Gula, T, & Lovric, M. (2024) Quality Numeracy Tasks: Development and Stress Test of a Rubric for Teachers and Designers. *Adults Learning Mathematics: An International Journal*, *18*(1), pp.71-87.

# Quality Numeracy Tasks: Development and Stress Test of Three Rubrics for Teachers and Designers.

**Taras Gula**
George Brown College, Toronto, Canada
tgula@georgebrown.ca
**Miroslav Lovric**
McMaster University, Hamilton, Canada
lovric@mcmaster.ca

## Abstract

The development of a set of rubrics for teachers and curriculum designers interested in the quality of numeracy tasks emerged from our struggle with an attempt to clarify the distinction between numeracy and mathematics. Shifting our perspective from asking what numeracy and mathematics are to asking what they are about led to the conceptualization of a core mandate for numeracy scholarship and a recognition that a distinct feature of numeracy tasks is that they aim to inspire transfer from concrete to abstract thinking spaces and back.

Next, we used models of the thinking process for real-world problem solving developed by Verschaffel, Greer, & De Corte (2000), Burckhardt (2008), Blum and Ferri (2009), OECD (2017) and especially Wolfram (2020) to shape the development of Rubric 1 and 2 for discerning quality numeracy tasks. These models helped break down the idea of transfer from concrete to abstract thinking spaces and back into well-defined thinking actions. We then added a third rubric that aims to help teachers decide whether to use a particular numeracy task in their classroom setting.

In a conference workshop at ALM 30 (in 2023) we presented the rubrics and a series of tasks from a variety of settings as tools to help with the stress test. This paper first outlines the background and rationale for numeracy tasks as a distinct type of mathematics task, then presents an updated and scaled down version of the three rubrics, demonstrates them in action with a series of distinct mathematics tasks, and reports on what we learned from stress testing.

Key words: numeracy tasks; word problems; quality numeracy tasks; rubric development; transfer.

## From numeracy to numeracy tasks – a foundation.

One of the core motivations for numeracy in its original conceptualization (Crowther, 1959; Cockroft, 1982) came from the realization that school-mathematics skills were not automatically being transferred to the real world. Since the 1980s scholarship in numeracy has grown in scope and in depth, but the idea of transfer remains at its core.

This new term "numeracy" was needed to highlight a gap in mathematics education, not in mathematics. While mathematics is very useful in providing us with a way to think about (to make sense of) the concrete world around us, numeracy's goal is to help us think about transfer, an educational, not mathematical challenge. Indeed, the term numeracy is not one that mathematicians use when participating in the discourse of mathematics. Numeracy is not a child of mathematics, but rather they appear to exist on different planes of human endeavour.

Disentangling the two is easier when we consider what they are about and sets the stage for making numeracy tasks a well-defined subset of mathematics tasks.

In Gula and Lovric (2024) we set the stage for the development of numeracy task rubrics by gathering the threads of the scholarship of numeracy and its most prominent family members: Quantitative Literacy (QL), Quantitative Reasoning (QR), Mathematical Literacy (ML) and numeracy's word problem cousins. We coined the term 'word problem cousins' to indicate that scholarship focused on word problems has some of the same concerns as numeracy scholarship, however these scholars rarely refer to each other's work. We did not attempt another definition of numeracy as many numeracy scholars agree that there is no consensus as to a unifying definition (Geiger et al., 2015), even referring to the construct as 'slippery' (Coben et al., 2003 p. 9).

Instead of framing the question as 'What is numeracy?' we asked: 'What is numeracy about?'. This reframing helped us outline six areas of consensus in numeracy scholarship (Gula and Lovric, 2024, p. 2-9) that served as a foundation for a proposed core mandate:

> Our core mandate as scholars of numeracy is to investigate how our educational institutions and teachers within and outside of those formal systems can improve citizen (and student) use of mathematical abstractions to help make sense of the concrete world that they/we live in and to interpret the abstractions of others. (Gula and Lovric, 2024 p. 9)

A key aspect of the core mandate is a reframing of transfer as the shifting from concrete to abstract thinking spaces and back, instead of transfer from school-mathematics to real world as originally conceived. This reframing provides the groundwork for conceptualizing numeracy tasks as ones that provoke or inspire a specific series of thinking actions in the student.

Before moving on, we would like to acknowledge other numeracies that have emerged including statistical literacy and reasoning (DelMas, 2002), financial literacy (Titko & Lace, 2013), health numeracy/literacy (Schapira et al., 2014), data literacy (Gummer and Mandinach, 2015) and academic numeracy (Brady, 2016). Each aims to extend numeracy into new contexts or settings. While they broaden the scope of numeracy scholarship, we do not see them as changing the core mandate of numeracy scholarship.

## Shaping quality numeracy tasks as a distinct form of mathematical task.

A large body of scholarly work examines problem solving in mathematics and many of the thinking actions that good problems (good tasks) are to inspire in students (Polya, 1945; Gravemeijer, 1997; Căprioară, 2015; Hendriana, Johanto, & Sumarmo, 2018). Thinking actions gleaned from those include: reasoning (deductive, inductive), generalizing, making conjectures, persevering, experimenting, and proving. Similar thinking actions were proposed in a discussion of the mathematical habits of mind in Cuoco et al (1996). The fact that these actions are desirable and relevant whether the problem requires transfer (i.e., the shifting from concrete to abstract thinking spaces) or lives exclusively in the abstract suggests that they aren't helpful in shaping qualities of a numeracy task nor in developing a system of assessment specifically for numeracy tasks.

In her book "Thinking as Communicating" Anna Sfard (2008) describes mathematics as a discourse with its own thinking space. She points out that "Bertrand Russell … put it bluntly, mathematics begins where the tangible real-life objects end and where reflection on our own discourse about these objects begins." (Sfard, 2008, p. 129) Mathematics education is about bringing students into the discourse of mathematics, and good mathematics tasks (problems) can do that in part by living in the abstract.

Task 1 (below) for example, is a very basic arithmetic task that stumps many students (all high school graduates from a variety of countries) in a college mathematics class taught by one of the authors (many would put 10 and 18 in the two blanks). It comes from a mathematics sophistication index designed for prospective elementary teachers (Szydlik, 2009). The task can assess student understanding of the meaning of the equal sign if it involves some sort of reflection as part of the exercise, and in that way can bring students into mathematical discourse. Task 2, a much more complex and open-ended task provides many more opportunities for mathematical discourse, and for engaging the thinking actions important to mathematicians.

**Task 1:** 3 + 7 = ____ + 8 = ____. What numbers go in the blanks? (Szydlik et al., 2012)

**Task 2:** Inscribe a square in a given triangle. Two vertices of the square should be on the base of the triangle, the two other vertices of the square on the two other sides of the triangle, one on each (Polya 1945, p. 31).

Neither of these tasks aim to inspire transfer and thus they are clearly not numeracy tasks - they are not set in nor about the real (concrete) world. Although they are different from each other, they are united in that they exist purely in the abstract thinking space, the discourse, of mathematics.

Polya (1945, p. 93) described two types of mathematics problems, the purely abstract mathematics problems (like Task 2 above) and what he called "practical problems…acknowledging that they [practical problems] are different in various respects from purely mathematical problems".

We propose two word problem examples (Task 3 and 4a) that are pretty standard in a variety of algebra and pre-calculus courses as examples of practical problems.

**Task 3:** A farmer has 2400 ft of fencing and wants to fence off a rectangular field that borders a straight river. He needs no fence along the river. What are the dimensions of the field that has the largest area? (Stewart, Clegg, & Watson, 2020, p. 337)

**Task 4a** (a related rates problem): imagine a spherical snowball with a radius of $x$ cm. It is melting so that the radius shrinks at a constant rate of $y$ cm per hour. How fast is the volume of the snowball decreasing?

Intuition suggest that these are, at best, contrived real-world problems, and at worst, ones that promote the plugging of numbers into a formula or some other standardized procedure approach to problem solving. Furthermore, these mathematics problems do not inspire the thinking actions described above. Evans (2000) and Jablonka (2015) help give a foundation to these intuitions when describing a tendency for mathematics teachers and faculty to give primacy to mathematical structures rather than concrete situations when posing mathematics problems set in real-world contexts. Indeed, word problems in Calculus textbooks (often referred to as "story problems") are typically about using a specific mathematical idea, structure or calculation rather than about the context described, and that fits with their course objectives. The point here is not to critique calculus textbooks, but to show that numeracy tasks have a different goal.

From this perspective, word problems such as Task 3 and 4a are poor numeracy tasks even though they are set in real-world contexts in part, because they are about abstract mathematical structures. In the first stage of a solution, students are expected to "strip" the problem of words to find out what mathematical structure they are supposed to be using. In being defined as a related rates problem task 4a's mathematical nature is made explicit, confirming that the problem is not about melting snowballs. The student who would go out and study snowball melting in practice would be considered distracted by the noise (the details) of the context.

The bias to see the real world as composed of mathematical structures that we must recognize when solving problems is seen in numeracy scholarship as well. In their detailed analysis of Quantitative Literacy assessments among Health Numeracy scholars, Vacher and Chavez (2009, p. 38) used the following mathematical structures as markers of numeracy skill: numeration and counting, hierarchy, arithmetic, multistep arithmetic, and probability. This kind of list is not an uncommon way of presenting the skills of numeracy and suggests that mathematical structures are what many real-world word problem tasks are about. Furthermore, even those who contribute to a sophisticated conceptualization of numeracy and its many dimensions indicate a bias towards seeing the real world through the eyes of mathematical structures. For example, Geiger states (our emphasis): "What it means to be numerate is in a state of constant revision because of the *mathematical demands* of an increasingly globalized world…" (Geiger 2016, p. 252). Note that we are not suggesting that knowledge of mathematics and procedures is not important to numeracy. Instead, we are suggesting that key aspects of numeracy's core mandate are being neglected in many tasks designed to inspire transfer.

Perhaps it is this bias that has prevented many from seeing two distinct types of numeracy tasks: one type of task is about mathematical structures and their relations, while a second is about the concrete context the task is set in.

Dahl et al. (2023, p. 29-30) use a Problem Based Learning (PBL) perspective that inadvertently helped support the conceptualization of numeracy tasks as a distinct type of mathematics task by delineating three types of problems in mathematics. Two of the three are designated as internal to mathematics, echoing the work of Anna Sfard (2008) who would describe them as ones that aim to bring students into mathematical discourse. The first of these two is set purely within the abstract world of mathematics (e.g., the arithmetic 'sophistication' task – Task 1, or Polya's inscribe a square into a triangle task – Task 2) and the second of the two is set in an artificial real world in order to motivate specific mathematical actions (e.g., the fence question – Task 3a and the melting snowball question – Task 4a). For example, in the melting snowball problem, the snowball is simply a convenient and interesting representative for a sphere rather than the sphere being a convenient representation of a snowball. Students can read the intent and respond in kind.

The third mathematics problem type includes those that are external to mathematics, fitting with the conceptualization of numeracy tasks as ones that inspire transfer from concrete to abstract thinking spaces and back. These problems/tasks are rooted in a concrete situation where mathematics can help make sense of and resolve any challenges faced by the protagonist(s). A correct solution will involve the use of mathematics, but the focus of the solution will be on the concrete world external to mathematics. This conceptualization provides a way of delineating two types of world problems, one about the real world (we call these numeracy tasks) and a second set in the real world, but about mathematics.

Dahl et al. (2023) also provide support for numeracy tasks as a type of modelling problem, one which promotes the use of abstract mathematics to make sense of the concrete. Calculus textbooks typically present tasks internal to mathematics (like Task 4a) where the use of elegant models that can be generalized to other situations is important. This is also a common approach to modelling problems in mathematics courses.

Let us look at a revised version of Task 4a which we have reset to be about the concrete context and thus framed as a numeracy task.

**Task 4b (snowball melting version 2): Assuming that you live in a place where there is currently snow, go outside and make a few different snowballs of various shapes and sizes. Take them inside and place them on a funnel over a measuring cup and watch them melt**

**tracking a rise in water level in the cups (or make a video watching them melt in the same setup). Is there some sort of commonality to their rates of melting? Is the level of water in the measuring cup related to the change in size of the snowball?**

Task 4b is framed as 'an investigation of melting snowballs' problem in which their shape, composition, density, etc. play a role in melting. In Task 4a the person completing the task is invited to play in the abstract side of the divide and engage in the internal discourse of mathematics, and in task 4b - external to mathematics - about the concrete situation presented, in which playing around in the concrete situation is more important than the mathematics used to do so. Both involve modelling, but only Task 4b fits with Dahl et al. (2023) as a problem external to mathematics discourse, and with the conceptualization of numeracy tasks that emerge from the core mandate of numeracy.

At this point we have established numeracy tasks as a distinct subset of mathematics (and mathematical modelling) tasks, and a distinct subset of word problems. Perhaps there is also a case to be made for numeracy tasks as distinct from mathematics tasks altogether, but that is not the goal here. Whether you choose to see them inside or outside the set of mathematics tasks really will not matter to the design of the numeracy task rubric, as the goal is to systematize the recognition of high-quality tasks that inspire the shift between concrete and abstract thinking spaces and give primacy to the concrete in creating numeracy tasks.

There are many elements to being numerate that have been set aside in our approach. For example, mathematical knowledge and skills, dispositions, contexts and cognitive processes, as suggested by the Common European Numeracy Framework (Hoogland et al. 2019). Those are important to consider as we work to help (and inspire) students to become better at shifting between concrete and abstract thinking spaces. Having a strong sense of what a quality numeracy task looks like is an important and often neglected aspect of numeracy education.

## Shaping components of transfer as a thinking process.

The impetus for creating a rubric for classifying the quality of numeracy tasks came from a frustration in seeing tasks that claimed to be numeracy tasks (often framed as real-world word problems) in published materials and presentations that many colleagues intuitively identified as weak, artificial, or lacking meaning without having any well-defined criteria to support those intuitions. A well-developed rubric for assessing students' quantitative literacy (AACU, 2009) and the rationale for Mathematical Literacy items used in PIAC assessments (OECD, 2017) provided guidance towards what those criteria may be, but we could not find a formalized system designed to assess the quality of the numeracy tasks themselves.

The reframing of transfer, key to the core mandate, points us towards a focus on thinking actions that make up the shift between concrete and abstract thinking spaces when considering the aboutness of numeracy tasks. Thinking actions specific to numeracy and thus to transfer are grounded in the work of a wide variety of scholars, yet show a remarkable commonality.

In Gula & Lovric (2024) we gathered a series of five thinking process models for what we call numeracy tasks and other scholars identify as word problem solving (Verschaffel, Greer, & De Corte, 2000), modelling (Burkhardt, 2008; and Blum & Ferri, 2009), mathematizing (OECD, 2017), and computational thinking (Wolfram 2020). These thinking process models gave us confidence to start the process of developing a rubric for the classification of numeracy tasks that would inspire students to experience and practice transfer. Though many of the models visualized the idea of transfer as shifting between real world and mathematics world, the visualizations and steps made it clear that thinking of these as shifts between concrete and abstract thinking spaces was more accurate. Though they do not mention the term transfer explicitly, each articulates a clear (though implied) delineation between concrete and abstract thinking spaces and the need to shift from one to the other when solving problems in the real (or material, physical, concrete, immanent) world.

Conrad Wolfram (2020) outlined a thinking process model that breaks down the solution to a good numeracy task (he calls it a computational thinking task) into four thinking actions: define, abstract, compute, and interpret. This model is the one that most directly informed the development of numeracy task Rubric 2 and implies transfer as the shift from concrete to abstract (through define & abstract actions) and back (interpret results).

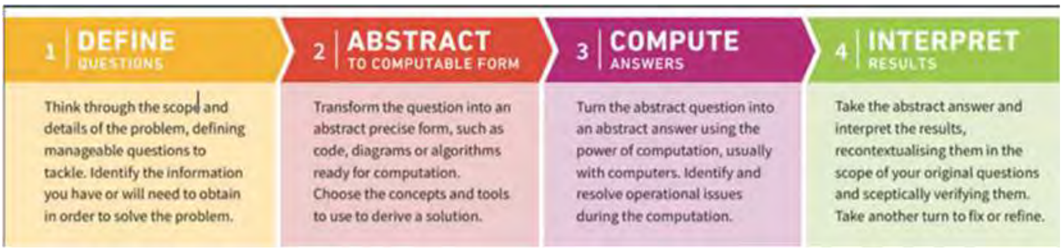

*Figure 1: Wolfram's model of computational thinking process (Wolfram 2020, p. 51)*

It is our hypothesis that many mathematics courses, or units within them that aim to inspire transfer tend to focus on the actions of computation and interpretation, and neglect the define and abstract thinking actions, which are much more difficult to set up as learning outcomes. The core mandate of numeracy (presented above) challenges us to think about how to engage students in the experiences of using mathematics to solve concrete problems, by engaging transfer between abstract and concrete thinking spaces. Ideas from problem-based learning and the thinking process models provide support for this conceptualization. The thinking actions described by Wolfram (2020) break the process down into elements that provide a path that helped us design Rubric 2 in particular, and can help us rethink curriculum in many courses.

## Demonstration of a task that inspires the four-stage thinking process.

We introduce Task 5 from an introductory statistics course in Health Sciences at an Ontario College to illustrate each of Wolfram's four stages from Figure 1 in action.

Note that variations of this task can be designed to inspire just one or two of the four thinking actions he describes. For example, a task that asks students to analyse real data - common practice in introductory statistics courses - allows faculty to develop rich tasks that inspire computation and interpretation actions. We suggest that making up data in Stage 2b as a part of a complete numeracy task gives students practice with abstraction without diminishing define and abstraction actions, and slightly diminishing interpretation since the results are known before the action takes place.

**Task 5:** Design and carry out a research study and write up results on the following topic: Does exercise help university students with sleep? You will complete the work in 4 stages with the following tasks for each stage:

**Stage 1 (define):** product is a clear research question that can be used in data collection.

- The process involves establishing the scope and parameters of the terms (What does exercise mean? How would one measure time of exercise or ask a survey question about it?) Establishing a connection between the research question and variables is not a trivial process and involves a continuous shifting back and forth between concrete and abstract thinking spaces in students. What assumptions are built into the research questions? Are they potential biases? Does the research question get at the actual problem?

**Stage 2a (abstract):** product is a data set template, and a plan for data collection (structured as a proposal for research and including survey/questionnaire or observation method).

- The process will necessitate thinking ahead to data analysis since the data that will be collected (and especially its type – categorical versus measurement) will define the computation to come and to choose the output (charts/graphs) that will help answer the research question. How will data about quality sleep be collected (as a yes/no question, a measurement/score)? The data set template has to fit with the data that will be collected, and with the framing of the research question.

**Stage 2b (abstract):** product is a data set with data made up by the student and entered into the data set template. Having the student make up data stimulates abstraction and is highly efficient in avoiding formal research ethics approval process, and allows for topics which would be impractical if data collection was needed).

- The process of making up data requires more than consideration of data type for each variable. Students must also take into consideration that summary statistics and visuals their made up data produce need to be grounded in the concrete real world – e.g., for a strong positive association between exercise and sleep the results need to show that exercise helps with sleep more than a little bit (strong positive association) without losing touch with what is realistic (e.g., sleeping 14 hours a day is not a good thing either).

**Stage 3 (compute):** product is output (usually software will produce summary statistics, charts/graphs and visuals) that could help answer the research question from stage 1.

- The process will require a comfort with the computation method used, but more importantly the ability to decide what output helps answer the question about exercise and sleep.

**Stage 4 (interpret):** product is a written report answering the research question with evidence. Furthermore, students are asked to consider how their made-up data might differ from what they would find if they actually collected data.

- Process will start with looking for anomalies (e.g., outliers or results that are not realistic – since data is made up) in output as a way of checking for validity, and if needed stepping back to stage 2b and 3 to fix any of these.
- After data validity is established, the student is expected to produce a written report demonstrating that they have answered the research question with the data at hand and to provide evidence to support their answer. Given that the data is made up, students are also asked to reflect on the extent to which the made up data fits with what would be expected in the population they are purportedly studying.

Now that we have established a clearer description of the four thinking actions that numeracy tasks can inspire, we will go to the next step of describing the development of criteria for a potentially useful rubric.

## Rubric development and description

A search in the literature yielded no formalized system that helps evaluate numeracy tasks posed in formative and summative assessments and exercises, though there were several articles that pointed to characteristics of good numeracy, QL, QR, ML or word problem tasks (Hoogland et al. 2018; Geiger 2016; Follette et al. 2015; Gaze et al. 2014; Grawe, 2011). Others focused on evaluating student responses to problems posed. In particular, the Quantitative Literacy Values Rubric (AACU, 2009) proposes the following Quantitative Literacy skills (similar to

Define/Abstract/Compute/Interpret actions): assumption, application/analysis, representation, calculation, interpretation, communication.

We were able to identify nine characteristics of a good numeracy task (Gula & Lovric, 2024, p. 17) that provided the impetus to develop criteria of a numeracy task rubric – Rubric 1. After a few preliminary sketches we realized that instead of one rubric we would need two, then by the sixth draft we saw the necessity for three rubrics that each answered a separate question. We introduce each rubric below by highlighting the questions they are to answer. Each rubric has multiple criteria with 3 descriptors for high/medium/low quality. The rubrics are not meant to be used independently of each other, rather to be used in succession.

We present the revised post stress test version of three rubrics below, showing the descriptor for high quality only. The complete rubrics with descriptors for 'medium and low quality' levels can be found online at Gula and Lovric (2023).

**Rubric 1 (for the teacher or designer):** Is the task posed clearly a numeracy task? Does the task inspire the shifting from concrete to abstract thinking spaces and back? This rubric is for the teacher or designer aiming to help assess the extent to which a numeracy task is well designed and to identify areas that can be improved. For each criterion there are three descriptors from high to low quality.

**Rubric 1. Purpose: for the teacher/designer to assess the degree to which a task is well designed in general over four criteria.**

| Criteria | High quality |
|---|---|
| **Nature of context (concreteness)** | The task description takes the student (or group of students) out of school (or away from their desk) and explicitly into a concrete context (without necessarily doing so physically)<br>It is external to both the class and any mathematics being studied in the class. |
| **Task description: content (aboutness)** | The task presented demands a focus on and response to a concrete situation within the scenario.<br>Though it demands the use of quantitative methods and/or mathematical tools, it is not about the mathematics that is/needs to be used.<br>Description makes it easy to assess whether a student (or group of students) is familiar with the language/culture of that context or has experience in it. |
| **Authenticity of scenario in context** | The scenario is presented in a way that it would appear in the given context.<br>The challenge/problem is one that would arise in that context. |
| **Task description (required actions)** | The task described makes explicit the need for responses (however trivial) to each of the following actions of the adapted Wolfram computational thinking process (Define, Abstract, Compute, Interpret - see breakdown in Rubric 2).<br>Multiple responses to the posed task are possible, though some may be judged as stronger (or weaker) if errors are made.<br>A solution using informal methods of abstraction or computation is just as valid as one using formal mathematical methods. |

**Rubric 2 (for the teacher or designer):** To what extent does the task inspire responses to each of the four thinking actions necessary for exercising transfer: Define, Abstract, Compute, Interpret? Note that we are not suggesting that each task must inspire all four. In many circumstances teachers may choose to get students to practise individual actions before putting them all together.

**Rubric 2. Purpose: for the teacher/designer to assess the degree to which the task is able to inspire responses to each of the four thinking actions associated with transfer: Define, Abstract, Compute, Interpret.**

| Criteria | High quality |
|---|---|
| **Define the question.** | Scenario presented such that student will need to reframe the scenario into quantitative friendly format in order to complete the task.<br>The student will need to describe assumptions or simplifications made to reduce ambiguity inherent in task description. |
| **Abstraction:** | Scenario requires reframing into a formal or informal mathematical model (i.e., in preparation for computation) and allows for at least one correct approach.<br>No new mathematics needs to be developed by the student to solve the task(s) presented in the scenario. |
| **Computation and result(s):** | Scenario requires calculation(s) without indicating need for technological help.<br>The calculation method used by the student can be similar to what would be expected in the context of the scenario, but does not need to be so.<br>Expected results of calculations can include visual and or numerical formats, charts and/or graphs as warranted and appropriate. |
| **Interpret results: Task response expectations** | A correct response to the scenario is context based, and requires more than just presenting the result of calculations including some discussion of define and abstract actions, justifying any decisions made.<br>May include analysis, discussion, justification with logically sound narratives. |

**Rubric 3 (for the teacher who is deciding whether to use the task in class):** Is the task designated as good enough by Rubrics 1 and 2 a good one to use for the students I am teaching? Once a numeracy task passes through the first two filters as high (or high enough) quality, the teacher needs to make sure that it is right for their setting/context and their students. If not, these tasks can be tweaked by the teacher to reduce unnecessary barriers.

**Rubric 3. Purpose: for the teacher who is deciding whether to use a quality numeracy task** in a class. Assumption: the task passed through Rubrics 1 and 2 as good enough.

| Criteria | Yes: use it |
|---|---|
| **Appropriateness of context** | The concrete context that the student is placed in is one that the student (or group of students) has experience with directly or indirectly or has the capacity to grasp. |
| **Accessibility of language and terminology** | The language used (especially technical terminology native to the context) is familiar to the student (or group of students) in the classroom setting for the course being taught. |
| **Mathematical expectations** | The formal or informal mathematics needed to solve the problem are familiar to the student (or group of students) to whom the task will be presented. |

## The rubrics in action - a demonstration with three tasks

We have chosen three very different tasks to analyse through the lens of the rubrics. They were chosen not only to give the reader a sense of how each rubric works, but also to demonstrate their effectiveness at recognizing high quality numeracy tasks (i.e., ones that fit with the conceptualization of them as distinct in their aboutness), filtering out poor tasks and pointing to areas for improvement in task design. For each task we will go through the three rubrics criterion by criterion.

**Task 4a (Melting snowball problem – original version) versus Rubric 1:** The Melting snowball problem was introduced earlier as a related rates problem, suggesting that this task is about abstract mathematics rather than the concrete world of melting snowballs. Thus, we would not expect it to be judged as a high quality numeracy task.

Criterion 1 (concreteness): This task does not take the student away from their desk as the framing is not of any actual space, time or particular snowball. This task framed as a melting snowball may be an interesting thing to think about, but more as a puzzle than as any melting snowball that exists and for which someone needs to figure out the speed that its volume is shrinking. Rating: very low.

Criterion 2 (aboutness): The discourse of this task is far from the discourse of melting snowballs. Terms like 'spherical' and 'diameter' are from the discourse of mathematics - they are indicators that the snowball is simply there to motivate a specific mathematics routine or set of procedures using particular structures of mathematics. Given that there is no plausible concrete context (e.g., engineers getting ready for a snowball fight with rising temperatures?). Rating: very low.

Criterion 3 (authenticity): There is no context in which concern about the speed of melting of a snowball might be of value. Rating: very low.

Criterion 4 (required actions): In short: Define – no need; abstraction – minimal need; computation – yes; interpret – cut and paste interpretation will suffice. The problem does inspire some shifting between concrete and abstract, but very little. Rating: low.

In conclusion, the melting snowball task (#4a) fails to make it through Rubric 1 and would not be recommended to be used in a numeracy course, or in a mathematics class where the teacher is interested in getting students to experience transfer from concrete to abstract thinking spaces and back. However, it may be an acceptable and useful mathematical problem in other aspects where building understanding of mathematical structures is the goal.

We will not take this task to Rubric 2 or 3 as it qualifies as a very low-quality numeracy task.

**Task 6 (Bat and Ball task):** A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

The Bat and Ball task is used in a variety of settings as a test of cognition and is considered to be quite effective. It is part of the Cognitive Reflection Test (CRT) first described by Frederick (2005) in which he examines the extent to which people challenge their intuitive thinking. It challenges the habit of applying a mathematical procedure (structure) before taking time to define the question being posed. We first came across this task in Daniel Kahneman's (2011) "Thinking Fast and Slow" as a task that exposes fast thinking. We have used it as an exercise in building student self-awareness of their responses in mathematics problem solving but have not thought about it as a potential numeracy task until development of the rubrics.

**Bat and Ball task versus Rubric 1:**

Criterion 1 (concreteness): The task does not make any effort to take one out of the classroom and into another setting other than one in which the cost of a ball is of concern. It suggests the use of a mathematics formula, and thus is not external to mathematics. Rating: low.

Criterion 2 (aboutness): The scenario does demand a focus on and response to a concrete situation, but it is not about that concrete situation, it is about developing awareness of habits where fast thinking wants to use a convenient, but incorrect procedure from arithmetic. The lesson has nothing to do with purchasing baseball equipment, it inspires a need to slow down our thinking to consider the concrete. Language used in description is pretty clear unless it is being read by a person with no baseball or cricket knowledge. Rating: medium.

Criterion 3 (authenticity): there is no real context in which one would be interested in the cost of a ball or be told that the bat costs $1 more. Rating: very low.

Criterion 4 (required actions): Define – no need as scenario is straightforward. Abstraction – this is where the challenge lies for the person completing the task - informal methods are possible. Computation – straightforward. Interpret – the task as posed does not require self-reflection, or justification but can be extended to do so. Rating: hard to assess as there is a need for only two of the four thinking actions – a more thorough assessment can be done with Rubric 2.

**Bat and Ball task versus Rubric 2:**

Criterion 1 Define the question: no need for reframing, nor for statements of assumptions/simplifications to reduce ambiguity. Rating: low.

Criterion 2 Abstraction: The requirement for abstraction involves focus on the concrete scenario and avoidance of how fast thinking sends the reader to subtraction (incorrectly). For many (including one of the authors) appropriate model building begins after first making the fast thinking error. The algebraic model ($x + y = \$1.10$ and $x - y = \$1$) is one possible goal in a classroom, but there are informal ways to represent solutions to this task. No new mathematics is needed for someone with basic arithmetic. Rating: medium to high.

Criterion 3 Computation and results: most basic arithmetic calculations are all that is needed, or perhaps an algebraic representation, but there is no authentic approach as the context is not one that would appear in any situation outside of the classroom. Requirement is a simple numerical result with dollar (or cents) sign. Expectations are simple and straightforward. Rating: high.

Criterion 4 Interpret results: Correct response is just a number with unit and the most commonly made error ($0.10 for price of ball) can be double checked, but that is not required in the task – it is presented as if it was a straight calculation question. Rating: low.

This task does not inspire all four of the thinking actions that we propose are important to numeracy tasks. Thus, we would not suggest promoting it as a complete numeracy task. This does not mean that it would not be a really good task that could promote the need for slow thinking when working on abstractions (building models) of concrete phenomena. Its usefulness is in challenging students to make sure they were not just applying the first mathematical structure that popped into their heads when solving problems in concrete contexts.

**Task 7 (Rental Car task):** CTV News (2022) reported the following story (shortened here):

> A woman who rented a car in Toronto said she was charged $8,000 after the company made an error. Giovanna Boniface said she rented a car from Avis for three days after she travelled to Toronto Pearson earlier this month to help get her daughter settled at university driving a total of about 300 kms.

> Boniface said she prepaid about $1,000 to rent the car. While waiting to board a flight, she said she checked her credit card statement online to make sure the charge had been processed correctly.
> "That's when I notice this charge for over $8,000 from Avis," Boniface told CTV News Toronto on Thursday. The company had charged her for driving 36,482 kilometres at a rate of 25 cents per kilometre.
> She tried calling the Avis location at the airport, but no one would answer the phone, finally getting through on their general phone number, but still had issues sorting out the problem.
> "They didn't seem to really get what my issue was and I really needed them to remove this $8,000 charge," she said.

Construct an argument that Giovanna could use to show that the rental company made a serious mistake.

This task comes to us from a popular numeracy course (named “Numbers for Life”) created and taught by the second author at his institution.

**Rental Car task versus Rubric 1:**

Criterion 1 (concreteness): The task does take the student working on it out of the school setting into a concrete real world context by suggesting that the student needs to help the subject make their case to the car rental company and by using a scenario published by local media. It is unlikely that this task would be presented within the context of any mathematics topic, and in this way is external to “school mathematics”. Rating: high.

Criterion 2 (aboutness): This task demands a focus on and response to the subject in trying to help correct the error made (and get the money back). There are a few quantitatively grounded reasons available to her, and deciding which is best is not based on any particular mathematical procedure or structure. Even though renting a car is a fairly common scenario in wealthier countries, not every student would have been exposed to the details of credit card payments and reimbursement, and car rental (and payment by km etc), however, the details described make it relatively straightforward to decide whether the students we are teaching will find the setting a barrier, and the universality of trying to make one’s case in being wronged by a large corporation is pretty close to universal in our world. Rating: high.

Criterion 3 (authenticity): The problem is presented in a way that would appear within the given context as it is a redacted quote from a published news report. People pay for something using a credit card, and most then check to see if they were charged the right amount. Mistakes happen, and it is not inconceivable that a charge on one’s credit card contains an error. Although there is no specific mention of which documents Giovanna could present, or had access to (for instance, the exact number of kilometres she drove), the article mentions the information that she was able to gather from looking at the invoice. This information fits within the expected process of correcting a wrong. We would give a higher rating if the actual invoice was presented to demonstrate the values presented by the rental company and maybe a map of the area in which the subject travelled. Rating: medium to high.

Criterion 4 (required actions): Define – need for investigating multiple options – thus yes! Abstraction – no real need for formal mathematical structure, the approach could be informal, but will involve abstractions - basic arithmetic will suffice, so no new mathematics is needed. Computation is straightforward, though visualizations may take time. Interpret – needed to construct a convincing argument that the rental company made a mistake will inspire some reflection on the decisions that were made. Rating: medium to high.

**Rental Car task versus Rubric 2**

Criterion 1 Define the question: A student is expected to go over multiple representations (abstractions) and think about which is the most straightforward and producing a compelling narrative to help Giovanna convince the company that they made an error. Lots of shifting to abstract and back. Not too many assumptions need to be declared, and the scenario is simple enough that there isn't much need to reduce ambiguity. Rating: medium to high.

Criterion 2 Abstraction: There are multiple possible approaches that could help build a solid argument for the Giovanna, none of which require any new mathematics for a vast majority of those who have experience with formal education. There may be no obvious correct approach, but instead many possible good (and correct) approaches from which to choose. For instance, one abstraction could involve a comparison between the total kilometres driven as claimed by the rental company and some familiar distance. An appropriate framing (such as a comparison to a known quantity) would be expected – so the decisions made in abstraction requires thinking ahead to interpretation and back to definition. Switching between the concrete and abstract thinking space will not necessarily be linear in this case. Rating: high.

Criterion 3 Computation and results: Depending on the approach, this task may not require any computation, but just a comparison of quantities (distances). Thus, no technology is needed to answer this task. Producing visualization may be aided by technology where possible. Rated high.

Criterion 4 Interpret results: Given that the student is asked to construct an argument, they will need to do much more than just present numbers – that approach seems not to have worked for Giovanna. In framing their argument using a comparison with a known distance, the student may choose to argue that "the distance of about 36,000 km is nine times the approximate driving distance from Lisbon to Kyiv, (for Europeans). or Vancouver to Toronto (for Canadians) Just presenting the facts will likely not work, otherwise Giovanna would not need other people's help – which adds to the complexity of the response needed. Rating: high.

**Rental Car task versus Rubric 3:** no ratings are provided as they would be dependent on the context the reader is considering. The discussion simply provides some evidence to its utility, and explanation of how this rubric can be useful.

Criterion 1 Appropriateness of context: This task would not be a useful task for engineering students for example, nor one that is set in an elementary class, or remote community without any of the issues confronted by Giovanna, not because of the mathematics involved, but because it would be unlikely to engage the students in wanting to help.

Criterion 2 (Accessibility): The task works for students in North America where renting a car is something that many people do, or at least is available as an option as the discourse is straightforward without technical language particular to any discipline (high accessibility). Given that fewer individuals get their news from reading news reports, the presentation of information may be uncomfortable for some students.

Criterion 3 Mathematical expectations: The mathematical expectations are not dictated by the scenario, with many options available, especially informal ones.

## Rubric development and stress tests

The authors developed multiple drafts of the rubrics through a less formal stress testing with a variety of tasks before we introduced it to a wider audience. We have had two opportunities to present the rubrics to a knowledgeable audience: the first at the National Numeracy Network conference in the United States in the fall of 2022, and the second at the Adults Learning Mathematics conference in Ireland, in the summer of 2023. In each we presented all three rubrics with 12 tasks in one-hour face to face workshop sessions we called stress tests.

Our observation of the interactions and discussions made it clear that for almost all attendees much more time was needed to absorb the complexity and intent of the three rubrics as well as the tasks.

What we learned about setting up stress tests (from notes we made rather than direct suggestions from the attendees):

- It is best if attendees read through the rubrics and their intent before stress testing them. Due to the novelty of our rubrics, the need for presentation of the theoretical underpinnings took time from the workshop sessions, and thus from working on assessing their effectiveness.
- Present a scaled down version of the three rubrics to work with and have the complete version available in the background. Having all three rubrics with all three levels (high, medium, low) available to attendees meant that time was wasted in trying to read through and absorb all the rubrics and their criteria.
- Carefully curate the tasks that are presented to attendees. We presented 12 tasks and allowed attendees to choose which to use. This meant that time was spent by the groups to select, solve and enjoy the tasks before putting them to use in assessing the rubrics. Choosing fewer tasks that are standard and familiar to the audience will help attendees focus attention on assessing the rubrics.
- The moderator of the session needs to direct the discussion. Many of the attendees spent time playing with the tasks we presented, enjoying their challenges - as mathematicians are happy to do - rather than examining the rubric's effectiveness. This was a surprise, but should not have been.

What we learned about the rubrics was generated from comments of the attendees during the sessions and in written responses post-session. Almost all suggestions have been addressed in the rubrics presented above in this document. It is worth noting that the feedback and suggestions we received were exclusively focused on the form of presentation of the rubrics rather than their content. We have taken this as an indication of the general agreement of those in attendance with the approach and ideas we presented, though we do not have any formal evidence to support this contention.

- The rubrics were generally very well received – there was interest in its uniqueness given that we do have rubrics for student responses to tasks, but no system of evaluation for the tasks themselves. Interestingly, we did not get a sense from the attendees that assessing the tasks/problems/exercises we give to our students was something they felt a need for in their teaching or research work.
- The novelty and complexity of the rubrics (three rubrics with a total of eleven criteria between them) made it difficult for many of the attendees to get familiar enough with them to meaningfully put them to use. Perhaps this is why some suggested reducing the scope of what the rubrics were intending to do.
- Many suggestions to reduce the wordiness and verbosity of the descriptors – having long descriptors contributes to the need for more time to absorb and acts as a disincentive for those that would like to try to make sense of the rubric quickly.
- Bullet points make information much more accessible than long text.
- It is crucial to have consistency in language use across levels moving across from high to low quality. The improved versions are not pictured in this document, but accessible online (Gula and Lovric, 2023).
- Phrasing the descriptors positively and avoiding the use of negative phrasing like 'unrealistic' or 'impossible' helps open discussions on improvement of the task.

- Rubric 1 and Rubric 2 need to be presented in a way that the distinction between them is clearer. This will help the user know how to use them and when.

## Conclusion

The goal of this paper is to provide a rationale for a numeracy task rubric, to introduce you to 3 rubrics that emerged as we worked on developing our original version, and to demonstrate their potential efficacy through their use with three tasks and by describing outcomes of the stress tests. We recognize that the process of validating these rubrics is in its infancy.

Support for the content validity of the rubrics comes from a theoretical foundation that justifies numeracy tasks as a distinct form of mathematics tasks, specifically as a distinct form of a mathematical modelling task. Key to the theoretical foundation is a reframing of transfer as a shift from concrete to abstract thinking spaces and back, building on the original conceptualization of transfer as a shift from school mathematics to the real world outside the classroom. This reframing is rooted in the work of a variety of scholars in mathematics education, in particular those that broke down the shifting between concrete and abstract thinking spaces into distinct thinking actions. The usefulness of the theoretical foundation was demonstrated by examining four tasks showing how the reframing of transfer helps see that numeracy tasks form a distinct subset of word problems (or modelling problems) and mathematical tasks.

The rubrics were introduced by describing the process of their development including a rationale for the need for three rubrics as opposed to one.

We also presented demonstration of the utility of the three rubrics, challenging Rubric 1 to assess the extent to which three tasks inspired the shift between concrete and abstract thinking spaces, challenging Rubric 2 to isolate the four thinking actions important to this shift, and challenging Rubric 3 to help teachers gauge the appropriateness of those tasks for their particular classroom.

The stress tests we put the rubrics through helped us streamline and improve the rubrics and their framing, but we cannot take them as indicative of a general consensus in the mathematics education community as to their effectiveness as tools to help improve the numeracy, QL, QR, ML or word problem solving of students in the courses where teachers use them. There is much more work to be done in that area. Anecdotally we have found that there seemed to be general agreement that the rubrics are a useful addition for teachers interested in improving their understanding of the numeracy tasks they present to their students in mathematics or math-related courses, but few expressed feeling an urgent need to use them in design or reflection of their own work.

There are many potential avenues of exploration from this starting point. A more formal research project that helps establish the validity of the rubrics, but more importantly of the reframed transfer metaphor, would be a natural next step. Another avenue, which many teachers and curriculum designers would find useful is to work on curating a series of high-quality numeracy tasks set in a variety of contexts (and appropriate for a variety of levels) that teachers can use in class or use as models to develop their own. Most importantly, we aim to inspire faculty to use these rubrics and provide feedback as to their effectiveness to help students strengthen their understanding and use of mathematics in solving concrete problems.

## Acknowledgements

Our research was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) grant 890-2015-2034. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of SSHRC.

We are particularly grateful to Kees Hoogland who provided comprehensive and detailed written suggestions in the weeks after the conference ended.

## References

AACU (Association of American Colleges and Universities). (2009). Quantitative literacy value rubric. https:// www.aacu.org/value/rubrics/quantitative literacy

Blum, W. & Ferri, R. B. (2009). Mathematical modelling: Can it be taught and learnt? *Journal of Mathematical Modelling and Application*, *1*(1), 45–58.

Brady, K. (2016). Developing students' academic numeracy skills: Taking a whole-of-institution approach. *Proceedings of The Australian Conference on Science and Mathematics Education* (pp.176-181).

Burkhardt, H. (2008). Quantitative literacy for all: How can we make it happen. In Steen, L. (Ed.), *Calculation vs. context*, (pp. 137-162). Mathematical Association of America.

Căprioară, D. (2015). Problem solving: Purpose and means of learning mathematics in school. *Procedia - Social and Behavioral Sciences*, *191*, 1859-1864.https://doi.org/10.1016/j.sbspro.2015.04.332

Coben, D., Colwell, D., Macrae, S., Boaler, J., Brown, M., & V. Rhodes, V. (2003). *Adult numeracy: Review of research and related Literature*. London: National Research and Development Centre for Adult Literacy and Numeracy.

Cockroft, W. (1982). *Mathematics counts*. London: Her Majesty's Stationery Office.

Crowther, G. (1959). *Ministry of Education: 15 to 18: A report of the central advisory committee for education* (England). London: Her Majesty's Stationery Office.

Cuoco, A., Goldenberg, E. P., & Mark, J. (1996). Habits of Mind: An organizing principle for mathematics curricula. *Journal of Mathematical Behavior*, *15*, 375-402.

CTV News (2022). Woman charged $8,000 after rental car company claims she drove 36,000 kilometres in three days. Toronto. https://toronto.ctvnews.ca/woman-charged-8-000-after-rental-car-company-claims-she-drove-36-000-kilometres-in-three-days-1.6042079

Dahl, B., Hüttel, H., Rasmussen, J. G., & Rasmussen, M.G. (2023). Contextualisation in university level problem-based learning. *For the Learning of Mathematics*, *43*(3), 27–31.

delMas, R. C. (2002). Statistical literacy, reasoning, and thinking: A commentary. *Journal of Statistics Education*, *10*(2). https://doi.org/10.1080/10691898.2002.11910679

Evans, J, (2002). *Adults' mathematical thinking and emotions: A study of numerate practice*. London and New York: Routledge.

Follette, K. B., McCarthy, D. W., Dokter, E., Buxner, S., & Prather, E. (2015). The quantitative reasoning for college science (QuaRCS) assessment, 1: Development and validation. *Numeracy: Advancing Education in Quantitative Literacy*, *8*(2).

Frederick S (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, *19*, 25–42. doi: 10.1257/089533005775196732

Gaze, E. C., Montgomery, A., Kilic-Bahi, S., Leoni, D., Misener, L., & Taylor, C. (2014). Towards developing a quantitative literacy/reasoning assessment instrument. *Numeracy* *7*(2): Article 4. https://doi.org/10.5038/1936-4660.7.2.4

Geiger, V., Goos, M., & Forgasz, H. (2015). A rich interpretation of numeracy for the 21st century: A survey of the state of the field. *ZDM*, *47*, 531–548. https://doi.org/10.1007/s11858-015-0708-1

Geiger, V. (2016). Teachers as designers of effective numeracy tasks. *Mathematics Education Research Group of Australasia*. Paper presented at the Annual Meeting of the Mathematics Education Research Group of Australasia (MERGA) (39th, Adelaide, South Australia.

Gravemeijer, K. (1997). Solving word problems: A case of modelling? *Learning and Instruction*, *1*(4), 389-397.

Grawe, N. D. (2011). Beyond math skills: Measuring quantitative reasoning in context. *New Directions for Institutional Research*, 149 41-52.

Gula, T. and Lovric, M. (2023). Health numeracy project (Projet de numératie de la santé): Numeracy task rubric. https://healthnumeracyproject.com/numeracy-task-rubric_full/

Gula, T. and Lovric, M. (2024) Numeracy tasks: Inspiring transfer between concrete and abstract thinking spaces. *Numeracy 17*, 1(4). doi: https://doi.org/10.5038/1936-4660.17.1.1447

Gummer, E. S., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record*, *117*(4), 1-22.

Hendriana, H., Johanto T., & Sumarmo, U. (2018). The role of problem-based learning to improve students' mathematical problem-solving ability and self confidence. *Journal on Mathematics Education*, *9*(2), 291-300.

Hoogland, K., Pepin, B., de Koning, J., Bakker, A., and Gravemeijer, K. (2018). Word problems versus image-rich problems: an analysis of effects of task characteristics on students' performance on contextual mathematics problems. *Research in Mathematics Education 20* (1): 37-52.

Hoogland, K., Auer, M., Díez-Palomar, J., O'Meara, N., & van Groenestijn, M. (2019). Initiating a common European numeracy framework. *Eleventh Congress of the European Society for Research in Mathematics Education* (No. 8). Freudenthal Group; Freudenthal Institute. ERME.

Jablonka, E. (2015). The evolvement of numeracy and mathematical literacy curricula and the construction of hierarchies of numerate or mathematically literate subjects. *ZDM, 47*, 599-609.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

OECD, (2017). PISA 2015 Mathematics Framework, in *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem* Solving. OECD Publishing. https://doi.org/10.1787/9789264281820-en

Polya, G. (1945). *How to solve it: A new aspect of mathematical method*. Princeton University Press.

Schapira, M. M., Walker, C. M., Miller, T., Fletcher, K. E., Ganschow, P. S., Jacobs, E. A., Imbert D, O'Connell M,. & Neuner, J. M. (2014). Development and validation of the numeracy understanding in medicine instrument short form. *Journal of Health Communication*, *19*(sup2), 240-253.

Sfard, A. (2008). *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511499944

Stewart, J., Clegg, D. K., & Watson, S. (2020). Calculus: Early transcendentals (9th ed). Cengage Learning.

Szydlik, J. E., Kuennen, E., & Seaman, C. E. (2009). Development of an instrument to measure mathematical sophistication. *Proceedings for the Twelfth Conference of the MAA's Special Interest Group on Research in Undergraduate Mathematics Education.*

Szydlik, J. E., Kuennen, E., Belnap, J., Parrott, A., & Seaman, C. E. (2012). Conceptualizing and measuring mathematical sophistication. [Unpublished manuscript] http://tinyurl.com/429z9bma

Titko, J., & Lace, N. (2013). Financial literacy: building a conceptual framework. *Proceedings of the 11th international conference Liberec Economic Forum;* (pp. 585-592).

Vacher, H L, & Chavez, T. (2009). Quantitative literacy on the Web of Science, 2: Mining the health numeracy literature for assessment items. *Numeracy 2* Iss. 1(5). doi: http://dx.doi.org/10.5038/1936-4660.2.1.5

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

Wolfram, C. (2020). *The Math(s) fix: An education blueprint for the AI age.* Wolfram Media, Incorporated.