Reinforcing Webb's Depth of Knowledge:
Laterally Extending DOK by Acknowledging Proficiency's Impact on Cognitive Demand

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 American Education Research Association in Chicago, Illinois.

Abstract

Norman Webb's (2002) Depth of Knowledge is the most commonly used typology of cognitive complexity across the assessment industry. Unfortunately, it is widely misapplied, inflating the reported DOK levels of test items – a problem that is commonly understood but rarely publicly acknowledged. We lay out common misunderstandings/misapplications of DOK and offer a more robust system for classifying items by cognitive complexity by returning to Webb's original conceptions and descriptions of DOK. This method of classification recognizes the impact of increased proficiency on reducing cognitive load, thus the fact that cognitive complexity is as much a product of test taker proficiency as the tasks given to test takers. We show the feasibility of this approach with a simple interrater reliability study.

Validity is the alpha and omega. It is, "the most fundamental consideration in developing tests and evaluating tests" (American Educational Research Association et al, 2014, p 14; 1999, p. 9). It is, "the most important consideration in test evaluation" (AERA et al, 1985, p. 9). The 2014 Standards also say, "The match of test content to the targeted domain in terms of cognitive complexity…[is] also [an] important consideration" (p. 26), explaining that cognitive complexity is one aspect of validity.

Furthermore, the United States Department of Education (USED) requires evidence of cognitive complexity to be presented as part of its *Assessment Peer Review Process*. The very first bullet point in the *Validity* section of its *A State's Guide to the U.S. Department of Education's Assessment Peer Review Process* says, "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). USED is clear that examples of the kinds of evidence that states must submit about their tests include aspects of cognitive complexity.

> Documentation of procedures to review items for alignment to: (1) academic content standards, intended levels of cognitive complexity, intended levels of difficulty, construct-irrelevant variance, and consistency with item specifications, such as documentation of content and bias reviews by an external review committee. (p. 39)

In fact, this *Guide* uses the phrase "depth and breadth" 23 times, and mentions "cognitive complexity" nearly 20 times.

Cognitive complexity is an important part of validity, alignment and alignment studies because alignment has too often been accepted when items and tests have just somewhat been in the neighborhood of state standards for learning goals. Webb (1997) went further, "For most of the states, frameworks and assessments were judged to be aligned if goals and learning objectives were considered in some way in the design or selection of the assessment instruments." Reviewing items for their alignment – in terms of cognitive complexity – with their respective standards is a way to hold claims of alignment's feet to the fire.

In Webb's "revolutionary article on alignment" (Forte, 2017, p. 6), he explained how important alignment – and therefore validity – was at the end of the 20th century.

> 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 (1997, p. 1).

Alignment and cognitive complexity have only become more important since 2002's *No Child Left Behind Act* (Flowers, Wakeman & Browder, 2009).

Understanding and operationalizing validity and cognitive complexity in test development and test evaluation is, therefore, incredibly important. *The Standards* and USED both stress that cognitive complexity is a part of validity. And yet, conceptions of cognitive complexity are marginalized and have largely become a resented hoop to jump through. The most commonly used typology for cognitive complexity, Depth of Knowledge (DOK), is misunderstood and misapplied and therefore offers little to actually help ensure validity of items, tests and/or the inferences made upon them.

In this largely conceptual paper, we review misunderstandings of Webb's DOK (2002), examine Webb's original Depth of Knowledge construct (wDOK) to highlight its central thrust and reinforce DOK by offering revised Depth of Knowledge (rDOK) structure that laterally extends (Bechard, Karvonen, & Erickson, 2021) that central idea consistently through the entire DOK structure. We explain how this rDOK can be applied both to recognizing the cognitive complexity of standards and to recognizing the cognitive complexity of test items.

Literature Review

There have been many typologies to classify sorts of cognition, and comparisons between them go back at least 50 years. Gall (1970) compared eight different systems, a generation before Webb introduced his Depth of Knowledge (1999, 2002, 2005). By 1985, Bloom's Taxonomy was already dominant. "It would be difficult to find a more influential work in education today than *[Bloom's] Taxonomy of Educational Objectives*" (Paul, 1985, p. 36). Despite the fact that Bloom's Taxonomy was originally intended for use in assessment (Bloom, 1956), Webb's DOK has been the dominant classification system in large scale assessment for over 10 years.

Bloom's Taxonomy

Bloom and his colleagues ambitiously sought to develop a taxonomy of all educational objectives, with volume I representing *the Cognitive Domain* (Bloom, 1956). The Taxonomy (i.e., Bloom's Taxonomy) was originally aimed at university examiners for use in designing and developing their assessments. However, it quickly moved beyond that limited context. "What made the taxonomy easy to implement, at least in some form, was crafted intentionally. Bloom and colleagues, though they envisioned their work used chiefly among university examiners, sought to establish something that could be used in any context" (Schneider 2014, p. 38).

Anyone who has worked in schools, districts or in teacher education know the ubiquity of Bloom's Taxonomy. By 1985, it had already been pivotal for decades.

> A generation of teachers have now come of age not only familiar with and 'acceptant' of the general categories of the Taxonomy, but also persuaded that the Taxonomy's identified higher-order skills of analysis, synthesis, and evaluation are essential to education all levels. For these teachers, critical thinking is essential because higher- order skills are essential. To learn how to think critically, in this view, is to learn how to ask and answer question of analysis, synthesis, and evaluation. (Paul, p. 36)

There is no question that our own understandings of the goals of schooling have been strongly shaped by The Taxonomy – likely even before we were aware of its existence – by teachers who took it quite seriously.

Two Taxonomies

Bloom's Taxonomy was altered in 2001 (Anderson et al, 2001; Krathwohl, 2002) by a team led by one of Bloom's original colleagues in the developing The Taxonomy. Though the Revised Bloom's Taxonomy (RBT) is two-dimensional rather than unidimensional and alters the order of the top two levels, The Taxonomy remains immensely popular among curriculum developers, school district personal and teachers. Some continue to use the 1956 original and some use RBT (Schneider, 2014).

Table 1

Two Bloom's Taxonomies

	Bloom's Taxonomy (1956)	RBT (2001)
Levels (i.e., Cognitive Process Dimension)	1. Knowledge 2. Comprehension 3. Application 4. Analysis 5. Synthesis 6. Evaluation	1. Remember 2. Understand 3. Apply 4. Analyze 5. Evaluate 6. Create
Second Dimension (i.e., Knowledge Dimension)	None	A. Factual B. Conceptual C. Procedural D. Metacognitive

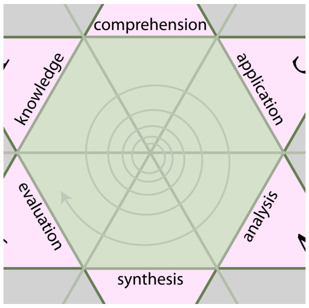
The Taxonomy's Central Problem

The main problem with Bloom's Taxonomy is its basic validity. The team that developed the taxonomy knew, "A taxonomy must be so constructed that the order of the terms must correspond to some 'real' order among the phenomena represented by the terms" (Bloom, 1956, p. 17). Despite the fact that that the group "spent considerable time" trying to justify the ordering of the levels," they were unable to find "a sound basis" for their work. They recognized that their levels were not nearly as distinct as so many would take them to be. "No entirely clear lines can be drawn between analysis and comprehension at one end or between analysis and evaluation at the other" (p. 144) – a range that covers five of The Taxonomy's six levels!

Researchers have tried to confirm the hierarchy described by The Taxonomy and at least since 1970 have generally found it not to be (Schneider, 2014) well founded. Even the original team saw this problem – though their reasoning was problematic. "If this is the real order from simple to complex, it should be related to an order of difficulty…Our evidence on this is not entirely satisfactory" (Bloom, 1956, p. 17). This issue is sometimes subtly acknowledged. As in the middle of *Bloom's Rose* (Kennedy, 2008) contains a spiral that indicates there is no beginning or end.

Figure 1

Bloom's Rose



Although Bloom's Taxonomy has long been treated as progression through which to lead students (Schneider, 2014), it simply is *not* the hierarchy that people take it to be (Clinton &

Hattie, 2021). Nonetheless, it is used, extended and adapted (Flowers et al, 2009; Webb, 2006) and remains popular around the world (Crowe, Dirk & Wenderoth, 2008; Zeng, Lawhorn, Lumley & Freeman, 2008).

Despite its popularity among educators, assessment developers need a better typology for cognitive complexity.

Webb's Depth of Knowledge (wDOK)

While Bloom's Taxonomy may be the most well-known view of cognitive complexity – especially among teachers – Norman Webb's Depth of Knowledge is the most widely used in the field of large scale assessment (Flowers, Wakeman & Browder, 2009; Wyse & Vigor, 2011). It is so commonly assumed that it is widely used without even attribution (e.g., Achieve, 2006). It so commonly assumed to be synonymous with *cognitive complexity* that a scholar cited in this article assailed USED for requiring it in Peer Review at the Council of Chief State School Officer's 2022 National Council on Student Assessment– in spite of the fact that it is not so much as *mentioned* (not even as example) anywhere in USED's (2018) peer review documentation. USED does *not* require its use, as explained by a USED representative at that session. But USED might as well require it, considering its ubiquity. Of course, the scholar's mistake was quite understandable, as CCSSO itself has long stated a preference for Webb's DOK (CCSSO, 2009), and it "underscores much of the US K-12 Common Core [assessments]" (Clinton & Hattie, 2021, p. 3). Furthermore, item writers seem to prefer DOK, though that may be in part due to their familiarity with it (Schneider et al, 2013).

Unfortunately, like so many tools and ideas that dominate the practice of content development, there is far less literature *about* and/or *examining* wDOK than the tools of psychometrics. Instead, the actual use and understanding of DOK is taken for granted in the literature that acknowledges its existence. That is to say, Webb's Depth of Knowledge is often cited, but rarely the subject of study, itself.

Short History of wDOK

Webb's own Depth of Knowledge typology is most clearly laid out in his 2002 *Depth-of-Knowledge Levels for Four Content Areas,* though it has origins going further back than that. His revolutionary 1997 article outlined his thinking about validity, including cognitive complexity. He observed that in the mid-1990's, 20% of states did not even have mathematics standards and those that did varied quite a bit in scope and/or specificity – and yet nearly all states had math assessments! Poor alignment practices were the rule. His new proposed approach included six different criteria to consider on alignment of assessment to standards, of which Depth of Knowledge was just one.

- Balance of Representation
- Categorical Concurrence
- Depth of Knowledge Consistency
- Dispositional Consonance
- Range of Knowledge Correspondence
- Structure of Knowledge Comparability

Eventually, Webb (2007) settled on just four criteria (i.e., Categorical Concurrence, Depth of Knowledge Consistency, Range of Knowledge Correspondence, Balance of Representation).

Webb's 1997 description of Depth of Knowledge was a bit different than what came later. Even then, though he was clear that Range of Knowledge was not adequate, in and of itself.

> The depth of knowledge required by an expectation or in an assessment is related to the number of connections of concepts and ideas a student needs to make in order to produce a response, the level of reasoning, and the use of other self-monitoring processes. In addition, other factors influence the cognitive demands of performance including the social or contextual requirements, the variety of representations students are expected to use (written, verbal, pictorial, and variations within each), and requirements for transfer and generalization to new situations. (p. 15)

Webb's goal, even then, was recognizing when assessments "require students to demonstrate only a small part of what the standards intend" (p. 15) and use cognitive complexity to acknowledge such limitations. He also saw novelty (i.e., "new situations") as contributor to this Depth of Knowledge.

By 1999, Webb had developed wDOK further, as least for math and science. This early formulation (p. 3) contained the now famous four wDOK levels, with brief science- & math-oriented descriptions totaling just 78 words

Table 2
Webb's Original Four DOK Levels

Levels		Description
1	Recall	Recall of a fact, information, or procedure.
2	Skill/ Concept	Use of information, conceptual knowledge, procedures, two or more steps, etc.
3	Strategic Thinking	Requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer; generally takes less than 10 minutes to do.
4	Extended Thinking	Requires an investigation; time to think and process multiple conditions of the problem or task; and more than 10 minutes to do non-routine manipulations.

Webb has always been explicitly clear that his typology was a tool to be used when examining "assessment systems—all forms of gathering information about students' learning" (1997, p. 2). That is, he did not believe that large scale standardized assessment *could* address the full range of learning expectations listed in standards and/or expected by educators. For all its problems, even his 1999 version made clear that Level 4 KSAs (i.e., knowledge, skills and abilities) or tasks would not be found on large scale assessments. He expected *other* components of the greater system of assessments (e.g., including classroom assessments of various sorts) to produce evidence of those KSAs.

By 2002, Webb expanded his definition of the four levels. Mathematics' explanation totals over 700 words, with math-specific examples and explanations. Science took over 1100 words, and he also included this kind of fuller explications for Social Studies, for Reading and for Writing. These are not simply expanded definitions, but actually differ in significant ways from the 1999 presentation. For example, by 2002 Webb has abandoned defining how much time *Extended Thinking* entails and no measure of time is included in any way. The idea of *steps* (i.e., originally part of the level 1 & 2 definitions) remains – though only for math and science; more attention is given to the “more demanding reasoning” (p. 6) that he was trying to get at, rather than simple step-counting.

Recognizing DOK

Since 2002, more summaries and recapitulations of wDOK have been offered by innumerous parties, and almost none of them bear serious mention. Two, however, do.

First, the DOK Wheel is a common explanation for wDOK, and no one knows where it came from. Walkup (2014) lays out everything anyone needs to know about the DOK wheel.

> In December 2013, I emailed Norman Webb for his views on the DOK wheel chart. Apparently, he is asked quite regularly for permission to use the chart in the mistaken belief that he created it. In his response, Webb explicitly stated that he considers the wheel chart misleading and has always discouraged its use.

We had a similar email interaction with Webb in 2011. Among other things, he wrote, “Although it references my work and uses DOK, the wheel itself misrepresents my work by only focusing on the verbs.” We have seen it falsely credited to him and the Wisconsin Center of Educational Research, but it should be ignored.

Second, Karin Hess and others have developed their *Cognitive Rigor Matrix* (Hess et al., 2009), which combines the widely venerated Bloom’s Taxonomy and wDOK. However, this approach builds upon wDOK, and is not an explanation of DOK. Hess and others have developed this approach for each of the content areas, but their work is different than wDOK, itself.

Our own summary of wDOK (Wine & Hoffman, 2022) is shown below. Webb offers subtleties and details that are particular to the different content areas, but Table 3 contains the essential consistent structure (at 134 words).

Table 3

Webb’s Depth of Knowledge (2002)

Level	Name	Description
wDOK 1	Recall	Recitation or recognition of facts, basic reading comprehension, rote use of algorithms or procedures. Includes recitation or identification of explanations learned previously.
wDOK 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.
wDOK 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.
wDOK 4	Extended Thinking	Thinking that is extended across multiple contexts or concerns in ways that connects those contexts or concerns. Arriving at generalizations based upon a range of information or ideas. Analysis that includes multiple factors or issues, and accounts for those issues in the final product.

Obviously, this Wine and Hoffman summary of wDOK is quite credible, accurate and wise. Webb himself offered a 690 word summary of the four DOK level in 2007, a summary that is consistent with his 2002 explanations, though without those subject-specific examples and contextualized explanations – and it lacks a nice summary table. We leave it to the reader to decide which might be superior.

Generally, any time cognitive complexity is presented with four level and mentions *recall* and *strategic thinking* as different levels, it is an attempt to make use of Webb's Depth of Knowledge, whether it is cited or not. Any other approach, be it with different level or more dimensions or one that focused primarily of key words (i.e., as opposed to describing cognition more clearly), it is not Webb's DOK, regardless of how it is presented.

Misunderstandings of wDOK

Though it is underexplored in the literature, misconceptions of Webb's Depth of Knowledge abound, as is clear to anyone who has worked with content alignment reviews (e.g., with teachers on content validity committees/panels). Some assessment professionals share in these misconceptions and even perpetuate them. Unfortunately, only Wyse and Vigor (2011) have formally studied and presented these misunderstandings. They looked at how a group of item writers understood DOK, after they received training on DOK (i.e., including differentiating DOK from Bloom's). This group included both those with previous experience with DOK and those who were new to it.

Wyse and Vigor identify a number of "misconceptions" (p. 185), a list that shows some misunderstanding they themselves had(see below).

- Conflating DOK with task difficulty – the most persistent misunderstanding.
- Changing the testing population can alter the DOK level of an item.
- "…relat[ing] that thinking and DOK to the student and not the item or content standard" (p. 194).
- DOK is determined by "the level of cognitive ability…that an item requires" (p. 194).
- DOK levels sometimes overlap.
- High level of concurrence between Bloom's Taxonomy and wDOK.
- Length of student response – perhaps in time and perhaps in number of words – determines DOK level.
- There are different definitions for DOK, depending on the level of the assessment.
- The most important thing is matching verbs between standards and items.

Webb himself (2007) listed two particular misconceptions about using his DOK when considering alignment.

- Pre-determining DOK distribution in advance or for a grade level, without actually examining the official expectations (e.g., state standards) carefully.
- There is a single "decision rule" (p. 19) regarding the appropriate share items at the aligned standard's DOK level, regardless of the purpose of the test.

Unfortunately, Wyse and Vigor's exploration shows that even scholars themselves sometimes misunderstand Webb's Depth of Knowledge. For example, they present and misinterpret a quote from Webb (1997) and claim,

> The fact that an individual student used or may use a simpler or more complex approach to solve a problem does not change the DOK level of the item for that

> student because the DOK level refers to the baseline level of cognitive processing required to provide a correct response. (p. 188)

This view that the relevant "cognitive processing that an item…requires an examinee to engage in" (p. 188) can be pinpointed to a unique cognitive path simply does not hold up to rigorous examination (Hoffman, 2022). Similarly, Wyse and Vigor prioritize the number of steps an item requires, without unpacking the question of what counts as a step and for whom. They wave away the empirical fact of frequent of disagreements about the DOK level of an item or a standard, insisting that there *no room* for seeing overlap between the DOK levels. They claim that "the DOK level of a test item does not inherently change if the examinee population changes" (p. 189). These beliefs do not stand up to careful consideration and analysis, even if they may seem true on the surface.

Automaticity

The idea of automaticity in cognitive processing is widespread throughout cognitive psychology, though it has a number of names. Singer (2002) points out that even when applied in sports psychology, it is described in many ways.

> …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)

Regardless of its name, what we refer to as *automaticity* is the idea that the application of KSAs and/or ideas can be done consciously and with deliberation, or they can be done automatically and with less consciousness – what Price & Driscoll (1997) called "mindful" and "automatic" (p. 473) thinking. While formal *Automaticity Theory* (Moors & De Houwer, 2006; Stanovich, 1990) generally views this as two distinct forms of cognition, many see this deliberation-automaticity distinction as more of a continuum (Logan, 1985).

This idea is not limited to Automaticity Theory. It appears in 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. Evans and Stanovich (2013) explain how all of these areas and more come under the broader umbrella of *dual process theories*. They, too, see these two modes of cognition as "two poles of a continuum of processing styles" (p. 226). They make clear that the more deliberative mode (i.e., Type 2) is *not* necessarily superior.

Schema Theory explains that this kind of shift in cognition is the result of "well-structured schemata that are automatically activated during problem solving" (Moreno & Park, 2010, p. 12). ACT* theory (Anderson, 1990, 1992) highlights that this this greater automaticity results from the consolidation of many steps into a single cognitive step. Deliberation calls on "greater cognitive capacity usage" and "greater mental effort expenditures" (Salomon and Perkins, 1989, p. 125). Increased proficiency leads to less cognitive work to produce a similar result (Anderson, 1982; Fitts & Posner, 1967) because with that proficiency, "Learners no longer need to concentrate" (Ericsson, 2014, p. 82). Each of the dual processing theories explains that the shift to more automatic cognition results in lower cognitive burdens, reduced cognitive demands and lower cognitive complexity.

The shift from deliberation to automaticity results in "fast, effortless (from a standpoint of allocation of cognitive resources), and unitized (or proceduralized)" cognition (Ackerman, 1987, p.

4). Salomon and Perkins (1989) showed that this is not limited to simple procedures, that even complex skills "eventually become routinized" (p. 130). This does not even require increases in proficiency, as experience or practice can increase automaticity without necessarily increasing 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).

We explore the (obvious) application of dual process theories to DOK below.

Goals of and Constraints on this Project

The basic goal of this revised Depth of Knowledge project (rDOK) has always been to develop an approach to classifying cognitive complexity that could replace the problematic use Depth of Knowledge system that dominates large scale assessment and peer review. The various mistakes and compromises that have been made in using DOK has meant that industry practices has drifted from Webb's original wDOK. We use the term *iDOK* to refer to standard practices under the label of DOK. They are based in wDOK and are *claimed* to be Webb's Depth of Knowledge, but they have drifted to something a bit different.

The requirements of this project have been as much defined by the constraints from such use as by that basic goal. Yes, it is focused on offering a more rationalized and consistent take on cognitive complexity that respects the goals of Webb's original work. However, we have not been free to develop simply *any* sort of new typology of cognitive complexity that fits our own sense of the most important potential meaning of "cognitive complexity." That is, we have *not* tried to define or operationalize anything like an ideal typology that recognizes all the various dimensions of cognitive complexity that may be important.

Rather, we were trying to develop a useful and thoughtful drop-in replacement for iDOK that helps to identify discrepancies between standards and items or tests. This is why this typology is based so deeply on wDOK, itself. rDOK is revision of wDOK, as RBT is a revision of Bloom's taxonomy (Anderson, 2001; Krathwohl, 2002). Moreover, rDOK has to fulfill USED's (2018) requirements for its peer review process. Last, rDOK has to confront the all-too-common misunderstandings of wDOK.

rDOK Cannot Reduce to Item Difficulty

The most common error in categorizing cognitive complexity is mistaking difficulty for cognitive complexity. Quite simply, everyone does it – at least from time to time. In our own work, we sometimes have to stop and ask ourselves whether we just conflated difficulty and cognitive complexity.

In fact, item difficulty already exists an empirical measure. Cognitive complexity *must* mean something different from item difficulty. That is not so say that these two constructs must be entirely orthogonal, but cognitive complexity must be largely distinct from item difficulty.

rDOK Cannot Reduce to Grade Level

There is no question students are expected to engage in more complex work as they progress through the grades. Nonetheless, even at the highest levels, there is some low cognitive complexity work. Advanced science and math include memorization of key formulas and facts. Social studies also include memorization. Vocabulary demands of texts in ELA are always increasing, with fluent reading and writing depending on memorized knowledge of vocabulary – even to the point of internalized understanding that contributes to automaticity.

Furthermore, cognitive complexity ceases to be a generally useful construct if high cognitive complexity tasks and standards are only found in higher grades. Educators believe that higher order thinking skills are taught at lower grades, and we do, too. There are tasks that are quite different than the fluid recall of memorized facts and formulas that epitomize low cognitive complexity tasks at *every* grade level.

rDOK Cannot Reduce to Test Taker Proficiency

It is tempting to view higher cognitive complexity as the domain of the most proficient and skilled students and test takers. That is, the approaches taken by the most advanced students are those marked by the greatest cognitive complexity. However, we see this view contradicted from the beginning in wDOK. Furthermore, we believe that the various approaches in psychology to thinking about proficiency (see above) also undermine this idea. Research has long made clear that cognitive demand and cognitive complexity of tasks *lessens* as proficiency and experience increases.

This is perhaps the least intuitive (and most radical) aspect of this project. The most advanced and successful students, the highest scoring the test takers, and/or the kinds of people who might read *this* paper (i.e., you!) are *not* necessarily engaging in tasks with the most complex cognition. Their very proficiencies enable them to successfully complete those tasks with *less* mental effort and cognitive complexity than the struggling and/or diligent test taker who more carefully figured out what to do, checked their work and labored find a cognitive path to a successful response.

rDOK Must Fit USED's Peer Review Requirement

Regardless of our own or anyone else's feelings about the utility, value or return on considering cognitive complexity in assessment development, the United States Department of Education (2018) requires it. Hence, rDOK cannot merely be useful for examining, exploring and understanding items and assessments, and must fit the expectations and workflows of USED's peer review. While USED does not require a hierarchical and unidimensional scale, in practice this is what assessment developers and alignment studies have grown to expect.

One the great attractions of wDOK is the simplicity of it scale. That is, it offers four levels of a single dimension (at least in reporting), and one of them is not even applicable to on-demand large-scale standardized assessment. While cognitive complexity is clearly a broad topic that can be recognized in multiple ways, a typology for cognitive complexity *meant for assessment development work* will not be adopted if it is significantly more cumbersome than wDOK. The unceasing demands of the content development cycle makes it difficult for CDPs to find the time to adopt and apply a multi-dimensional scale that takes significantly longer to use than iDOK.

In this context of USED's peer review process, a cognitive complexity typology is not merely meant to describe tasks. Rather, it is used evaluate whether items and/or assessments match the cognitive complexity of some goal or reference. It must provide a yardstick that test developers may use both to define some requirement(s) for cognitive complexity (i.e. as indicated in standards) and to recognize whether that requirement(s) has been met (i.e., in the cognition prompted by tasks and items). A multi-dimensional construct inevitably will provide contradictory answers – enabling its own marginalization from being a driver for increased validity. (This has been one problem with wDOK, as Webb's explanations at times present multiple determiners of cognitive complexity.)

rDOK Should Not Incentivize Fraudulent Inflation of Cognitive Complexity

While the wDOK structure does not itself incentivize fraudulent inflation of reported cognitive complexity for items, it is hard to deny that it has been used to do just that. External policy preferences for using large scale standardized assessments as the singular measure of student achievement and/or learning have pressured assessment developers to support claims that these assessments measure things they cannot, as Webb pointed out in 1997. Thus, a structure originally designed to highlight the limits of this one class of assessment has been compromised to support a policy approach that we lack the technology (i.e., large scale assessment capabilities) or resources (i.e., time, money, expertise, testing time) to implement.

wDOK offers too many distinct ideas about contributors to cognitive complexity to be used rigorously and reliability amid such pressures. For example, the wDOK reading construct incudes length, purpose, amount of organizational structure, form, use of external tools and more. It just is too easy to find some element of the assessment task that might be in the neighborhood of some aspect of Level 2 or Level 3 and thus claim that the item is at the level. Obviously, Webb intended a good faith weighing and recognizing of the various aspects of his DOK construct, but amid financial, policy, time and other organizational pressures, it is too easy to resort to a vaguely colorable – though in fact quite poor – argument.

Ideally, a practical typology for cognitive complexity would be simpler, giving fewer potential arguments that an item has reached some higher level. While it may depend on professional judgment, it would include fewer ambiguities to leverage for inflated claims.

rDOK Must Correct the Common Misunderstanding of wDOK

This project was focused on addressing six common misconceptions about Webb's Depth of Knowledge.

First, cognitive complexity is *not* item difficulty. Item difficulty is empirically measured. There are easy and difficult problems of low cognitive complexity. Naming the current President of the United States is low cognitive complexity and low difficulty, whereas naming the 21st President of the United State is of the low cognitive complexity but is much more difficult. Naming the capital of the 41st state is *still* of low cognitive complexity, and is even *more* difficult. Schneider et al (2013) showed exactly this: that a range of difficulty certainly exists at the lowest DOK levels, and at higher levels as well.

Second, cognitive complexity is *not* determined by the number of steps in a solution path. More proficient students are able to combine steps, making more progress with fewer steps. In fact, this is often what it means to be more proficient (Anderson, 1990, 1992). When Webb wrote of the number of steps in a problem, he meant distinct processes that must be put together. For beginning students, adding a pair of 4-digit numbers requires a number of steps, but for more advanced students no one considers that to be more than single step – if that. Though Webb *did* mention this concept of steps and being aware of them, DOK level is not really a function of the number of steps a problem requires. To be sure, problems for which the test taker/student applies a single known procedure – a rote memorized cognitive path – *is* wDOK 1 and figuring out how to put together multiple such procedures rises above wDOK 1. However, counting the steps is no more useful than counting the tines on a fork to figure out its use (Marshall, 1990).

Third (and similarly), the scale of a task does *not* determine cognitive complexity. Webb (2002) was clear, "if the required work is only repetitive" then the cognitive complexity is not increased (p. 2, and (ironically), pp. 4, 7, 8). More of the same is merely more of the same. A larger canvas may allow for work of greater complexity, but does not require it. Furthermore, quite a lot of complexity can fit on a small canvas. Wyse and Vigor (2011) pointed this out as well, "DOK is related to cognitive processing and thinking, and it does not change depending on context (e.g., a recall item is a recall item)" (p. 196).

Fourth, the time required by a task is *not* a determiner of cognitive complexity. Webb (2002) addresses this point explicitly. He explains that his Level 4 (i.e., *Extended Thinking*) requires significant amount of time, but it is the nature of the thinking that matters and not the amount of time. "The extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher-order thinking" (p.2, and repeated).

Fifth, the context in which a skill is applied is *not* determinative of its cognitive complexity, just as the scale does not determine the cognitive complexity. Correcting spelling of a word as part of a long essay is no more cognitively complex than spelling it in isolation – even if the larger task is more complex. Solving a triangle is no more complex for being part of large surveying project than as part of a single homework problem. The application of a KSA (or group of KSAs) is *not* made more complex for it being embedded as part of some larger whole. If the unit of analysis is a single item, then it is *the application of the targeted cognition* part of cognitive path prompted by the item that determines its cognitive complexity.

Sixth, the verbs used in a standard or item are *not* sufficient to determine cognitive complexity. Webb knows it, Brabrand and Dahl (2009) know it, everyone should know it. For example, there are cognitively simple inferences and there cognitively complex inferences; *infer* does not have single cognitive complexity. *Solve* can be applied to cognitively simple tasks or to tasks of enormous cognitive complexity. *Name, write, add, determine, change, explain* and *embed* (i.e., just some of the verbs used in this section) can all be used to prompt cognitive simple tasks or to prompt cognitively more complex tasks.

Seventh, there is *not* necessarily a singular and definitive path through an item to a solution. "Students may use multiple strategies, take different paths, or think in many unique ways when coming up with a response to a test item. Moreover, these strategies may be more or less complex routes to producing a response" (Wyse & Vigor, 2011, p. 188). Some try to pretend that this is not the case, but they merely are choosing some arbitrarily preferred path (Hoffman, 2022). Taking notions of *Fairness* seriously requires acknowledging that every aspect of alignment and validity – including cognitive complexity classification – is conditional on which of multiple student paths are considered.

rDOK Basics

Just as RBT (Revised Bloom's Taxonomy; Anderson et al, 2001, Krathwohl, 2002) was developed to address shortcoming with Bloom's original Taxonomy (Bloom, 1956), we have developed rDOK (revised Depth of Knowledge) to address particular issues with the use and application of Webb's Depth of Knowledge typology of cognitive complexity (2002). This lateral extension (Bechard, Karvonen, & Erickson, 2021) of wDOK aims to simplify some of the decision making about classification of cognitive complexity by recognizing the central thrust (i.e., the dominant dimension) of wDOK and amplifying its place in in the typology. It also recognizes that "DOK is about the process invoked by the item" (Wyse & Vigor, 2011, p. 197) and follow through on what that implies.

rDOK is not dramatically different then wDOK, at least on paper. Rather, it is a refined approach to using essentially the same Depth of Knowledge tool. As such, it is at least as much an effort to address problem with iDOK (i.e., typical contemporary use of Depth of Knowledge in a standardized testing context) as it is an effort to alter wDOK. That is, it is built incredibly closely on Webb's Depth of Knowledge framework, and its limited alterations of wDOK only yield significant differences because they are focused on the use and application of DOK in assessment development – recognizing the differences in application of a single construct across *multiple* different contexts/content areas (Wine & Hoffman, 2023).

Recognizing Automaticity as the Central Thrust of DOK

rDOK is the application of dual process theory – what we refer to as *deliberation vs. automaticity* – to Webb's DOK structure. This is on obvious approach, as his Level 1 has always been called *Recall* (1999, 2002, 2005) and his Levels 3 & 4 have been called *Strategic Thinking* and *Extended Thinking*, respectively. Webb was always focusing on more automatically applied cognition vs. more deliberate cognition. He differentiated Level 1 from Level 2 by explaining,

> 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)

The "rote" nature of Level 1 cognition is clearly automaticity! While use of a "well-known algorithm" is not quite as automatic as a "rote response," it is clearly more automatic than the decision-making that Webb says typifies Level 2 cognition.

Unfortunately, Webb misnamed his Level 2. He originally (1999) called it *Skill/Concept*, but then varied more. In 2002, he did not name Level 2 for reading or for writing, continued with *Skill/Concept* for both mathematics and science, but called it *Basic Reasoning* for social studies. Many practitioners have simply missed the sentenced quoted in the previous paragraph, despite his numerous repetitions of it, verbatim (1999, 2002, 2005, 2006, 2007[1]). When applying a skill or concept is done in the form of "perform[ing] a clearly defined series of steps," it is *not* wDOK Level 2 cognition. Many have seen that his Level 1 definition includes, "a one-step, well-defined, and straight algorithmic procedure" (2002, p. 3) and extrapolated from that and the name *Skill/Concept* that multiple steps automatically means cognition beyond wDOK's Level 1 – *despite the fact that Webb said explicitly otherwise on the very next page*. This confusion has prevented some from seeing that Webb's DOK focuses intensely on this *deliberation vs. automaticity* issue as its central thrust.

Hence, this application of the deliberation vs. automaticity of dual process theories as the primary lens through which rDOK determination are made is *not* a radical or new intervention into DOK. It is a quite conservative approach that simplifies the meaning of cognitive complexity, befitting its use in large scale assessment development.

rDOK as the Complexity of Test Takers' Cognition

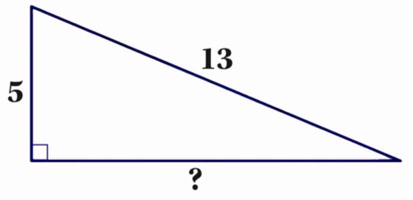
The real notable shift in rDOK is that rDOK takes seriously that DOK is a measure of *cognitive* complexity, rather than task complexity. That is, the complexity in question when

[1] Google Scholar actually lists nine works with Webb as an author that contain this exact sentence, in addition to however many times he used the sentence that would not be cataloged by Google Scholar. (It also lists four works by others that quote the sentence in its entirety.)

classifying items is the complexity of the cognitive paths taken *by test takers* (i.e., particularly of the application of the targeted cognition). It is not the complexity of the paths most desired by educators and content development professionals. It is not the paths that make best use of the supposedly aligned standards or KSAs. It is not the complexity that adults imagine that they might have taken when they were that age[2]. rDOK determinations should be made about the actual plausible cognitive paths of test takers (i.e., across the range of typical test takers).

Consider the following item:

What is the length of the third side of the following triangle?



A) 5
B) 7
C) 12
D) 25

This item presents multiple potential solution paths, including:

- Application of the Pythagorean Formula, taking the square root of 144
- Backsolving though the Pythagorean Formula without ever taking a square root
- Visually estimating
- Reasoning based upon the fact the length of any two sides of a triangle must be greater than the length of the third side.

Each of these are valid paths that make use of knowledge and skills learned in math class and lead to a successful response. However, while the first two approaches are DOK Level 1 (i.e., in this case, backsolving does significantly reduce the complexity of the solution path), the last approach is Level 2.

The RTD[3] approach to content development is *always* based in the cognitive paths of test takers and acknowledges that different test takers can have quite different cognitive paths – thus the challenge of writing items that elicit evidence of the targeted cognition *for the range of typical test takers*. With this item, students who do not know the Pythagorean Formula will likely take a different cognitive path to a solution that those who do. Cognitive complexity is *not* a feature of items, but a feature of the *interaction* between test takers and items. As Hoffman (2022) exhaustively demonstrates, there is no singular definitive cognitive path to a solution to be preferred over other plausible paths that test takers might take and claiming otherwise is a self-

[2] Considering one's own cognitive path can be an important step in recognizing the cognitive paths that test takers might take, but this is far from adequate to uncover those paths (Hoffman & Wine, 2021).

[3] The RTD (Rigorous Test Development) project is our effort 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 large scale test development, today.

centered act that is entirely at odds with the recognition of the variation among students that Fairness requires.

This means that rDOK classification decisions about items can result in multiple rDOK levels for a single item *when the item allows for multiple solution pathways,* which is *quite* often the case. Certainly, test takers who come to items with different levels of proficiency may solve items with different levels of automaticity. Note that this is *not* a foreign addition to Webb's Depth of Knowledge typology. wDOK points to the importance of "rote" applications and how "well-known" an algorithm is. Webb clearly did not mean that the solving for the third side of a right triangle is wDOK Level 1 and finding the surface area of a sphere is wDOK Level 2 just because the former is a well-known algorithm and the latter is not. His definitions have long pointed to differences in deliberation and decision-making *on the part of test takers*, not simply whether some field has figured a pre-defined solution path. Test takers and learners are often expected to carefully reason though what is automatic and virtually thoughtless for experts. Webb's DOK recognizes this distinction, as does rDOK.

Of course, applying these ideas in practice, as in classification projects, is quite the challenge. The problems with wDOK have mainly appeared when people have tried to apply in practice, as opposed to just when reading or hearing about it.

Context and Cognitive Complexity

When considering the cognitive complexity of an item or a task, one must be careful to focus on the part of the cognitive path that the item is meant to examine. That is, the cognitive complexity of *the application of the targeted cognition*. A larger complex task that only makes use of the targeted cognitive in a cognitively simpler manner should be classified at that simpler level. There is no amount of *other* cognitive complexity that can make up for a too cognitively simple application of the targeted cognition, as that would shortchange the aligned standard.

For example, hammering a nail may seem like a cognitively simple task. Certainly it can be. When framing a house, nailing 2x4's together is so cognitively simple that we have invented dumb tools that can do it (i.e., nail guns). On the other hand, hammering a nail into a wall to hang a picture is notably *more* complex, despite the fact that the larger context of building a house is *far* more complex than the larger task of hanging a picture. You do not want to hit the nail so hard as to damage the wall, you do not want to drive the nail in so far that you cannot get the wire to hang from it, and yet you want to get it done quickly. You have to get a feel for how hard the wall is and pay attention to reduce your force when getting closer to completion – while perhaps paying attention to the possibility of stud behind the wall that will call for *increased* force. All of that is separate and apart from decisions about where to drive the nail – be they aesthetic or mechanical. Just the hammering itself call for *more* deliberation and attention in the *simpler* context.

Allowing the complexity of the larger task to drive classification decisions instead of focusing on the application of the targeted cognition allows for construct-irrelevant cognitive complexity to take the place of *all* on-standard cognitive complexity across a whole test. It falsely inflates the reported cognitive complexity of both applications that are appropriately low rDOK and those that are inappropriately low. Thus, it can suggest that well aligned items are *not* aligned and that poorly aligned items *are* well aligned. Respect for the standards – which are democratically sanctioned though adoption by state legislatures – requires this kind of classification of *the application of the targeted cognition*, every time.

rDOK's Typology

Like wDOK, rDOK has four levels. In fact, rDOK is quite compatible with wDOK in most ways. Table 3's summary of wDOK (above) applies to rDOK as well. Table 4 (below) emphasizes the role of automaticity vs. deliberation in differentiating the different levels. Wine and Hoffman (2023) dives deeper into each of the content areas with explanations of how wDOK applies there, including examples for each content area.

Table 4

rDOK Typology

Level	Name	Description
rDOK 1	Recall	KSAs applied by rote, automatically and/or formulaically.
rDOK 2	Basic Decision Making	The use of tools, skills and/or concepts *with deliberation* and requiring decision-making in their application. Can include the thoughtful selection (i.e., a conscious decision) of the appropriate well-known conceptual tool to address a novel problem.
rDOK 3	Strategic Thinking	Prospective or retrospective reflection about novel work, including planning a multi-step path and/or explanations after-the-fact of the decision-making that went into a solution path.
rDOK 4	Extended Thinking	Multi-context or multi-factor novel work that is too complex to be held in one's head at once and is unlikely to be completed in a single sitting. *Note: Additional required time is not sufficient to make for wDOK 4 cognition; more of the same is merely more of the same.*

Novelty is an important factor in this kind of cognitive complexity because the more readily a problem is recognized, the more automatically a solution path is identified and the more rote the cognitive work is, the less complex the cognition is. Practice increases automaticity (Ericsson, 2014; Logan, 1985) and novelty is – by definition – unpracticed.

Hierarchy of Deliberation, *Not* a Hierarchy of Value

Lamentably, DOK has long been treated as a hierarchy of desirability, as though more cognitively complex cognition is *better* cognition. Certainly we have emphasized the importance of "critical and analytical thinking skills" in our teaching – *a lot!* But greater cognitive complexity is *not* necessary better cognition. Every content area prizes proficiency marked by rote recall and automatic applications. Sight reading words and command of math facts are not only desirable, but they are *critical* foundations for later learning and work. There is no height in any field at which that kind of easy access facts, formulas, declarative knowledge and/or techniques is not a vital part of the work.

The fact that various constraints on assessment content development has made it notably difficult to include rDOK 3 items – and often even rDOK 2 items – has combined with this mistaken undervaluing of highly proficient/low complexity cognition to incentivize the inflation of DOK classifications. There is nothing wrong with less complex cognition, *so long as the standard calls for it.* So long as the items are aligned to the standards – without the alternative paths to solutions that undermine the item's ability to elicit evidence of *the targeted cognition* – rDOK classification should be an easy check. Inflating the classification of standards and/or expecting large scale assessments to assess what they cannot only makes all of this more difficult – if not impossible.

Classification with rDOK's Four Levels

Each of the rDOK levels can be assigned to standards and to test taker cognition.

When applied to standards, rDOK is about the degree of automaticity vs deliberation with which their KSAs (knowledge, skills, abilities) are expected to be applied. Some standards clearly describe skills that call for high levels of automaticity, whereas others are about deliberation. Recognizing which is which – and when appropriate proficiency with a standard[4] may include a range of automaticity (and therefore a range of rDOK levels) – requires deep and *grade level specific* knowledge of content and expectations about student cognition.

When applied to items, rDOK is about the cognitive paths towards solutions that the items prompt in *test takers* – like all forms of alignment reviews should be. It is *not* about some favored or preferred cognitive path on the part of content development professionals or educators. While those involved in rDOK classification projects have to decide what range of students to consider, well-considered use of rDOK cannot avoid the fact that different students may follow paths of different cognitive complexity in response to the same item. But this is no different than acknowledging many items allow for multiple solution paths, some of which make use of the targeted cognition and some of which do not. As with standards, rDOK classification of items requires deep and *grade level specific* knowledge – in this case about student cognition.

rDOK 4 – Extended Thinking

Level 4 is the easiest to recognize and classify, regardless of whether it is standards or items in question.

Large scale standardized assessment, being built of a collection of on-demand tasks to completed in a single session, simply is not compatible with wDOK 4 cognition. Other forms of student work *can* be classified as wDOK 4, including much of what educators often call *projects.* This kind of work contrasts with the more regular – sometimes repetitious – work of lower levels of cognitive complexity. Such projects are generally multi-part, multi-stage efforts that bring a variety of knowledge, skills and/or concepts together.

Efforts of higher cognitive complexity are invariably made of up sub-tasks of lower cognitive complexity. For example, the work of large and complex research paper incudes the cognitively simple sight reading of words, the automatic spelling and (hopefully) cognitively effortless typing. This in no way prevents the larger task from including rDOK 4 cognitive complexity, as those lower cognitive complexity tasks are planned and put together with high levels of deliberation – hence the greater cognitive complexity.

There are few K-12 standards that call on wDOK 4 levels of cognition, but they are easy to recognize. None of the CCSS-M Content Standards call for rDOK 4[5], though many of the Standards

[4] Standards generally do not declare the level of proficiency that is expected. That is, whether it is sufficient for a test taker to muddle through to a successful response (as opposed to easily and confidently getting directly to a successful response) is an important question to be answered by test owners.

[5] The high school probability and statistics standards *could* be applied at an rDOK 4 level, but even these standards do not require it.

for Mathematical Practice *allow* for rDOK4 applications. Similarly, NGSS's Disciplinary Core Ideas do not call for rDOK 4 cognition, though many of the Science and Engineering Practices *could* be applied at that level. NGSS's Performance Expectations are often compatible with rDOK 4, though rarely *require* it. The kinds of larger projects that call for rDOK 4 are compatible of many of the CCSS-L standards, but – again – few require it. Though few standards require this kind of management and coordination of sustained, multi-part work, it has long been a part of the middle and upper grades instruction and learning.

rDOK 3 – Strategic Thinking

rDOK 3 is also generally easy to recognize – or at least its potential applicability can be. Tasks that require planning *before* the implementation work can begin are rDOK 3, in contrast to the kinds of tasks that students and test takers just dive into. *Planning* is one of the two hallmarks of rDOK 3. Alternatively, *reflection back on prior work* is the other hallmark of rDOK 3. These kinds of prospective and retrospective reflection are the rDOK3 level of deliberation. Both this sort of prospective and this sort of retrospective reflection involve a sort of metacognition that thinks *about* the work in addition to performing the work.

The planning of rDOK 3 cognition is more than simple selection a tool to solve a problem. Rather, it entails *developing* a plan for the multiple steps that the task requires – quite a bit more cognitively complex than grabbing a preexisting algorithm or solution. rDOK 3 requires looking down the road (i.e., the cognitive path) and consciously planning what is to come. For example, the idea of using an outline before drafting a paper is not rDOK 3 cognition, as it is merely a routine decision to use a common tool. However, developing the outline for the particular topic and purpose – and perhaps even audience – *does* constitute making a plan for the draft before writing the draft; this is prospective reflection that makes for rDOK 3 planning. Because the planning itself is the rDOK 3 cognition, developing such a plan can constitute rDOK 3 even without the plan being enacted.

rDOK 3 can also take the form of explaining prior work or revisiting prior work in order to revise it. Note that explaining prior work is different than merely showing one's work along the way; being more explicit is not the same thing as being reflective. Furthermore, offering a canned explanation for an approach is merely an act of recall/repetition (i.e., rDOK 1) and not truly reflective at all. Bloom (1956) acknowledged that one cannot necessarily identify the cognitive path simply by looking at the product, as an insightful analysis could be a product of the student's own work or could merely be something they recall from their prior lessons. Reviewing work to catch and correct errors is not necessarily rDOK 3 cognition. Spelling checking and other line editing is *not* truly reflective, whereas re-evaluating the appropriateness of an approach or the validity of an argument *is* sufficiently reflective – even if one concludes that it *is* valid. (For more discussion of the ambiguity of cognitive complexity in test takers' final work products, see *rDOK 2*, below.)

rDOK 3 standards explicitly call on this kind of planning and/or iterative revisiting of work for substantive evaluation. They are among the easiest standards to classify, as *their* verbs are often quite explicit about planning and/or reviewing and the rest of their language makes clear they are focused on metacognition (i.e., thinking about thinking).

rDOK 1 – Recall

The time limitations on large scale standardized assessment combine with psychometric requirements (i.e., which encourage more shallower items instead of fewer deeper items) to make rDOK 1 tasks and cognition the dominant level of cognitive complexity on tests. Unfortunately, this has historically led to the inflation of reported DOK levels to use more of the 4-level scale more often. This has always undermined Webb's (1999, 2005) goal of highlighting the limitations of large scale standardized assessment so that educators and policy makers can be sure to assess more complex forms of cognition using *other* classes of assessments.

rDOK 1 – like wDOK Level 1 – includes a broader range of cognition elicited by assessments than many people intuit. It includes the automatic word recognition of sight reading and quick recall of math facts. It also goes up at least through so many math and science solutions' "Oh, wait. I know how to do that. What's that formula again?" Following standard procedures – not unlike following a recipe – has always been part of this lowest level of cognitive complexity, given Depth of Knowledge's four levels (Webb, 1999, 2002, 2005, 2006, 2007). Yes, this is a *range* cognitive complexity, but it all falls in rDOK 1's range.

Because cognitive complexity is *not* about the complexity of the final result, unquestioningly following a long recipe or pre-defined procedure is rDOK 1 cognition – regardless of the complexity or length of the task. Even if successfully following some of those steps are difficult and call on great skill, it remains rDOK 1 cognition. Cognitive complexity is not about difficulty and is *not* necessarily positively correlated with proficiency, the most typically impressive work may be a product of the lowest level of cognitive complexity.

rDOK 1 standards describe knowledge, skills, abilities, concepts and approaches that students are meant to internalize to the level of regularized – and perhaps even rote – application. These include both factual declarative knowledge *and* procedural knowledge. Standards that lay out KSAs that proficient students should be able to do automatically and with ease are rDOK 1 standards[6]. As explored in Wine and Hoffman (2023), K-12 mathematics traditionally places a particularly heavy emphasis on rDOK 1 standards, with automaticity being the hallmark of math proficiency. Other content areas place heavy emphasis on rDOK 1, as well. Surface level reading comprehension usually has the automaticity of rDOK 1. The various declarative knowledge of social studies, and the standard formulas and routines of science are all found in rDOK 1 standards.

rDOK 2 – Basic Decision Making

The most challenging DOK classification decisions have always fallen around Level 2. Without Level 4 being highly germane, the tough cases are always going to fall on the 1-2 line or the 2-3 line, but it is even trickier than logical inevitability suggests. rDOK 2 is about basic conscious decision making, but that is not always easy to distinguish from its neighbors.

While the highly proficient may engage in some targeted cognition more or less by rote (i.e., rDOK 1), the less proficient may need to take a more deliberate path to use the same tools and ideas. Whereas the highly proficient student may automatically select the right formula or offer the correct spelling, the less proficient student may struggle a bit to be sure and have to consciously think through what the correct choice is (e.g., "Is that $4\pi r^2$ or is that $(4/3)\pi r^3$? Wait. Cubed is volume and squared is area. Got it."). That is, this kind of conscious decision making may stem from diligent care by a less confident student *or* from the novelty of a task even for a more

[6] If test owners decide that sufficient proficiency with a standard include both high automaticity and more deliberate use of the specified KSAs, a standard can allow for multiple rDOK levels. That is, like individual items, a single standard can be classified with multiple rDOK levels.

proficient (and confident) student. In either case, the student has to stop consciously weigh their options at one or more points along the way.

We call this *tactical* decision making, because it is made in the moment, without the kind of planning of rDOK 3. Working through a task and making each decision as it comes (e.g., "Which formula now?...Looks like I need to do this, now…Yup, that's gotta mean I do this…") is rDOK 2 cognition. Conscious decisions to deviate from even a simple recipe may be rDOK 2, such as trying to figure out how to substitute for a missing ingredient or how to alter a recipe to better match local tastes. When done by rote (e.g., automatically upping the garlic, as some do) it is not, but when done more deliberately (e.g., tasting a dish at the end and figuring out what addition is needed to make it *just* right) it can be. Of course, *planning* at the beginning how to alter a recipe to achieve some different goal – as one of us is so fond of doing – is rDOK 3 cognition.

rDOK 2 cognition can be quite difficult to distinguish from rDOK 3 cognition, merely by examining a test taker's work product. The final product could be product of a carefully worked out process by a less proficient student or of an easy fluid effort by more proficient student who could just wing their way to the end. Cognitive complexity is a trait of the cognitive path, not of the final destination.

Despite all of this, these sorts of classifications are not necessarily difficult to determine, once one has clearly identified a cognitive path. Rather, the reality of the variation across test takers makes it difficult to limit potential rDOK 2 classifications to *just* rDOK 2. Professional judgment resting on deep expertise in the content area and/or broad experience with test takers often reveals such a variety of authentic cognitive paths that individual items may require multiple rDOK levels. Ideally, items present sufficiently novel problems that even highly proficient students must stop and take stock without simply racing through with rote applications of targeted cognition. Unfortunately, this is rare.

Consider the right triangle in the problem above as an archetypical example. It is not the cliché 3-4-5 triangle that is so well known as to be instantly recognized by most students, and the arithmetic of a 5-12-13 triangle is not so complex as to trip up many students. It seems to be at a sweet spot. But *some* students will simply remember that 5-12-13 is a Pythagorean Triple and will simply rely on that memorized knowledge, almost as a reflex. The fact of variation across students and the importance of recognizing cognitive complexity as a trait *of the cognitive paths taken by test takers/students* makes rDOK 2 the most contestable level of cognitive complexity. It also means that many standards and many items are legitimately classified in multiple levels.

rDOK 2 standards describe decision-making around options, but not the longer view planning of a longer cognitive path (i.e., rDOK 3). Frankly, there is a question of professional judgment about whether many standards should be classified at this level at all. One can look at standards that describe what would *ideally* be mastered to the automaticity of rDOK 1 and conclude that they also allow for less fluid and sure application – so long as that cognition still leads to successful final product. For example, is it sufficient to be able to *work out* which was President Andrew Johnson and which was President Andrew Jackson, or would such a standard require more automatic command of the difference? Must one know all the perfect squares up to 14^2 both forwards *and* backwards, or would be sufficient to count up through them to find that $\sqrt{169} = 13$?

Hence, projects to classify standards by rDOK cognitive complexity requires deciding the level of proficiency that a standard demands. There are sustained and important decisions to made in advance of any rDOK classification project.

Inter-Rater Reliability

We conducted a very small inter-rater reliability (IRR) study to confirm the feasibility of this approach, more a proof of concept than even a pilot study. Rather than using the standard practice of content review committees and validity studies of having raters confer and see if they can come to an agreement (Herman, Webb & Zuniga, 2007), the two raters we used for each of Mathematics and ELA did *not* confer. The two Mathematics raters *had* done a similar activity six months earlier with quite different items, but had not communicated since – allowing for some drift from their previous calibration. The two ELA raters had discussed the rDOK construct and reviewed our documentation on applying it to ELA (Wine & Hoffman, 2022, 2023), but had never compared rDOK rating for items and therefore had never calibrated against each other. We did *not* use the industry-standard "consensus process" (Webb, Herman & Webb, 2007, p. 18) in which raters discuss their thinking together and are allowed to alter their ratings in response to those discussions. Therefore, we consider this quite a stringent test of the feasibility of the construct.

Mathematics

For this portion of the IRR study, we used the most recent New York State Regents Algebra I Exam (University of the State of New York, 2023a). We selected this exam because the items are publicly available, it includes a mix of selected and constructed response items (see Table 5), and public documentation lists the standards to which each item is aligned. Five of the constructed response items explicitly ask multiple questions, each of which we treated as a separate item to be rated for cognitive complexity.

Table 5

Math Items

Response Type	Number
Selected Response	24
Stand Alone Constructed Response	8
Grouped Constructed Response	15 (i.e., 4 + 3 + 2 + 2 + 4)
Total	47

There were calibration issues between the two raters, with one being more likely to see rDOK 1 path to successful responses and the other more likely to see rDOK 2 paths to successful responses, as shown in Table 6. Nonetheless, they agreed in over 90% of their rating decisions. (Because the rDOK levels are *not* mutually exclusive, each rater made four decisions about each item, totally 188 rating decisions across the test.) They agreed on all four rating decisions for over two-thirds of the items, in spite of not using a consensus process.

Because our protocol was more stringent than conventional practice, there is no clear reference for what agreement rates we should see. Herman, Webb and Zuniga (2007) set a bar of 65% agreement across a larger group for agreement for individual items. Admittedly, that is quite a different context, but it suggests that 65% perfect agreement (across four decisions for each item) across an entire test is *not* disqualifying of the construct or documentation.

Table 6

Math Items rDOK Ratings (Summary), Perfect and Overall Agreement Rates

Rater	rDOK1	rDOK2	rDOK 3	rDOK4	Perfect Agreement	Overall Agreement
1	30 (63.8%)	32 (68.1%)	2 (4.3%)	0 (0%)		
2	34 (72.3%)	26 (55.3%)	2 (4.3%)	0 (0%)		
Agreement	37 (78.7%)	39 (83.0%)	47 (100.0%)	47 (100.0%)	32 (68.1%)	170 (90.4%)

Of course, the real question is Cohen's Kappa (κ)– to address whether those agreement rates are simply of function of chance (Cohen, 1960; Hsu & Field, 2003)

Table 7

Inter-rater Agreement Rates and Cohen's Kappa (Math)

		Rater 1							
		rDOk 1		rDOK 2		rDOK 3		rDOK 4	
		rDOK 1	Not rDOK 3	rDOK 2	Not rDOK2	rDOK 3	Not rDOK 3	rDOK 4	Not rDOK 4
Rater 2	rDOK	27 (57.4%)	7 (14.9%)	25 (53.2%)	1 (2.1%)	2 (4.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
	Not rDOK	3 (6.4%)	10 (21.3%)	7 (14.9%)	14 (29.8%)	0 (0.0%)	45 (95.7%)	0 (0.0%)	47 (100.0%)
	Chance Agreement	56.2%		51.9%		91.8%		100%	
	Actual Agreement	78.7%		83.0%		100%		100%	
	κ	**51.4%**		**64.6%**		**100%**		-	

Again, these κ values (i.e., 51.4%, 64.6%, 100%) show that the rDOK construct for math is viable, even with the potential drift here due to lack of any calibration efforts between math raters for over six months. At worst (i.e. rDOK 1), they are in the *moderate* range, and they rise through *substantial* to *nearly perfect* (Landis & Koch,1977) – or *actually* perfect.

English Language Arts

For this portion of the IRR study, we used the most recent New York State Regents ELA Exam (University of the State of New York, 2023b). We selected this exam because the items are publicly available, it includes a mix of selected response and constructed response writing items (see Table 8), and public documentation lists the standards to which each item is aligned. The two constructed response items were aligned to 16 standards, each. We counted each as a separate rate decision because rDOK classification is about *the application of the targeted cognition.*

Table 8

ELA Items

Response Type	Number
Selected Response	24
Grouped Constructed Response	32 (i.e., 16+16)
Total	56

Unfortunately, the two ELA raters never engaged in any calibration training, only reading and discussing the application of rDOK to ELA items *in principle*. Nonetheless, they agreed in over

90% of their rating decisions. (Because the rDOK levels are *not* mutually exclusive, each rater made four decisions about each item x aligned standard, totally 224 rating decisions across the test.) They agreed on all four rating decisions on nearly two-thirds of the items, in spite of the lack of both calibration training or consensus process.

Table 9

ELA Items rDOK Ratings (Summary), Perfect and Overall Agreement Rates

Rater	rDOK1	rDOK2	rDOK 3	rDOK4	Perfect Agreement	Overall Agreement
1	55 (98.2%)	49 (87.5%)	24 (42.9%)	0 (0%)		
2	54 (96.4%)	48 (85.7%)	18 (32.1%)	0 (0.0%)		
Agreement	53 (94.6%)	47 (83.9%)	46 (82.1%)	56 (100%)	35 (62.5%)	202 (91.2%)

Because our protocol was more stringent than conventional practice, there is no clear reference for what agreement rates we should see. These rates exceed Herman, Webb and Zuniga's (2007) 65% -- though, again, that was for quite a different context.

Of course, the real question remains Cohen's Kappa.

Table 10

Inter-rater Agreement Rates and Cohen's Kappa (ELA)

		Rater 1							
		rDOk 1		rDOK 2		rDOK 3		rDOK 4	
		rDOK 1	Not rDOK 3	rDOK 2	Not rDOK2	rDOK 3	Not rDOK 3	rDOK 4	Not rDOK 4
Rater 2	rDOK	32 (57.1%)	1 (1.8%)	44 (78.6%)	4 (7.1%)	16 (28.6%)	2 (3.6%)	0 (0%)	x (0%)
	Not rDOK	2 (3.6%)	0 (0.0%)	5 (8.9%)	3 (5.4%)	8 (14.3%)	30 (53.6%)	0 (0%)	56 (100%)
	Chance Agreement	94.7%		76.8%		52.6%		100%	
	Actual Agreement	94.6%		83.9%		82.1%		100%	
	κ	-1.2%		30.8%		62.4%		-	

These κ values (i.e., -1.2%, 30.8%, 62.4%) are not as impressive as those for mathematics. We expect that with calibration training, these could likely be increased significantly – and certainly using a consensus process would increase them further. However, the math of κ calculations make it difficult to improve on chance 96% or 98% of items fit a category[7].

Final Discussion

Webb has always been clear that the purpose of alignment studies – and simply of thinking about alignment issues – is to highlight the discrepancies between large scale

[7] We are quite confident that the way that proficient reading works, the grade level of this test and the standard-specific demands of this test's reading passages combine with the use of multiple choice items such that almost every selected response item is addressable by proficient test takers with rDOK 1 cognition. We do not doubt that cognitive complexity of these items are, in fact, quite low for most test takers. Nor do we doubt that this a problem on many large scale standardized ELA test – a problem that has long hurt their facial validity with ELA educators. In this case, the items that raters did not classify as rDOK 1 have alignment issues and/or lack a definitive key.

assessments and the expectations for learning laid out in state standards and seen in classrooms. He pointed out that alignment studies – including examinations for cognitive complexity – are *supposed* to reveal shortcoming because evidence of some learning goals simply cannot be elicited in by large scale assessments, for any number of reasons. Webb's answer to this was that large scale assessments must be part of a cohesive system of assessments.

> Not all learning can be assessed by large-scale assessments. Teachers are in a much better position to assess important learning such as how well a student is able to perform a scientific inquiry or devise a mathematical proof. Aligning the assessment system with expectations serves as an inventory to help assure all outcomes are being assessed in some way. That is, the expectations are covered by the assessments. (1997, p. 2)

It is folly for anyone think that *every* important standard is appropriate for the kinds of large scale, standards-based, on-demand, timed, standardized assessments that we see today and/or have existed in our lifetimes. It is not news that standardized assessments are limited (e.g., see Lindquist, 1951). The lack of rDOK 4 and rDOK 3 items, and the ubiquity of rDOK 1 in ELA, is an accurate reflection of the cognitive complexity elicited by large scale standardized assessments.

While one can assail assessment development professionals for undermining the basic goals of cognitive complexity and alignment studies by bastardizing DOK definitions to inflate reports of cognitive complexity, we think that that would be a mistake. For all of our own frustrations with iDOK and misapplication of cognitive complexity, we know that the fault lies in the pressures put on assessments to report what they simply are incapable of reporting. So long as large scale assessment is viewed as the ultimate – or only – credible account of student proficiency, achievement and/or learning, *any* careful and thoughtful systems for evaluating item or test validity will be undermined. We fervently wish that every psychometrician and content development professional could rise together as one to educate the public and policy-makers about the limits of large scale assessment – especially give the budgetary and time constraints put on assessment development, assessment delivery and assessment scoring – and even refuse to contribute at all to contemporary misuse of assessment. But we know that such a thing is not possible. We ourselves strive to delivered improved assessments that come closer to their goals, as we are under no delusions that demand for large scale assessments will go away – nor that it should.

Rather, we ask for what we often ask for in professional work: humility. There are limits to what any structure can make a particular round of review catch. We acknowledge the limits of using cognitive complexity as an alignment tool, and do not at all suggest that is sufficient. In fact, we believe that well aligned items should sail through cognitive complexity checks. The goal of cognitive complexity review to catch those most egregious violations of alignment – the ones that sell students short and cover for less challenging curriculum that undermines the deep and lasting lessons that people who love their content areas want students to retain for a lifetime. Undermining the rigor of typologies of cognitive complexity ends up mislabeling failures to align educational efforts toward the "worthy residue" (Sizer & Sizer, 2000, p. 50) of education that remains long after the lessons have been forgotten.

rDOK is offered as a simplification of wDOK, removing some of wDOK's ambiguities that stem from the multiple traits that Webb listed for different DOK levels – particularly Level 2. As such, rDOK is itself offered as a humble tool for its simplicity and as one that recognizes the challenges and pressures faced by assessment professionals. Perhaps in its simplicity, it will better protect classification of cognitive complexity from those pressures.

References

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

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.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37-46.

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.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A case study. *Applied Measurement in Education*, 20(1), 101–126.

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/

Hoffman, A. & Wine, M. (2021). *RTD Item Alignment Examination Annotated Procedure*. [White Paper]. AleDev Research and Consulting. DOI:10.13140/RG.2.2.15152.43528

Hsu, L. M., & Field, R. (2003). Interrater agreement measures: Comments on Kappan, Cohen's Kappa, Scott's π , and Aickin's . *Understanding Statistics*, 2(3), 205-219.

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.

Kvålseth, T. O. (1989). Note on Cohen's kappa. Psychological reports, 65(1), 223-226.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Linquist 1951Lindquist, 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.

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.

University of the State of New York (2023a). *Regents High School Examination: Algebra I (Wednesday, January 25, 2023).* https://www.nysedregents.org/algebraone/123/algone12023-exam.pdf

University of the State of New York (2023b). *Regents High School Examination: English Language Arts (Tuesday, January 24, 2023).* New York State Education Department. https://www.nysedregents.org/hsela/123/reela12023-exam.pdf

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. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, 26(2), 17-29 ,

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

Wine, M. and Hoffman A. (2022). *RTD Approach to Using Webb's Depth of Knowledge (DOK) Typology To Classify ELA Items* [White Paper]. AleDev Research and Consulting.

Wine, M. and Hoffman A. (2023). *Reinvigorating Webb's Depth of Knowledge in Three Content Areas.* Paper presented at the annual meeting of the National Council of Measurement in Education in Chicago.

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.