grade level specific* knowledge of content, expectations about student cognition and even instruction. This is no keyword-based approach that can be used by those with just passing knowledge of the range of typical test takers or their educational experiences. It relies on the well-earned professional judgment of experienced educators – though not necessarily just practicing K-12 teachers. In fact, this relies on knowledge that no single teacher develops on their own. Instead, they learn from each other[12] about a broader range of children, children's learning experiences and they various ways that children approach and work through the tasks they are given.

The RTD Project from which this rDOK arose is about *increasing* item validity – their ability to elicit evidence of the targeted cognition for the range of typical test takers – so that the inferences made from large scale assessment are based upon better building blocks than the assessments that are so distrusted by K-12 teachers and those who are deeply committed to their content areas. This requires deep knowledge of content, deep knowledge of student cognition and deep knowledge of how assessment works.

Considering the authentic cognitive paths that test takers might take when responding to an item requires careful examination of the wording, ordering of elements of an item and the various and subtle ways that that an item can signal or suggest a path to test takers. Even items that follow the exact same template can prompt different sorts of paths because of this this kind of subtle communication, often as a result of the personal or educational background and experience of different test takers. Just as high quality Fairness reviews rely on careful and deliberate thinking about each item, so should other forms of item review.

While we see the efficiency gains to made with further automation of assessment development, greater use of templated items and even the use of artificial intelligence, we are *quite*

[12] Similarly, those without experience as teachers can learn about common instructional approaches and about the cognition of students/test takers. This allows teachers and non-teachers alike to learn about levels and children with which they lack first hand experience. If course, it requires openness – perhaps even eagerness – to learn from colleagues who *do* have this expertise.

concerned that such gains will continue to come at the expense of item validity and of test validity[13]. Standardized testing is not going to gain credibility with the public or with educators by being more efficient. Improved credibility requires tests that *better* reflect the goals of the content area(s) and better assess the various kinds of cognition – successful and unsuccessful – that that test takers engage in when trying to do the work of the content area. It is vanishingly unlikely that that can be accomplished so long as efforts to improve test development practices are aimed at doing it cheaper and making tests faster – as opposed to making items and tests more valid.

We are fully aware that this approach to classifying cognitive complexity swims against that tide. Rather than making cognitive complexity review an easier hoop to jump through, rDOK forces alignment reviewers to think more careful about test taker cognition and how they might interact with items. We consider that its greatest strength.

[13] Anything that increases the predictability of content on a test or the manner in which it is assessed undermines assumptions about a test being a random sampling from a content domain, increasing opportunities for inappropriate test preparation and undermining inferences that generalize from test taker test performance to broader proficiency in the content domain.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1985). Standards for Educational and Psychological Testing. Washington, DC: AERA

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA

Anderson, L.W., Krathwohl, D.R., Airasian, P, Cruikshank, K, Mayer, R., Pintrich, P. Raths, J. & Wittrock, M. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of Educational Objectives (Complete edition)*. New York: Longman.

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.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Handbook I: cognitive domain*. New York: David McKay.

Clinton, J. M., & Hattie, J. (2021). Cognitive complexity of evaluator competencies. *Evaluation and Program Planning*, 89, 102006.

Clinton, J. M., & Hattie, J. (2021). Cognitive complexity of evaluator competencies. *Evaluation and Program Planning*, 89, 102006.

Council of Chief State School Officers. (2009). *Models*. Retrieved from http://www.ccsso.org/Projects/alignment_ analysis/models/418.cfm

Crowe, A., Dirk, C. & Wenderoth, M.P. (2008) Biology in Bloom: Implementing Bloom's Taxonomy to enhance student learning in biology. *CBE-Life Sciences Education* 7, 368-381.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3), 223-241.

Flowers, C., Wakeman, S., Browder, D. M., & Karvonen, M. (2009). Links for academic learning (LAL): A conceptual model for investigating alignment of alternate assessments based on alternate achievement standards. *Educational Measurement: Issues and Practice*, 28(1), 25-37.

Forte, E. (2017). Evaluating Alignment in Large-Scale Standards-Based Assessment Systems. Council of Chief State School Officers.

Hess, K., Jones, B., 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*. ERIC Document (Online Database).

Hoffman, A. (2022, December 21). Do items have a central complexity? *Complex Variety*. https://www.aledev.com/blog/2022/12/21/do-items-have-a-singular-cogntive-complexity/

Huraian Sukatan Pelajaran Sejarah Sekolah Menengah Atas (2012). Retrieved from http://web.moe.gov.my/bpk/sp_hsp/sej/hsp_sej_f4b.pdf

Kennedy, J. (2008). Bloom's rose [online image]. *Wikimedia Commons*. https://commons.wikimedia.org/w/index.php?curid=4000460

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4), 212-218.

Lindquist, E. F. (1951). Preliminary Considerations in Objective Test Construction. In Linquist, E. F. (ed), *Educational Measurement*. American Council on Education.

Marshall, G. (1990). *Pretty Woman*. Touchstone Pictures; Silver Screen Partners IV; Regency International Pictures (uncredited).

Miles, A., Charron-Chénier, R., & Schleifer, C. (2019). Measuring automatic cognition: Advancing dual-process research in sociology. *American Sociological Review*, 84(2), 308-333.

Paul, R. W. (1985). Bloom's Taxonomy and Critical Thinking Instruction. *Educational leadership*, 42(8), 36-39.

Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). Evidence-centered design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (p. 40–63). Routledge/Taylor & Francis Group.

Schneider, J. (2014). *From the Ivory Tower to the Schoolhouse: How Scholarship Becomes Common Knowledge in Education*. Harvard Education Press.

Schneider, M. C., Huff, K. L., Egan, K. L., Gaines, M. L., & Ferrara, S. (2013). Relationships among item cognitive complexity, contextual demands, and item difficulty: Implications for achievement-level descriptors. *Educational Assessment*, 18(2), 99-121.

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.

Sizer, T & Sizer, N. (2000). *The students are watching: Schools and the moral contract*. Beacon Press.

US Department of Education (2018). *A state's guide to the US Department of Education's assessment peer review process*.

Walkup, J.R. (2014, December 24). Bad DOK Chart Sabotages Understanding of Depth of Knowledge. *Cognitive Rigor to the Core!* http-//cognitiverigor.blogspot.com/2014/04/by-john-r.html

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science* (CCSSO and NISE Research Monograph No. 6). Madison/ University of Wisconsin–Madison, Wisconsin Center for Educational Research.

Webb, N. L. (1999). *Alignment of Science and Mathematics Standards and Assessments in Four States. (Research Monograph No. 18)*. National Institute for Science Education University of Wisconsin-Madison. Council of Chief State School Officers.

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

Webb, N. L. (2006). Identifying content for student achievement tests. *Handbook of Test Development*, 155-180.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied measurement in education*, 20(1), 7-25.

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied Measurement in Education*, 20, 7–25

Wine, M. & Hoffman A. (2022). RTD Approach to Using Norman Webb's Depth of Knowledge (DOK) Typology of Cognitive Complexity [White Paper]. AleDev Research and Consulting.

Wyse, A. E., & Viger, S. G. (2011). How item writers understand depth of knowledge. *Educational Assessment*, 16(4), 185-206.

Zheng, A. Y., Lawhorn, J. K., Lumley, T., & Freeman, S. (2008). Application of Bloom's taxonomy debunks the" MCAT myth". *Science*, 319(5862), 414-415.