United Nations Educational, Scientific and Cultural Organization

UNESCO INSTITUTE FOR STATISTICS

GLOBAL ALLIANCE TO MONITOR LEARNING

SUSTAINABLE DEVELOPMENT GOALS

Method for Developing an International Curriculum and Assessment Framework for Mathematics

GAML Fifth Meeting
17-18 October 2018
Hamburg, Germany

Summary

Background

This paper is presented to explain the methodology followed during the:

1) creating of a content and skills framework for mathematics from cognitive theory and various national curricula, and
2) development of a coding scheme to map various national assessment frameworks (NAFs) onto the framework.

The approach followed was intended as a way to model inter-jurisdictional mathematics assessments during the first eight years of formal schooling, within the broader objective of monitoring progress towards SDG 4, Indicator 4.1.1:

4.1: By 2030, ensure that all girls and boys complete free, equitable and quality primary and secondary education leading to relevant and effective learning outcomes.

4.1.1 Proportion of children and young people: (a) in grades 2/3; (b) at the end of primary; and (c) at the end of lower secondary achieving at least a minimum proficiency level in (i) reading and (ii) mathematics, by sex.

Defining Mathematics

There are various theories about how people learn and do mathematics. The Global Content Framework builds upon three definitions:

1. ***Mathematics literacy*** was proposed by PISA developers to describe abilities one's capacity to formulate, employ and interpret mathematics in various contexts (OECD, 2012).
2. ***Mathematics proficiency***, on the other hand, is defined as what one knows, can do, and is disposed to do (Schoenfeld, 2007).
3. ***Mathematics competency*** combines a domain-based developmental interpretation of growth with a process-based interpretation of how children learn and do mathematics (Niss & Højgaard, 2009).

These theories focus on mental attributes believed to be engaged when one is actively negotiating solutions to mathematics tasks, but they generally lack detail about task types and requirements and they rarely include a developmental focus. Curriculum, however, provides sufficient detail and can be used to develop a tool to effectively address SDG 4 Targets.

Using curriculum as a basis

There are three things that make curriculum attractive as a basis for the development of the Global Content Framework:

1. Curriculum has far greater "cash value" than theory. Millions of children come to understand mathematics as a result of curricular interpretations. It is curriculum that *de facto* defines mathematics for the millions worldwide lucky enough to receive any instruction.
2. Curriculum documents are built around detailed learning expectations organized in developmental sequence. Similarities in curriculum construction present an

opportunity to identify a common international reference list of detailed learning expectations.
3. Curriculum designers are more concerned with practicalities about how we appropriate mathematical ideas and generally less concerned with ensuring the logical structure of mathematics as a discipline is preserved.

National Assessment Frameworks (NAFs)

In many jurisdictions teaching and learning intentions have shifted from a focus on inputs associated with educational access and moved towards measuring specific outcomes. And increasingly, outcomes are described in terms of specific competencies with which children must demonstrate some facility before it can be assumed they are mathematically competent. As a result, there has been an inexorable shift towards testing to provide evidence of systemic educational effectiveness. Just as curriculum documents can be likened to educational intentions, national and international assessment frameworks can be likened to jurisdictional testing intentions (i.e., intended question types and scope). Assessment frameworks are outlines, therefore, that provide examples of task-types deemed to be of sufficient jurisdictional importance to warrant testing. As mentioned, scope of testable tasks always represent sub-sets of learning expectations defined in related curriculum.

Outcomes

Reference List

1. English-, French- and Spanish-language curriculum documents were transcribed into a common framework and organized by year (years 1 to 8). A five-level framework (from broadest to most specific: Domains, Sub-domains, Constructs, Sub-constructs, and Action: Target) was selected from theory and various curricula to organize learning expectations. Few national curriculum documents were explicitly designed to fit the five-level framework. Therefore, protocols were established to address such emerging problems.
2. Using a constant comparison approach, three Reference Lists by language root (English, French, and Spanish) emerged, which helped to interpret differences within respective language-domain frameworks and as such to preserve national curriculum expressions.
3. The last stage involved condensing the three Reference Lists to a single representative list using the same constant comparison method (Glaser, 1965).

Coding Scheme to map NAFs

The Coding Scheme was created using the same Domain to Sub-Construct framework developed for the Reference List. It was intended to guide any coder to locate specific NAF item-types onto the curriculum side of the model.

The final framework is an arbitrary four-level framework: labelled as domains, sub-domains, constructs, and sub-constructs.

1. Domains are the broadest structural component, consisting of *Mathematical Proficiency, Number and Number Systems, Measurement, Geometry, Statistics and Probability* and *Algebra*.
2. Sub-domains, meanwhile, are broad content structures contained within DOMAINs
3. Constructs and Sub-constructs represent the broadest mathematical content structures existing within sub-domains and constructs respectively.

The make-up of the final four-level framework was informed by the selection of curriculum documents used to develop it (this is why it is essentially arbitrary).

Bibliography

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems*, 12(4), 436–445.

Niss, M. & Højgaard, T. (2011). *Competencies and mathematical learning: ideas and inspiration for the development of mathematics teaching and learning in Denmark* (No. 485–2011). P.O. Box 260, Roskilde DK: Roskilde University, Department of Science, Systems and Models IMFUFA.

OECD. (2013). *Pisa 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

Schoenfeld, A. H. (2007). What is mathematical proficiency and how can it be assessed? In A. H. Schoenfeld (Ed.), *Assessing mathematical proficiency* (Vol. 53, pp. 59–73). Mathematical Sciences Research Institute Publications. New York, NY: Cambridge University Press.

Method for Developing an International Curriculum and Assessment Framework for Mathematics

Malcolm Cunningham
Institute of Cognitive Science
Carleton University
Ottawa, Canada

As part of UNESCO's *Education 2030* framework for action, the Institute for Statistics (UIS) is leading development of global indicators intended to track the progress of *Sustainable Development Goal* four (SDG-4). The overall objective is development of comparable trans-national indicators by encouraging, where appropriate, new methods, statistical approaches and monitoring/reporting tools. We addressed the SDG-4 overall objective by creating a way to model inter-jurisdictional mathematics assessments during the first eight years of formal schooling. The approach was specifically intended to: (1) create a content and skills framework for mathematics from cognitive theory and various national curricula; and (2) develop a coding scheme to map various national assessment frameworks (NAF) onto the framework. This document is a critical appraisal of the methodology used to achieve these two goals.

Rationale

Not all children who study mathematics enjoy equal *opportunities to learn* (OtL) mathematics. While this may seem platitudinous, almost too obvious to state, it is a useful starting point in the current discussion. The pivotal phrase, OtL, is traditionally defined with respect to heterogeneity in external factors such as curriculum, schools, teaching approaches and social or cultural access (e.g., Carroll, 1963; Choi & Chang, 2011; Cross, 2009; Schmidt, Cogan, Houang & McKnight, 2011). Windfield (1987), for example, characterized it as the provision of adequate and timely instruction prior to tests or examinations. The term, however, can also be thought of in another way; one, moreover, that aligns more closely with modern testing. Carpenter and Moser (1983) studied students' learning differences in primary mathematics classrooms. They followed a cohort of children from Grade 1 and found that by the time students reached Grade 3, 11% had yet to master number facts to 10 and 30% had not mastered facts beyond 10. Perplexed, the researchers compared student responses to different types of problems to explain the observed variability but they finally conceded that there was no outward reason; a conclusion made all the more vexing as these students and their teachers had been offered intensive support and advice throughout the study period. Thus difference in this case could only be explained by internal, or cognitive, factors suggesting that one's opportunities to learn are also linked to differences in mental attributes (Cunningham, 2012). Although more often investigated as test results in psychological research, cognitive factors have been framed in terms such as differences in working memory capacity (e.g., Geary, 2004; Meyer, Salimpoor, Wu, Geary & Menon, 2010), cultural heterogeneity (e.g., Chui, Chow & McBride-Chang, 2007), or differences in ways in which individuals process conceptual and procedural knowledge (e.g., Hallet, Nunes & Bryant, 2010). The trouble is, educational testing and decisions that fall from test results rarely take into account the complexity of mathematics OtL (c.f., Pellegrino, Chudowsky & Glaser, 2001; Schoenfeld, 2007).

And this is especially problematic when designing an approach to meet SDG-4 goals because there is arguably far more complexity when making inter-jurisdictional comparisons than is normally encountered in studies of intra-jurisdictional differences. For in addition to human variability in cognitive factors, countries differ widely in their interpretation of such things as mathematics curriculum, teaching approaches, schooling, and testing; all of which conspire to make test interpretation and reporting more challenging. But if a nuanced model could be developed–one that accounts for some jurisdictional variation in external

factors–it would undoubtedly enhance our ability to interpret observed inter-jurisdictional differences and more firmly ground systemic reform efforts.

An effective, more nuanced, model should arguably be based upon a stable set of external reference indicators as they will make it possible to calibrate judgments about mathematics test results on some sort of standard. Inter-jurisdictional comparisons or specific recommendations for reform can then be relatively assessed and validated. But having said this, there are real limits in our current ability to capture complexity in external factors. Data constraints being what they are, it is difficult to explicitly model social, cultural, or structural educational variability. So instead, the reference list will be constructed from cognitive theory about how children learn and do mathematics and how this maps onto details of various national curricula. National curriculum documents provide evidence about what jurisdictions regard as important with respect to teaching mathematics and learning sequencing, and thereby contain important external OtL information. Ultimately, robustness of a curriculum-based reference set and any subsequent determinations derived from it will rely on international consensus about appropriate reference indicators. Then, using this reference set as a foundation, a coding scheme can be developed to map various national assessment frameworks. The resulting model is henceforth referred to as the *Reference List & Coding Scheme* or RL&CS framework (see Figure 1). To conclude, this approach addresses SDG-4 by establishing a more complex interpretive environment than is possible if trans-national indicators are founded on quantitative metrics alone. Complexity, however, introduces new assumptions, opportunities and threats. The current document is an appraisal of the methodology used in developing the approach.

Theoretical Background

The theoretical case for the RL&CS model ultimately rests on a cognitive understanding about how children learn and do mathematics. Thus theory is used to ground model claims. Details of various national curriculum documents can then be mapped onto the cognitive model to create a theory–curriculum reference list. This, in turn, forms the foundation of a coding scheme designed to map national assessment frameworks.

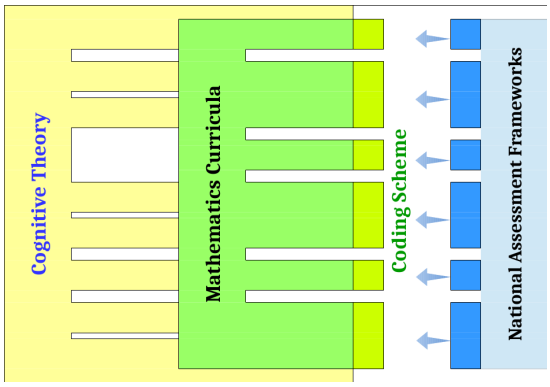

Figure 1. Illustration of the proposed RL&CS model for mathematics

Theories about Learning and Doing Mathematics

Theories about mathematical capability (also variously referred to as ability, literacy, proficiency and competency) provide conceptual descriptions for how people learn and do mathematics. They are used to rationalize such things as test design and item selection, and to shed light on the nature of mathematical cognition. Theory also often informs curriculum design.

Mathematics ability. E. H. Haertel and Wiley (1993) defined ability as the demonstration of procedural and conceptual knowledge and skills necessary for successful task performance. Ability is an estimable construct, however, only when analysis of performance is constrained to measurable features of tasks when completed in specific contexts. But it turns out that even if we constrain measurable features of tasks to be estimable there often remains difficult questions about what counts as knowledge and skills, the composition of estimable features, and the nature of appropriate contexts. Indeed, understanding how theory, measurable features of tasks and context coalesce to a working notion of ability inevitably involves a good deal of interpretation (E. Haertel, 1981).

Pellegrino et al. (2001) concurred with Haertel and Wiley's position adding that assessments of what children know and can do should be founded on a clear understanding of how interpretation is linked to cognition and observation. Hence, if cognition is a set of theories about how children learn and develop, and observation is defined as the kinds of tasks that evoke estimable demonstrations of ability then whatever notion

we hold about ability is at least partly based on theoretical rationale and partly on our interpretation of what we observe. Thus the concept of *ability*, its estimation and interpretation are very slippery things.

Mathematics literacy. Proposed by PISA developers to describe abilities of an average 15-year-old, mathematics literacy embodies one's capacity to formulate, employ and interpret mathematics in various contexts. These include reasoning, applying concepts, procedures, facts, and tools to describe, explain, and predict problem situations. A literate individual applies mathematical statements to and from the world reflecting well-founded capacity for judgment and decision-making that we would expect to be exhibited by constructive, engaged and reflective citizens (OECD, 2013).

The notion of literacy is founded on the existence of mathematical processes: students' abilities to (1) *formulate* situations mathematically; (2) *employ* mathematical concepts, facts, procedures and reasoning; and (3) *interpret*, apply and evaluate mathematical outcomes. PISA developers associated processes with mathematical knowledge domains to reflect major themes in various national curriculum documents. Developers called these domains Change and Relationships, Space and Shape, Quantity, and Uncertainty and Data. Presumably mathematical processes and knowledge domains are intimately linked so that a person who engages a question from any domain does so by marshalling mental attributes from among these mathematical processes. But ideas about mathematics literacy were principally developed to create robust test items, interpret students' responses and report results. Such things as details about the nature of items and how they are created is often proprietorial so there is little in the way of publicly available information about items and mathematical domains currently in use.

Mathematics proficiency. Schoenfeld (2007) defined the concept of proficiency to describe what one knows, can do, and is disposed to do. This idea embraces demonstrated facility with: (1) mathematical knowledge–facts and skills required for understanding mathematics in a specific context; (2) strategy use–formulating, representing, and solving problems; (3) metacognition–reflecting on problem solving progress during problem engagement; and (4) beliefs and dispositions–one's inclination to regard mathematics as sensible, useful and worthwhile.

As a theoretical entity, proficiency closely resembles the process component of mathematics literacy. It provides a framework for making sense of cognitive events which must be present as people engage mathematical tasks in the moment. Proficiency is not concerned, however, with content or skills associated with particular tasks or with how mastery develops over time. Instead, it is a theory which describes general cognitive features associated with learning and doing mathematics in any situation. Trouble is, when it comes to descriptions of different contexts, task types and mastery, interpretive differences inevitably arise.

Mathematics Competency. Jensen and Niss specify what it means to incrementally master mathematics through time with their cognitive theory of competence (Jensen, 2007; Niss & Højgaard, 2011). Broadly speaking, a mathematical competency is a well-informed readiness to act appropriately in mathematically challenging situations. It is associated with one's ability to ask and answer questions coupled with facility in the use of mathematical language and tools. Hence ability to ask and answer in, with and about mathematics is demonstrable through one's knowledge and skill with reasoning, modelling (i.e., ability to analyze properties of existing models and to actively model in given contexts and including ability to self-monitor), problem tackling (i.e., ability to pose and partly solve mathematical problems), and mathematical thinking (i.e., the nature, not content, of mathematical questions and answers). The ability to deal with mathematical language and tools, meanwhile, is demonstrated through various representations, symbols and formalisms, communication and the use of appropriate aids and tools. But generally speaking, competency theory is consistent with process components associated with other theories.

Competency stands apart from other theories in its linkage of cognition with content domains and mathematical development. Content domains are reminiscent of PISA, hence: Number Domains (i.e., concept of number and number systems); Arithmetic (i.e., four basic arithmetic operations, percentage and estimation and approximation); Algebra (i.e., characteristics of compositions applied to various sets of objects, operating rules, equations and solving problems); Geometry (i.e., properties of descriptive planar and spatial objects, geometric measurement, coordinate systems

and investigations); Functions (i.e., nature of functions, graphs of functions, linear and non-linear functions, rational functions, trigonometric functions, power functions and exponential functions); Calculus (i.e., continuity and limits of functions, differentiation, integration, differential equations and sequences); Probability (i.e. randomness and probability, combinatorial probabilities and finite spaces, stochastic models and probability theory); Statistics (i.e., organizing, interpreting and drawing conclusions about quantitative data, descriptive statistics, hypothesis testing, planning experiments); Discrete Mathematics (i.e., investigations of finite collections of objects, counting methods, combinatorics, number theory and graphs and networks); and Optimization (i.e., local and global extremes of real functions, optimization under constraints and linear programming) are organized in developmental sequence.

Niss and Højgaard (2011) identified and organized mathematics content domains in their model using language and interpretations that are consistent with mathematics as a discipline. Hence terms such as "algebra" or "geometry" are used instead of phenomenological descriptors like "form and shape" or "measurement." Table 1 summarizes the developmental architecture for the first 8 years of formal schooling.

Table 1
Illustration of Mathematics Competency content domains arranged in developmental sequence (Niss & Højgaard, 2011)

	Grades		
DOMAINS	1–3	4–6	7–8
Number	▲	▲	▲
Arithmetic	▲	▲	▲
Geometry	▲	▲	▲
Functions		▲	▲
Probability		▲	▲
Statistics		▲	▲
Discrete Math		▲	▲

Mathematics Curriculum Documents

Curriculum documents reflect societal agreement about what is important to teach, why, when it is to be learned, and how this all should be accomplished (Pratt, 1994; Tedesco, Opertti & Amadio, 2013). In other words, they are jurisdictional interpretations expressed as educational intentions roughly arranged in developmental sequence. This is the *intended* curriculum and not to be confused with how curriculum is actually implemented (i.e., what happens in classrooms) or what sense students make of what is taught (i.e., the attained curriculum ; Remillard & Heck, 2014). Needless to say, there is potential for substantial variability in how documents are interpreted and sequenced at each of these steps. But details and sequencing differences among inter-jurisdictional curriculum documents notwithstanding, four things make curricula very attractive templates from which to build an international reference list.

First, curriculum documents have a far greater "cash value" than theory. Consider the number of people who read about and understand cognitive theory versus the hundreds of millions of children who are subjected to curriculum each year. Regardless of whether or not cognitive theory presents a truer picture about how people mathematize, it is curriculum (and the myriad of tests designed from curriculum) that *de facto* defines mathematics for millions. Indeed, this is particularly true of mathematics as there is overwhelming evidence that outside of school, adults are increasingly less capable of offering effective help to struggling students beyond the elementary grades. Thus children make sense of new mathematical ideas as they are defined, presented and tested solely in classrooms (Schoenfeld, 2007).

The second reason curricula are attractive templates from which to build an international reference list has to do with details and sequencing. Many curriculum documents include detailed descriptions of knowledge and skills for teachers and test designers to construct appropriate instructional plans and assessment tools. And all of this detail is conveniently organized in developmental sequence.

The third reason has to do with broad similarities in the their construction. Again, national mathematics content details and sequencing differences notwithstanding, striking parallels exist among many jurisdictional expressions for mathematics. This is likely due

to the fact that, other OtL factors being equal and regardless of social or cultural makeup, human beings tend to develop mathematical expertise similarly. And it may partly explain why counting with natural numbers is easier to learn and precedes more difficult algebraic tasks. Broad similarity provides a common basis from which different curriculum expressions can be organized.

The fourth reason has to do with the scope of many curriculum documents. Curriculum designers are mainly concerned with how children appropriate mathematical ideas and are generally less concerned about the logical structure inherent in mathematics as a discipline. Thus we develop plans for students' *to experience ideas* where the term "experience" indicates a much broader vision than any topic list can encompass (Remillard & Heck, 2014). For example, curriculum documents often include such things as role playing, practice with summarizing, comparing different representational forms, and communication. Curriculum, therefore, involves a much wider scope than structural components of a particular subject matter. And this has important ramifications with respect to assessment because testable tasks are restricted to those which evoke estimable demonstrations (Pellegrino et al., 2001). Thus curricula, by definition, also encompass far more than their associated test regimes.

Assessment Frameworks

In many jurisdictions teaching and learning intentions have shifted from a focus on inputs associated with educational access and moved toward measuring specific educational outcomes (Tedesco et al., 2013). Increasingly outcomes are described in terms of specific competencies students' are required to demonstrate facility with before we can assume they are fully functioning members of modern society. In other words, there has been an inexorable shift toward results of testing to provide evidence concerning educational effectiveness.

Just as curriculum documents can be likened to educational intentions (i.e., attributes of the intended curriculum), assessment frameworks (whether national or international) can be likened to testing intentions (i.e., attributes of the intended test). And analogously, the intended test differs from the implemented test (i.e., what items are actually administered) and from the achieved test (i.e., students' response scores). Assessment frameworks are outlines, therefore, that provide examples of task-types deemed to be of sufficient jurisdictional importance to warrant testing. And it follows that assessment frameworks are restricted in scope to include only estimable intentions (i.e., subsets of related curriculum documents).

Theory, Curriculum and Assessment as a Single Model

Addressing the two goals of this project was accomplished by combining theory, curriculum and test information together in a single model. As noted, cognitive models for learning and doing mathematics tend to describe mathematical process. Niss and Højgaard's (2011) work stands apart because they link process with mathematics content domains in a developmental sequence. Their model, however, lacks sufficient detail and this is primarily why it is necessary to involve curriculum. So it is the combination of cognitive theory and curriculum that establishes the *Reference List* half of the RL&CS model (see Figure 1). Mapping NAFs to the Reference List (the second half of the model) is then facilitated by the creation of a coding scheme which acts as a bridge between the two model halves.

Modelling theory, curriculum and assessment together, however, raises important questions about integrity. First, there is the matter of how closely we can expect theoretical depictions of how people mathematize and detailed curriculum expectations will actually map. On the one hand we expect there to be inter-jurisdictional differences in curriculum content and sequencing, while on the other hand, they all map to just one theoretical description. Although limiting different curriculum expressions to fit within a single cognitive interpretation is potentially problematic it is arguably mitigated if theory is primarily used as a tool to organize details of various curriculum expressions–rather than provide a "true" or preferred structure–and, perhaps more importantly, a cognitive theory grounds the final model in literature. Mathematical domains in Niss and Højgaard's work provide roots for a decision tree that, with input from various curriculum documents, is successively refined, branching out to include more and more detailed sub-categories. So the power of this model lies in its ability to faithfully capture and organize *details* of various curriculum expressions.

Secondly, many countries include details and examples of what is important to teach and learn organized year-by-year in a sequence. These details often outline the kinds of experiences nations want students to have when engaging mathematical tasks. Many of these expectations are not tested or cannot be tested in subsequent national or international tests, however, so it is not immediately clear how non-overlapping details should be interpreted. Arguably, if a theory–curriculum decision tree faithfully captures something of the scope and complexity of learning expectations presented in curriculum it does so by faithfully reflecting differences and similarities whether they are tested or not. So the result is a detailed trans-national collection of learning expectations. It follows, if a particular document includes unique learning expectations not observed elsewhere, a new branch is added to the final decision tree. Or if a curriculum document includes the same or very similar expectations to others they are subsumed under the existing model framework. Hence the decision tree evolves in scope and depth to be more and more comprehensive as new curriculum data is incorporated. And the result is a comprehensive inter-jurisdictional description of mathematics education, broader than any single national expression.

Third, an assessment coding scheme should be based on the final Reference List. This suggests the need for consistency in the way we deal with expectations that address the same underlying concepts. And such consistency is arguably accomplished by dividing curriculum-based learning expectations into ACTION:TARGET pairs. An expectation such as *calculate the mean, median and mode of a data set* can be translated into a mathematical **ACTION**: *compute central tendency* and associated with a specific **TARGET**: *mean, median, mode*. This allows for different curricular interpretations to be consolidated into the same ACTION yet different TARGETs. Some curriculum documents, for example, may require students to compute central tendency using only the mean while others may require students to compute mode and mean. Identifying variable TARGET components under the same ACTIONs allows for a way to qualify inter-jurisdictional differences in the scope of particular learning expectations. It also lays the foundation of a quantifiable scale that can be used to compare jurisdictional variation with respect to particular ACTIONs.

Fourth, national assessment frameworks typically provide outlines of desirable item types. Often desirable item types can be translated into a required ACTION (e.g., add two numbers) and related scope or TARGET (e.g., up to a combined sum of 99). Similarity between curriculum- and assessment-based ACTION:TARGET pairs is facilitated by a coding scheme which provides a way of aligning the two halves of the model. In situations where this breaks down (e.g., item types are obtusely described) it may be necessary to establish and track interpretive protocols.

Lastly, since the scope of any jurisdictional assessment framework is, by definition, a subset of the content and skills coverage of the same jurisdiction's curriculum it follows that details in a final Reference List can be pragmatically thought of as the universe of ACTION:TARGET curriculum pairs. Thus details of NAF mappings can be regarded as samples of ACTION:TARGET test pairs. In other words, test pairs taken from any particular NAF are a subset of the total number of available test pairs. Moreover, since samples of NAF item type ACTION:TARGET pairs map onto corresponding ACTION:TARGET curriculum pairs via a coding scheme, the two model halves are intimately linked. Viewing the entire model invites questions such as how particular NAF item types may be distributed across the universe of curriculum-based learning expectations? Or, how do respective countries, when sorted along demographic dimensions like socio-economic status or gender, compare when considering estimable curriculum-based ACTION:TARGETs? The final RL&CS model, therefore, may be a suitable and promising diagnostic device which can be used to provide more nuanced information than tools which rely solely on quantitative metrics can possibly provide.

Method: Creating the RL&CS Model

Niss and Højgaard's (2011) mathematics domains for the first eight years of formal schooling served as a foundation. The final RL&CS model was then generated in two steps: (1) development of a final Reference List; and (2) development of a coding scheme to map NAFs. Each is briefly described. The completed RL&CS coding scheme protocol was then used to test model integrity against a set of new English-language NAFs.

Reference List

With Niss and Højgaard's (2011) model as a basis, the Reference List was produced in reductive phases: (1) Reference List data organized by language root (i.e., English, French, Spanish) and year (i.e., 1 to 8); (2) Reference List data organized by language root and content domain (e.g., Number Knowledge, Geometry); and (3) Reference List organized across language roots and content domains. Phases 1 and 2, when completed, served as data for the development of phases 2 and 3 respectively.

Phase 1: Curriculum data input by language root and year. English-, French- and Spanish-language curriculum documents were transcribed into a common framework and organized by year (years 1 to 8). English-language curricula included documents from Canada (Ontario), US Common Core Standards, Ghana, United Kingdom and Barbados. French-language curricula included documents from Canada (Quebec) and Democratic Republic of Congo and Spanish-language curricula included documents from Chile, Argentina and Guatemala.

A five-level framework: DOMAIN, SUB-DOMAIN, CONSTRUCT, SUB-CONSTRUCT, ACTION:TARGET was selected from theory and various curricula to organize learning expectations. Content DOMAINs are the broadest structural component (e.g., Geometry, Algebra) and there was little disagreement about domain-level topics across curriculum documents. SUB-DOMAINs represent broad content structures that exist within DOMAINs. CONSTRUCTs and SUB-CONSTRUCTs, meanwhile, represent the broadest mathematical content structures that exist within SUB-DOMAINs and CONSTRUCTs respectively. As a general rule, more detailed levels of the final framework were associated with more latitude in jurisdictional interpretation and the greatest differences.

Few national curriculum documents were explicitly designed to fit the five-level framework as described. Although there was broad agreement about domains, it was sometimes necessary to change different, ambiguous or very broad curricular expressions to fit a structure that was different than was originally intended. Some curriculum documents place ratio and proportion under DOMAIN:Measurement, for example, and not under DOMAIN:Algebra. Domain affiliation in this case does not alter the nature or details associated with ACTION:TARGET pairs so its reorganization to fit the final framework was regarded as a pragmatic detail with little consequence. It was also sometimes necessary to include learning expectations that, when translated verbatim, represent amalgamations of various components in the five-level framework. Since details of ACTION:TARGET pairs were preserved, these situations were also regarded to have little pragmatic consequence. So the purpose of the first phase was primarily (1) to organize details of various learning expectations onto a common, if arbitrary, organizational framework; (2) to identify potential translation problems from curriculum to the Reference List; and (3) establish protocols to address emerging problems.

Phase 2: Curriculum data input by language root and domain. Using a constant comparison approach, the three Reference Lists by language root and year from Phase 1 were adapted to reflect language root by content domains (Glaser, 1965). This adaptation shifts focus away from a year-by-year depiction of development typical of school organization to emphasize how learning expectations evolve within content domains. As before, DOMAIN to CONSTRUCT levels remained relatively stable (i.e., there were few instances where agreement was impossible); more variability was evident among SUB-CONSTRUCTS and below. ACTION:TARGET pairs, meanwhile, were arranged roughly in developmental sequence (e.g., skip-counting by 2's precedes skip-counting by 3's, 4's and 5's) in their respective SUB-CONSTRUCTs.

Three reorganized frameworks, each representing a different language root, emerged. It was sometimes necessary to interpret differences within respective language–domain frameworks to preserve national curriculum expressions. Spanish language documents, for example, sometimes integrated mathematical process ideas into specific domain structures of the framework. English-language frameworks, meanwhile, tended to keep process and content and skills descriptions quite separate.

Phase 3: Reference List data input across language roots and domains. The last stage involved condensing Reference Lists developed in stage 2 to a single representative list using the same constant comparison method (Glaser, 1965). As before, naming conventions for DOMAINS to SUB-CONSTRUCTS were

normalized to fit a common framework. And, as before, the scope of SUB-CONSTRUCTS increased to accommodate differences in language-based curriculum expressions. ACTION:TARGET pairs were modified where possible to represent the same or very similar learning expectations between Reference Lists developed in stage 2. Unique ACTION:TARGET pairs encountered in respective language-root documents were preserved in the final framework.

The final model, showing only DOMAIN-level correspondences, is illustrated in Figure 2. The Niss and Højgaard (2011) cognitive model includes nine distinct DOMAINS for mathematics whereas the curriculum-based Reference List includes only six. Agreement between models is accomplished by equating similar domains. Niss and Højgaard's *Number Domains* and *Arithmetic* are subsumed under the more general curriculum domain designation *Number Knowledge*. Similarly, Niss and Højgaard's *Probability* and *Statistics*, and *Algebra* and *Functions* are equated with Reference List designations *Statistics and Probability* and *Algebra* respectively. These differences arise from disagreement about how the structure of mathematics should be organized and interpreted. While Niss and Højgaard argued that domains should follow the organization of structures in mathematics as a discipline (e.g., algebra, geometry), curriculum documents are often organized around functional groupings (e.g., measurement, space and shape) that are more closely associated with the complexities of teaching and learning.

An illustration of organizational disagreement is evident when we consider the DOMAINs which stand alone in the final model (Figure 2). *Measurement* appears without a corresponding cognitive domain and *Discrete Mathematics* stands alone without a corresponding curriculum domain. Measurement is regarded by Niss and Højgaard as an example of a functional domain and has no designation in the structure of mathematics. This does not mean that measurement tasks are not important to teach but rather than they are not part of the formal structure of the discipline. Similarly, discrete mathematics is rarely taught in Elementary school as a coherent topic under that name. Despite this, students still engage in discrete mathematics activities when they plot linear functions or collect and analyze survey data under other domains.

The kinds of disagreements about the Reference List illustrate how unimportant its final structure was in terms of characterizing, say, an international curricular view. From the standpoint of the final product, it was far more important that its organizational form provided a clear way to logically sort and categorize details of ACTION:TARGET pairs.

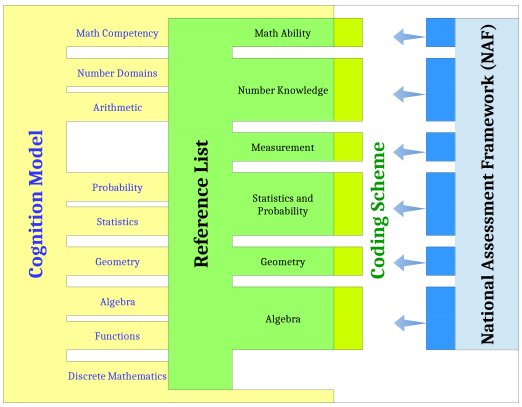

Figure 2. The *Reference List & Coding Scheme* (RL&CS) Model for Mathematics (years 1 through 8)

Coding scheme to map national assessment frameworks

The coding scheme was created using the same DOMAIN to SUB-CONSTRUCT framework developed for the Reference List. It was intended to guide a naive coder to locate specific NAF item-types onto the curriculum side of the model. A coder locates the appropriate DOMAIN to SUB-CONSTRUCT branch and compares the NAF item type to a list of ACTION:TARGET curriculum pairs. This positions the item type in the RL&CS model. If an item type had no match among ACTION:TARGET curriculum pairs, coders were asked to make note so that the Reference List could be subsequently modified. If an item type was ambiguously associated with more than one Reference List structure (e.g., associated with adding and subtracting natural numbers and, say, linear measures) coders were encouraged to duplicate the item type description to appear in all appropriate Reference List structures.

A total of 79 English-language mathematics NAFs were successfully mapped onto the Reference List by independent coders (Siakalli & Vaverek, 2017). Although many of the challenges noted above were iden-

tified there was general agreement that details in the RL&CS model contributed a great deal to its usefulness as an organizational tool. Indeed, coders encountered no NAFs that defied attempts to map.

Discussion

A critical appraisal of the methodology used to derive the RL&CS model is briefly sketched above. It was mainly concerned with a piece-wise discussion of model components (e.g., cognitive theory, curriculum), however, and not with robustness or utility of the model as a cohesive thing. This section addresses the model as-a-whole, opening with discussion about robustness of the final framework, followed by brief descriptions about how the model can be used, some implications and foreseeable challenges.

Robustness of the Model

The final framework was developed using qualitative research techniques primarily to ground results both in theory and in practice. A theoretical basis arguably stabilizes the final framework by providing a warrant for particular domains generally believed to be important mathematical components during the first 8 years of formal schooling (OECD, 2013; Schoenfeld, 2007; Niss & Højgaard, 2011). This anchors final model domains to existing literature. It also means that advances in our understanding of mathematics cognition can be easily incorporated to adjust things. Various curriculum expressions, by comparison, provide a modicum of flexibility. For particular curriculum expressions are informed by teaching and learning factors as much as they are by structural components of mathematics as a discipline. Inter-jurisdictional differences in perspectives, task details and sequencing introduces uncertainty which manifests as a churning effect mostly at the framework extremes. So the Reference List side of the RL&CS model reflects the structured formalism of mathematics as a discipline at the same time that it describes details of the relative informalism of mathematics as it is practised in schools. Both are necessary in the pursuit of UNESCO's *Education 2030* goals.

As mentioned, composition of the DOMAIN to SUB-CONSTRUCT framework is intended to be an organizing tool; one that sorts various learning expectations turned ACTION:TARGET pairs that are encountered in curriculum into a series of logical categories. But as this framework is itself derived from specific curriculum documents, any logical categories it defines are, in a sense, arbitrary. That is to say a different collection of curriculum documents may have resulted in derivation of a different framework. So the robustness of the RL side of the model rests principally on the trustworthiness of the final population of ACTION:TARGET pairs. Trouble is, learning expectations expressed as ACTION:TARGET pairs vary by jurisdiction so model robustness cannot rely on a finite set of "true" pairs because no such set exists. Instead, we are left with a population of ACTION:TARGET pairs whose trustworthiness amounts to their pragmatic value (Creswell & Poth, 2017). For it is schools that define mathematics for teachers and students through everyday interactions with mathematical tasks. Robustness of ACTION:TARGET pairs in the final model, therefore, amounts to their endorsement by international consensus; because we collectively agree that they are meaningful.

The assessment side of the RL&CS model is also based on the trustworthiness of the final set of ACTION:TARGET curriculum pairs. The coding scheme establishes a bridge between cognitive–curriculum and assessment sides of the model by linking curriculum-based ACTION:TARGET pairs to assessment-based ACTION:TARGET pairs. The assessment side of the model, therefore, amounts to a translation of information from one form to another. Assessment frameworks and, arguably most tests, are limited by learning expectations defined in associated curriculum documents. So what is taught, the limits of what is learned, and the scope of what is tested are circumscribed by the population of curriculum-based ACTION:TARGET pairs. And this is associated with how successfully assessment-based ACTION:TARGET pairs can be mapped. Robustness of the final model, then, depends on the comprehensiveness of the final set of ACTION:TARGET indicators and the robustness of the coding scheme and not on the final structural framework.

Practical Utility of the RL&CS Model

What about the practicality of this model? Utility of the RL&CS model is briefly discussed here in terms of opportunities and challenges. These topics are not

intended to be exhaustive but, rather, indicative of an evolving conversation as the framework is more widely available and different curricular expressions are applied.

Implications of the Framework

Two facets of the final model suggests opportunities to assume theory–curriculum, assessment or combined theory–curriculum and assessment perspectives. These views are complimentary and may facilitate jurisdictional or inter-jurisdictional curriculum content and sequencing comparisons, critical analyses of inter-jurisdictional testing intentions, item-level analyses, comparisons of various test results or any combination of these.

Modelling theory. There is some tension between theoretical and practical accounts about how students learn and do mathematics (e.g., Cobb, 1988). It may be, for example, that discrete mathematics does not appear in most curriculum documents because it is too abstract a notion for school children or that discrete mathematics as a stand-alone topic does not work in practice. A similar argument, though this time in reverse, can be made of measurement. Up to now, defined topics which exist in mathematics as a discipline have been largely isolated from topics which exist in curriculum. Differences between interpretative brands, therefore, may have more to do with the ways they are used. And differences in use is not limited to content domains. Cognitive theories often portray mathematical process elements in a discussion about how people mathematize. Curriculum documents, by contrast, present mathematics process elements variously as global competencies affecting all domains or as competencies specifically associated with particular task types. Reasons for these differences have not been well described or investigated. Yet a better understanding about these matters would undoubtedly inform both theory and practice.

The RL&CS framework, therefore, represents a way to organize theory and curriculum toward a better understanding of their relation. It is not so much that theory and curriculum should agree–constructivist views of learning differ from structural views of mathematics–but that their juxtaposition offers interesting fodder about the nature of the subject matter and how this is associated with practical challenges involved in learning and doing mathematics.

Modelling curriculum. The RL&CS framework can be used to study jurisdictional curriculum expressions in any or all of the identified mathematics domains. Since the cognitive–curriculum half of the model represents a wide selection of ACTION:TARGET pairs, it can be used to investigate distributional behaviour of various jurisdictional curriculum expressions. These can then be compared to NAF item-type distributions and even test outcomes.

It may be, for example, that certain jurisdictions share curriculum expectations but differ in NAF item-type distributions and test outcomes. Or the relative density of ACTION:TARGET pairs by domain between different curriculum expressions may provide important information concerning curriculum reform.

Modelling NAFs. In analogous fashion to curriculum, assessment frameworks (national or international) could also be modelled and compared using the assessment side of the RL&CS framework. Inter-jurisdictional variation in testing intentions may provide valuable information particularly when coupled with curriculum or test outcomes data.

An investigation of distributional overlap between NAF item types and curriculum or the overlap between NAF item types and actual test item-types may inform reform efforts in test design by jurisdiction. Indeed, how closely particular NAF item-types map onto associated curriculum or test design may be indicative of relative educational success.

Modelling item-level details and test outcomes. Just as NAFs can map onto the RL&CS framework using information about item types, if enough is known about actual item characteristics it is also possible to map details of administered tests and link these to analyses of student outcomes. For example, the Australian Council for Educational Research (ACER), in partnership with UIS and UNESCO, is developing a catalogue of items to measure learning outcomes (http://www.acer.org). This entails the compilation of achievement indicators used for elementary and secondary students from among 20 or so countries. Student responses will be analyzed and interpretation and reporting parsed by jurisdiction.

But ACER's set of mathematics items could be translated to ACTION:TARGET pairs and used to extend the RL&CS framework. Hence cognitive–curriculum learning expectations (i.e., *teaching and learning in-*

tentions) and details of NAFs (i.e., *item-types intended for testing*) could be extended to include details about administered tests (i.e, *administered items*) and item-level student responses (i.e., *test results*). Translating mathematics items into ACTION:TARGET indicators and linking these to students' responses introduces new opportunities. ACER could calibrate item selection on distributional characteristics of items when they are mapped onto the model. It would be possible, for example, to quantitatively describe how detailed attributes of items are related to the DOMAIN to SUB-CONSTRUCT components of the model framework. Is there sufficient content and skills coverage? (what counts as sufficient content and skills coverage?). It would also be possible to parse learning outcomes by item types; determining which, if any, item types tend to contribute more to outcomes of the global reporting scale.

An extended RL&CS framework allows for critical analysis of item-level test response data and how they relate to detailed item information and inter-jurisdictional learning expectations. This is important because there are substantial rifts between content and skills addressed by assessment instruments (particularly large-scale tests) and curricula (Pellegrino et al., 2001; Schoenfeld, 2007). Extending the model bridges this divide by introducing a single comparative platform. Details of item-level content and skills used to construct measures and item-level student outcomes gives information about test composition and relevance but also opens interesting possibilities for secondary analyses. It would be possible, for example, to conduct item-level latent trait analyses of student responses. Using this approach, we could model differences among students', say, geometry responses against factors such as gender or residence type (e.g., urban vs. rural).

Using the ACER reporting scale to extend the RL&CS model could be an exemplar for a more ambitious extension. For as long as item-level details are accompanied by student responses, any assessment instrument can be used to extend the framework. For example, governments who wish to learn more from their own national tests could extend the model to include proprietorial item-level information and use these to conduct a series of secondary analyses (e.g., student responses $\propto$ item format + SUB-DOMAIN item coverage). International agencies could extend the model in an analogous fashion by analyzing publicly available data. ACER's continued involvement in the design and implementation of the model extension allows for continuing development of a suite of tools that agencies and governments could deploy.

Utility of the Framework

The RL&CS framework integrity rests on the pragmatic value of the final set of ACTION:TARGET curriculum pairs. In other words, detailed ACTION:TARGET indicators are trustworthy only insofar as they continue to have value to stakeholders. Utility of the framework, therefore, is inexorably tied to its use. Three possible stakeholder groups are discussed.

National agencies and governments. National educational authorities can use the final framework to investigate effectiveness of curriculum-based learning expectations, testing intentions and, where appropriate, test results against other jurisdictional learning expectations, testing intentions and test results. The coding scheme is relatively easily learned and applied to jurisdictions with differences in curriculum, NAFs and test outcomes. Making the final framework available to a variety of national agencies and governments, however, presents a significant accessibility problem. For, on the one hand, integrity of the framework needs to be safeguarded while, on the other hand, it should be readily available.

An interactive web-based interface, if carefully maintained, could serve both integrity and accessibility functions. It could act as a source of information about the RL&CS model, provide information about, and opportunities to practice using the coding scheme, serve as a repository for inter/national results, and serve analysis templates to agencies and governments wishing to examine their own data. Moreover, it could function as a simple database, a repository of data to be downloaded as well as a simple tool to upload national data. A web-based interface could be maintained on an ongoing basis ensuring ACTION:TARGET indicator integrity while, at the same time, providing an accessible, modifiable and extendable platform. Above all, it would provide a cost-effective way to encourage international participation in realizing the SDG-4 goal.

International agencies. As with national authorities, a web-based interactive RL&CS client could also serve international agencies. In fact, there is no reason

why the front-end of a web-based interface could not simultaneously serve both national and international interests. Functional separation becomes imperative only when it is important to differentiate data derived from countries or agencies not wishing to share data from those that do. In such cases, separation would likely be guided by some kind of ethical protocol to manage informed consent requests.

Web-based data gathering, analysis and reporting protocols established for international and national users will define how the framework is used now and in future. Unlike paper documents, a web platform can be continuously revised as needs change. But similar to paper documents, a web platform–because it requires users to interact in particular ways–defines how the model is used. So while it would be possible to standardize model elements to ensure that, say, inter-jurisdictional comparisons are made more meaningful, standardization influences utility and, ultimately, integrity of the model.

Challenges

Just as trustworthiness of the final framework is tied to its use, threats to integrity are arguably tied to pragmatic issues of model design and use. For no single set of curriculum-based ACTION:TARGET indicators can objectively capture the diversity of worldwide teaching, learning and testing intentions. And this remains true regardless of whether or not the model is extended to include details of items administered to students or extended to include actual test results. So threats to integrity constantly change and, because they link pragmatics of data use with educational practice, evolve as new theoretical and jurisdictional interpretations about mathematics emerge.

RL&CS framework integrity is contingent on two main model components: Trustworthiness of ACTION:TARGET indicators; and integrity of the coding scheme linking the two halves of the final framework.

Trustworthiness of ACTION:TARGET indicators. Trustworthiness hinges on the extent to which ACTION:TARGET pairs adequately represent the scope and depth of international learning expectations. If this set is representative it can be treated as an indicator universe for the purposes of downstream model analysis and reporting. And we can use results to inform educational decisions with the assurance that they reflect a current understanding about mathematics curriculum. But the opposite is more likely the case. ACTION:TARGET indicators will not reflect adequate scope and depth in all respects. So it will be necessary to determine the point at which they are sufficiently adequate. The question, therefore, comes down to how we determine and maintain the final set of ACTION:TARGET indicators to be *reasonably representative* (Yarbrough, Shula, Hopson & Caruthers, 2010).

Complicating matters, mathematics curriculum documents continuously change so we can predict with certainty that representativeness of any final set of ACTION:TARGET indicators will drift over time. Mitigating this threat requires on-going work to refine and redefine such things as policies, decision-making, stakeholder roles, documents, curriculum-based descriptive language and content/skills details and test information. Stakeholders, both old and new, will provide valuable input about what is changing in mathematics education. New curriculum documents can be periodically mapped to ensure the model remains flexible enough to capture a broad diversity of jurisdictional expressions; previously mapped curriculum documents can be re-mapped. Changes in our understanding about assessment, test design, administration and reporting influence the selection of testable item types. Mitigating factors such as these can be included in a comprehensive, and ongoing, program evaluation approach (Yarbrough et al., 2010).

Arguably the greater the number of curriculum documents we map the more resilient and stable the model will become. As new information is mapped onto the final framework, misfitting ACTION:TARGET indicators become more apparent resulting in possible changes in structural components of the model framework itself. That said, however, pragmatic decisions have to be made as to how many (both old and new) curriculum documents are reasonable to include and with what periodicity (e.g., annually, semi-annually) they should be included. Outcomes of such pragmatic decisions define a reasonable accommodation threshold that, in turn, also influences model integrity.

Integrity of the coding scheme. Model integrity also depends on the effectiveness of the coding scheme. This is where structural components (i.e., DOMAIN ... SUB-CONSTRUCT) of the final framework, despite being essentially arbitrary, become important. Coders

rely on the framework structure to map assessment ACTION:TARGET pairs from various NAFs onto the model. If this structure is ambiguous or incomplete it is more likely to result in coding disagreements and errors. So fidelity of structural components turns out to have pragmatic relevance to model design.

Mitigating coding difficulties requires on-going support. Revised coding protocols resulting from changes to details of ACTION:TARGET pairs, new coders and new NAFs all conspire to challenge integrity. Similar to difficulty with ensuring reasonable representation of ACTION:TARGET curriculum outcomes, increasing the number of NAFs mapped to the model, and tracking anomalies and omissions, amounts to an integrity check. Provided the final model is a reasonable representation of international mathematics curriculum expressions, mapping should proceed with relatively few anomalies or omissions.

Conclusion

The methodology used to create a Reference List & Coding Scheme framework described here is a step toward establishing a complex interpretative environment from which to base educational decisions. It is ambitious in the sense that it encompasses learning expectations representative of three language roots yet not ambitious enough because it only encompasses three language roots. Likewise, it is flexible and extensible enough to reflect intended learning expectations, intended item types for assessment, details of administered items, and student response data. Yet it is not flexible or extensible enough to adequately capture more complexity in students' opportunities to learn. Moreover, as framework utility is for all intents and purposes a pragmatic property that is subject to drift, a great deal remains to be learned about effectively using it as we move forward. For our understanding about how students learn and do mathematics is still rudimentary so while this framework captures something of complexity it does not capture nearly enough. Clearly more work is needed.

References

Carpenter, T. P. & Moser, J. M. (1983). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*(3), 179–202.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record*, *74*(8), 723–733.

Choi, N. & Chang, M. (2011). Interplay among school climate, gender, attitude toward mathematics, and mathematics performance of middle school students. *Middle Grades Research Journal*, *6*(1), 15–29.

Chui, M. M., Chow, B. W.-Y. & McBride-Chang, C. (2007). Universals and specifics in learning strategies: explaining adolescent mathematics, science, and reading achievement across 34 countries. *Learning and Individual Differences*, *17*(4), 344–365.

Cobb, P. (1988). The tension between theories of learning and instruction in mathematics education. *Educational psychologist*, *23*(2), 87–103.

Creswell, J. W. & Poth, C. N. (2017). *Qualitative inquiry and research design: choosing among five approaches*. Sage publications.

Cross, D. I. (2009). Creating optimal mathematics learning environments: combining argumentation and writing to enhance achievement. *International Journal of Science and Mathematics*, *7*(5), 905–930.

Cunningham, M. (2012). *Student learning heterogeneity in school mathematics* (Doctoral dissertation, University of Toronto).

Geary, D. C. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, *37*(1), 4–15.

Glaser, B. G. (1965). The constant comparative method of qualitative analysis. *Social Problems*, *12*(4), 436–445.

Haertel, E. (1981). Developing a discrete ability profile model for mathematics attainment: Final report. *ERIC*.

Haertel, E. H. & Wiley, D. E. (1993). Representations of ability structures: implications for testing. In R. J. M. N. Frederiksen & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359–384). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hallet, D., Nunes, T. & Bryant, P. (2010). Individual differences in conceptual and procedural knowl-

edge when learning fractions. *Journal of Educational Psychology*, *102*(2), 395–406.

Jensen, T. H. (2007). Assessing mathematical modelling competency. *2007). Mathematical Modeling (ICTMA 12): Education, Engineering and Economics*, 141–148.

Meyer, M. L., Salimpoor, V. N., Wu, S. S., Geary, D. C. & Menon, V. (2010). Differential contribution of specific working memory components to mathematics achievement in 2nd and 3rd graders. *Learning and Individual Differences*, *20*(2), 101–109.

Niss, M. & Højgaard, T. (2011). *Competencies and mathematical learning: ideas and inspiration for the development of mathematics teaching and learning in Denmark* (No. 485–2011). P.O. Box 260, Roskilde DK: Roskilde University, Department of Science, Systems and Models IMFUFA.

OECD. (2013). *Pisa 2012 assessment and analytical framework: mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264190511-en

Pellegrino, J. W., Chudowsky, N. & Glaser, R. (2001). *Knowing what students know: the science and design of educational assessment*. Washington, DC: National Academy Press.

Pratt, D. (1994). *Curriculum planning: a handbook for professionals*. Wadsworth Publishing Company.

Remillard, J. T. & Heck, D. J. (2014). Conceptualizing the curriculum enactment process in mathematics education. *ZDM*, *46*(5), 705–718.

Schmidt, W. H., Cogan, L. S., Houang, R. T. & McKnight, C. C. (2011). Content coverage differences across districts/states: a persisting challenge for u.s. education policy. *American Journal of Education*, *117*(3), 399–427.

Schoenfeld, A. H. (2007). What is mathematical proficiency and how can it be assessed? In A. H. Schoenfeld (Ed.), *Assessing mathematical proficiency* (Vol. 53, pp. 59–73). Mathematical Sciences Research Institute Publications. New York, NY: Cambridge University Press.

Siakalli, I. & Vaverek, C. (2017). *Summary report for mathematics national assessment frameworks: content reference list and coding scheme*. Geneva, Switzerland: UNESCO International Bureau of Education. Retrieved from www.ibe.unesco.org

Tedesco, J. C., Opertti, R. & Amadio, M. (2013). The curriculum debate: why it is important today. ibe working papers on curriculum issues no. 10. *UNESCO International Bureau of Education*.

Windfield, L. F. (1987). Teachers' estimates of test content covered in class and first-grade students' reading achievement. *Elementary School Journal*, *87*(4), 438–445.

Yarbrough, D., Shula, L., Hopson, R. & Caruthers, F. (2010). *The program evaluation standards: a guide for evaluators and evaluation users* (3rd Edition). Thousand Oaks, CA: Corwin Press.