# Conceptualising the science curriculum - forty years of developing assessment frameworks in three large-scale assessments

PER MORTEN KIND
School of Education, Durham University, Durham DH1 1TA, UK

Published in Science Education 97:671–694, 2013

## Abstract

The paper analyses conceptualisations in the science frameworks in three large-scale assessments, TIMSS, PISA and NAEP. The assessments have a shared history, but have developed different conceptualisations. The paper asks *how* and *why* the frameworks are different, and seeks answers by studying their development. The methodology is document analysis by, first, tracing developments within each assessment, next, comparing developments and conceptualisation across the assessments, and last, relating the frameworks to trends of developments in educational theory. The outcome of the analysis provides a complex picture with the assessments following their own lines of development but with influence from trends in assessment and educational theory. Five main conceptualisations are found to have existed over time, with different definition of *scientific behaviour* and explanations to the relationship between knowledge and behaviour. The frameworkshave moved towards more elaborated *explanations* of the science domain, providing assessors with better support for operationalising learning objectives. Currently, the assessments are faced with a challenge of adapting to the "practice turn" in science studies and learning science and thereby accounting for scientific behaviour as a *community practice*. The paper concludes with suggestions for how frameworks may be improved to achieve this aim.

## Introduction

All fields need a common conceptualisation making communication possible without necessarily operating on a globally shared theory (Gruber, 1993). The current paper analyses assessment frameworks in large-scale assessments and uses the perspective that these provide *"an explicit specification of a conceptualization*" (ibid: 199). The context of the analysis is the importance large-scale assessments have had for educational policy, curriculum development and educational practices; which suggest *their* conceptualisations of the subject domain are likely to be adopted by many educators. In other words, that the *formal conceptualisation* developed in assessments projects are likely to shape and influence the conceptualisation used in the field more generally. The contention of the paper, however, is that this influence relates not just to the status of the large-scale assessments, but also to the format of the frameworks. The frameworks are similar to standards documents or curriculum guidelines setting out goals for an educational system, because of defining the same curriculum domain, but use a classification format that deviates from the listing of learning objectives typically found in these.  They organise the subject domain into *categories* based on some *organising principle* (Moseley et al., 2005). The format is used, partly, because the assessments operate across nations and/or curricula and therefore have to identify more general principles and structures, and, partly, because it supports assessors operationalising the subject domain into items and scoring rubrics. The outcome, however, is bringing out conceptual structures of the domain in a way that "provides support for thinking" (ibid: 34) and "facilitate the mental representation of a field". (ibid: 39). As will be shown in the analysis, the frameworks have developed over time towards more advanced structures and thereby are becoming more useful and relevant for educators. The influence may be positive or negative, but that it occurs at all justifies examining the frameworks critically to see *what* conceptualisation they offer. From this perspective, the paper presents an analysis of science frameworks used in three large-scale assessments: the *Trends in Mathematics and Science Study* (TIMSS), the *National Assessment of Educational Progress* (NAEP) and the *Programme for International Student Assessment* (PISA).

The reason for aiming the analysis particularly towards the TIMSS, NAEP and PISA is their dominance as "world-class standards" (DeBoer, 2011; Linn & Baker, 1995), but also their long and shared tradition for developing assessment frameworks. All three exhibit similar processes in their development, characterised by continuous revision of the same framework over many years and the involvement of academic experts across science education, science, learning psychology, assessment and policymakers. The *International Association for the Evaluation of Educational Achievement* (IEA), the provider of TIMSS, was established in the 1960s and developed a science framework to the *First International Science Study* (FISS) in 1970/71 (Comber & Keeves, 1973; Husen & Postlewaite, 1996). This framework has since been revised and rewritten many times, leading to the current TIMSS framework (Mullis, Martin, Ruddock, O'Sullivan, & Preuschoff, 2009). NAEP started at the same time as the IEA studies and published its most recent science framework in 2008 (NAGB, 2008). These two projects have had mutual influence due to their committees including some experts in common and from using benchmarking studies to compare contents (e.g. Neidorf, Binkley, & Stephens, 2006; Nohara & Goldstein, 2001; Wu, 2010). The first PISA survey was conducted in year 2000, so this project has a shorter history but fits the same overall pattern. Although other assessments could be mentioned, TIMSS, NAEP and PISA have been the most continuous and trend-setting large-scale assessments, suggesting their science frameworks are among the most thoroughly developed conceptualisation of the science domain.

The paper, however, contends that, despite their effort, the three large-scale assessments struggle to set an appropriate conceptualisation for the science learning domain. Difficulties are observed, firstly, in variation in concepts and organising principles between current versions of the frameworks (Mullis, et al., 2009; NAGB, 2008; OECD, 2006), which is surprising considering their mutual influence. Secondly, looking back on previous versions shows all three projects have changed substantially, seemingly struggling to find the “right” way of conceptualising the learning domain. Contrasting these difficulties, are researchers failing to recognise that differences exist. Nohara and Goldstein (2001), for example, when comparing TIMSS, NAEP, and PISA, notice that different “dimensions” are used and that these do not correspond well, but choose to ignore this in their analysis. These authors instead select one particular framework’s conceptualisation to analyse the content of other assessments.  This approach, which is common in benchmarking studies, serves to disguise rather than confront differences.

The paper, therefore, has two purposes. The first is to achieve better understanding of *how* and *why* assessment frameworks in the three projects are different. This understanding draws on backgrounds to the frameworks and how these have developed over time. Second, the paper evaluates the frameworks for appropriateness in guiding and informing science assessors and educators about the science learning domain. Criteria for the evaluation exploit work that informed Bloom, Engelhart, Furst, Hill, & Krathwohl (1956) when developing their taxonomy for the cognitive domain. They established three principles for a framework to satisfy: first, to use a conceptualisation that is relevant to and communicates with teachers, curriculum developers and assessors; second, to be logical and internally consistent; and third, to be consistent with current perspectives in educational theory. The second criterion means establishing terms and definitions for consistent use in different areas of the learning domain, and having an *organising principle* that provides a meaningful relationship between dimensions and categories. Working in the 1950s led to Bloom et al. anchoring their third criterion mainly in cognitive psychology, the dominant field of the day. They wanted to avoid psychologically untenable distinctions that teachers made regularly and to include distinctions important psychologically but not frequently considered by teachers. More generally, of course, the conceptualisation in the framework needs to reflect and support general educational theories and ideas of the era.

Bloom et al. (1956) prioritised these principles in the order presented above, i.e. paying more attention to communication with teachers than aligning the framework with educational theories. This paper reverses the order, because of difficulties science educators have experienced in implementing findings and new ideas from educational research into practice. Since assessment frameworks were first introduced, science education research has been through “paradigm shifts” relating to developments in *learning sciences* and *science studies*. A key issue explored is therefore to what degree and in what ways large-scale assessment frameworks have been aligned with and actively helped promote new ideas in science education. However, none of Bloom et al.'s three criteria can be ignored, so the challenge framework committees have faced is to establish assessment frameworks that give a plausible answer to them all.

As a context and reference, the next section of the paper summarises key theoretical developments in science education occurring during the period the frameworks developed. Thereafter, the paper outlines the methodology used, followed by the outcomes of the analysis. The final section points to possible future developments and suggests specific changes to the organisation of assessment frameworks in science education.

## From science processes to science practices

Three successive trends in science education research and practice have dominated science education since the large-scale assessment projects began. First, influence from *cognitive psychology*. This emerged in the late 1950s as a counter-reaction to *behaviourism,* defining achievement in terms of *mental structures* and *processes* acquired by students (Miller, 2003). Many processes were seen as "liberated from particular contents" (Inhelder & Piaget, 1958: 331) and therefore transferable between contexts. Gagne (1965), amongst others, used this rationale to suggest some *science processes*, that is, mental processes used by scientists when generating and testing theories, could be trained in education and applied to other life domains. Applied as *process science* this view gained popularity in science education research and practice far into the 1980s.

During the 1980s "process science" was criticised (e.g. Finley, 1983; Millar & Driver, 1987) for the implicit science philosophy and the learning psychology it promoted. Finely (1983), for example, claimed Gagne's conceptualisation reflected a *logical-empiricist view* of science that contemporary philosophers had since abandoned. He quoted Quine (1969) and other philosophers in noting meaningful observations occur only in the context of a conceptual scheme, in other words, that all observations are *theory-laden*. Simultaneously, research demonstrated that prior knowledge students bring to science learning was more important to learning outcomes than their ability to use cognitive operations (Driver & Easley, 1978; Osborne & Wittrock, 1985). The same criticism was raised outside science education by Brown and Desforges (1977) and Donaldson (1984) who showed that children's ability to reason depends on their understanding of content and contexts as well as their cognitive capabilities. Critics called for alternative approaches to explain both "learning" and "science".

A second trend developed in response, in which psychologists and science educators turned towards a "Kuhnian" view of science. This proposed that a person's science understanding develops gradually and, occasionally, revolutionises from simple to more advanced ideas. Learning, it was argued, is *less* about logical and abstract thinking and *more* about developing and understanding domain-specific knowledge. The trend became known as *conceptual chang*e (Hewson, 1981; Osborne, 1982; Osborne & Wittrock, 1985) and its underpinning philosophy related to *personal constructivism*. The links between cognitive psychology and science education were still strong, but attention in both areas moved towards content-rich and domain-specific environments rather than domain-general thinking (McCloskey, 1983).

A third and current trend developed somewhat later as an alternative answer to the criticism of process science and logical-empiricism, building on a socio-cultural view on science and learning. This view of *learning* is inspired by Vygotsky(1978) and Bakhtin(1981), who demonstrated how higher cognitive functions are learned through social interaction (usually in expert company) via a process requiring *communication* and involving *tools* (physical, symbolic, or both). This has shifted focus towards the function of *language* and the importance of playing *roles* in a community (Hughes, Jewson, & Unwin, 2007). The socio-cultural view of *science*, similarly, suggested science is *negotiated* within the scientific community: scientists construct tentative explanatory accounts of nature, drawing on information gathered, and put these forward for debate so the whole community can come to consensus (Kuhn, 1962; Latour, 1987). Combining these, Ford and Forman (2006) refer to the new trend as the "practice turn", and suggest science students learn to play interchangeable roles as *Constructors* and *Critiquers* of scientific claims: "The Constructor floats arguments and the Critiquers publicly identify errors in those arguments, at which point the presenter returns to production work and attempts to remove errors" (p15). More generally, much attention in science education has become directed towards establishing authentic discourse practices in the classroom by engaging students in *scientific argumentation* (Driver, Newton, & Osborne, 2000).

Thus, the last five decades have seen developments that have brought major challenges to the groups and committees responsible for setting assessment frameworks. In the behaviourist era, expectations and foci were on identifying *observable behaviour* avoiding the "murkiness" of the mental world. In the cognitive area, this changed to the opposite position, leading to domain general *mental concepts* and *processes* becoming key foci. Critique led first to a claim that reasoning is knowledge-dependent, and that science learning should be understood as *conceptual change*. Later, the focus on learning and science as *social* phenomena emphasises that an individual must adapt to community standards and learn to play roles as constructors and critiquers. Overall, this means shifts from learning as something occurring exclusively inside people's heads to events played out as practices in society, and similarly, from a "science-as-logic" to a "science-as-practice" conception of science (Lehrer & Schauble, 2006) have occurred.

This presentation is a simplified picture. Trends overlap and are more complex, but the outline serves as a basis for analysing patterns of development in the assessment frameworks. A question arises about how the frameworks have adapted to the changes and, in particular, how current versions of the frameworks have managed to include recent the "practice turn".

## Methodology

The study of the assessment frameworks used a three-stage document analysis. First, the frameworks from each assessment project were analysed separately to identify *how* they have changed over time. This meant placing versions of same framework side-by-side from the earliest versions onwards, looking for changes in organisation and labelling of dimensions and categories. This type of comparison is an established approach in benchmarking studies (e.g. Neidorf et al., 2006; Nohara& Goldstein, 2001; Wu, 2010), but commonly looking for changes in content rather than structure and comparing frameworks across rather within projects. The current study was looking in particular for structural and conceptual changes that affect the conceptual meaning of the domain. The outcome was a pattern of development for each project's framework.

The second stage was looking for explanations, still within each project, for *why* observed changes had occurred. Explanations were identified in the framework documents, but also in other documents either preparing for or discussing the versions of the frameworks. For example, prior to developing the 2009 version of NAEP a panel presented issues and recommendations (Champagne, Bergin, Bybee, Duschl, & Gallagher, 2004), and the 2006 version of PISA was influenced by the DeSeCo project (OECD, 2003a). Committee members have published several research papers explaining and discussing the frameworks. The outcome was "stories" following the pattern identified in stage one, describing the guiding rationales but also problems encountered when developing the frameworks. Recognising the limitation of this approach, of course, is important. No first-hand information, for example, has been gathered from actual meetings and debates between committee members. The outcome, therefore, is an interpretation restricted by the material made public.

Third, the frameworks were compared *between* the three assessment projects and evaluated by using the criteria set out in the introduction of the paper: a) relevance and alignment of the conceptualisation in the frameworks compared to the trends in science education research, b) logical consistency of the conceptualisations and c) their appropriateness for guiding educational practice. The comparison highlighted similarities and differences in the patterns of development of the conceptual structures, but also placed the frameworks into the wider context of developments in learning sciences and science studies.

The outcome was an identification of similarities and differences between the assessment frameworks, leading to a categorisation of different types of “common conceptualisations” for the science domain, and also an evaluation of the appropriateness of each of these.

## Framework development

The analysis presents first the developments of the frameworks. This is the outcome of stage one of the methodology, but including the underpinning explanations from stage two. What is presented is therefore the story behind the current versions of the frameworks in TIMSS, NAEP and PISA, identifying some of the thinking, problems and influences that have led to and shaped the particular concepts and structures used in the current frameworks. Conceptual trends across the three projects will be summarised after the presentation all assessments.

### The TIMSS Science Framework

IEA’s First International Science Study (FISS) was conducted in 1970/71, followed by the second (SISS) in 1984/85 and the third (TIMSS) in 1995. The last combined mathematics and science, leading to TIMSS meaning the *Third International Mathematics and Science Study*, and was repeated (TIMSS R) in 1999. From 2003, TIMSS became the *Trends in Mathematics and Science Study* with surveys arranged every four years (2003, 2007 and 2011).

The first FISS framework introduced a *two-dimensional matrix*, or a *table of specifications*, based on Tyler (1950) in combination with Bloom et al.’s (1956) taxonomy. The principle was to categorise learning objectives into *content*, meaning the specific subject matter to be conveyed, and *behaviour*, explained as, what the student should do with the material. Presenting these in a matrix produced cells combining every behaviour category with every content category, generating a “blueprint” to ensure content validity; that all aspects (cells) were included in the assessment. The matrix also underlined the inseparability of content and behaviour, demonstrating the impossibility of understanding content knowledge without using cognitive behaviours, and vice versa. Bloom's taxonomy added another principle by providing a hierarchical structure of the cognitive behaviour. This helped separate simple from advanced reasoning and therefore could be used as an expression for *cognitive demand*. The framework model set a standard that was used by IEA across all subjects and, as will be shown, has dominated many assessment projects since. The FISS framework included four content categories, *Earth sciences*, *Biology*, C*hemistry* and *Physics*, and four hierarchical behaviour categories, *Functional information*, C*omprehension*, *Application* and *Higher processes* (Comber & Keeves, 1973).

Bloom, Hastings, and Madaus(1971) published the *Handbook on Formative and Summative Evaluation of Student Learning* in the same year as FISS was carried out. This text was based on the Tyler-Bloom behaviour-by-content model and had invited authors to write chapters presenting assessment frameworks for school subjects. The authors, curriculum specialists with training in either a content area or educational psychology, approached their tasks differently (Haertel & Calfee, 1983). Psychologists tended to describe objectives in terms of mental structures, while content specialists looked towards the curriculum structure. Leopold Klopfer, the science chapter author, belonged in the latter group. His chapter started with a statement about inclusion of the "full range of student behaviors which may be sought as outcomes of science instruction in elementary and secondary schools" (Klopfer, 1971: 566); including:

- cognitive categories from Bloom's taxonomy;
- processes of scientific inquiry;
- skills of performing laboratory work;
- students' attitudes towards science; and
- students' orientations to the relationships of science to other aspects of culture and to the individual.

As a move to implement all these Klopfer created a two-dimensional framework, just as in FISS, but including all bullet points as categories in the *behaviour dimension.* This caused three essential problems. Firstly, a “dimensionality problem” arose because very different elements (Bloom’s cognitive categories, laboratory skills and attitudes) were placed in the same dimension; secondly, the hierarchical structure introduced with Bloom’s taxonomy was disturbed and thereby made it more difficult to express cognitive demand; and thirdly, categories *not* fitting neatly into the separation between *content* and *behaviour* were brought in. Examples of the last problem are *scientific inquiry* and *orientation* (i.e. knowing about science). Klopfer’s writing reveals a struggle to decide if either of these categories is “behaviour” or “content”. The final outcome was to place them both in both the behaviour and content dimensions. Despite these problems, Klopfer’s framework was adopted for IEA’s second science study, SISS, (Rosier, 1987), with only *re-labelling* as a significant change. The *Behaviour dimension* became the *Objective dimension*. This reflected Klopfer’s re-interpretation of the original Tyler-Bloom matrix: the new behaviour dimension read as a list of *objectives* rather than Bloom’s interpretation of *cognitive behaviour*. In contrast, however, categories in the behaviour dimension were renamed similarly to Bloom’s terminology, and made to *look* like cognitive behaviours. *Process of scientific inquiry,* for example, was renamed *Processes* and the *Application of scientific knowledge and method* renamed *Application*.

The two next revisions of the IEA science frameworks appear as attempts to solve Klopfer’s three problems. The third, TIMSS 1995, committee (Robitaille et al., 1993) focused on the “dimensionality problem”, resolved by splitting the behaviour dimension into two more coherent dimensions: *Performance Expectations*, which combined Bloom’s cognitive domain and scientific inquiry processes, and *Perspectives*, including *attitudes* and *orientation*. The solution was not ideal, making a complicated three-dimensional matrix (*Performance Expectations*, *Perspectives* and *Content*). The problem of a hierarchical performance dimension was also discussed, but dismissed at the time because the argument that science processes can not be ordered in this way held sway. Klopfer's third problem, that *science inquiry* is both "behaviour" and "content", was left untouched.

The committee responsible for the fourth revision in the TIMSS 2003 framework (Mullis et al., 2003) attacked, and solved, all Klopfer's three problems, however, at a certain cost. The solution, firstly, involved moving *scientific inquiry* out of the matrix to become a separate "overarching" dimension; "[overlapping] all of the fields of science and [having] both content- and skills-based components" (Mullis et al., 2001, p. 69). This alternative, (see below), was adapted from NAEP's 1996-2005 framework (NAGB, 2004). Secondly, the *Perspectives dimension*, with *attitudes and interests*, was excluded from the framework entirely. Together, these two moves re-established a two-dimensional matrix where both dimensions are more uni-dimensional (Klopfer's first problem); re-instated a hierarchical behaviour dimension (Klopfer's second problem); and removed topics which belonged to both dimensions (Klopfer's third problem). The behaviour dimension, labelled *Cognitive Domain*, included three categories simplifying Bloom's taxonomy:

- Factual knowledge,
- Conceptual understanding, and
- Reasoning and analysis.

For the TIMSS 2007 study (Mullis et al., 2005) categories in the behaviour dimension was re-labelled to match the revised version of Bloom's taxonomy Anderson et al. (2001) better, adopting:

- Knowing,
- Applying, and
- Reasoning

In summary, the IEA science framework have moved from and to a two-dimensional matrix based on the Tyler-Bloom model and defining behaviour as cognitive demand. Klopfer "disturbed" this model by attempting to include the "full range of student behaviour". This, however, was unsuccessful and only by classifying *scientific inquiry* separately and excluding attitudes and nature of science has the framework become conceptually coherent and functional.

## The NAEP Science Framework

The US-based NAEP science studies began in 1969/70 and have been repeated ten times over the last forty years. The first few surveys were carried out at irregular intervals, with individual states participating voluntarily. Since 2001 the reauthorization of the *Elementary and Secondary Education Act*, often referred to as *No Child Left Behind*, requires states' participation at grades 4 and 8 every four years in science and reading and mathematics biennially. The *National Assessment Governing Board* (NAGB) holds overall responsibility, while the assessment is carried out by the *National Center for Education Statistics* (NCES). NAEP results are known colloquially as the *Nation's Report Card*.

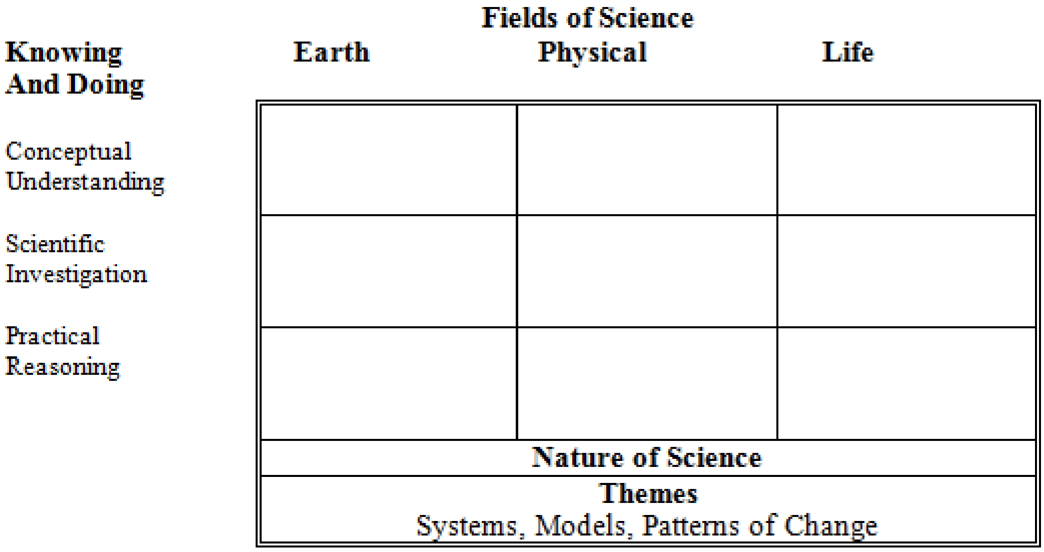
NAEP started with an open listing of objectives, styled like curriculum guidelines. A systematic categorisation developed after a few surveys with "dimensions and categories" similar to those in IEA's frameworks (NAEP, 1979), although not combining dimension in a matrix. Thus, the 1976-77 survey listed three dimensions separately:

- *Content* (the body of science knowledge),
- *Process* (the process by which the body of knowledge comes about)
- *Science and society* (the implications of the body of knowledge for mankind).

In 1981-82, a fourth dimension, *Attitudes*, was added (Hueftle, Rakow, & Welch, 1983). By having an open principle and using four dimensions, the framework omitted conceptual problems described in Klopfer's and IEA's two- and three-dimensional matrices above. However, the conflict between general cognition and scientific inquiry that tainted the IEA study was underlying and became apparent in the 1990 NAEP science framework, when *Process* was renamed *Thinking Skills* and given the three categories:

- Knowing Science,
- Solving Problems,
- Conducting Inquiries.

At this stage, NAEP also adopted a matrix structure like that of IEA, and a series of revisions were made leading to a new framework (Figure 1) in 1996 that was kept unchanged for nearly ten years (NAGB, 2004). This framework had a “content” dimension named *Fields of Science* and a “behaviour” dimension named *Knowing and Doing*. The *Attitudes* dimension from the previous framework was removed and the *Science and Society* became an “overarching” dimension called *Nature of science* outside the two-dimensional matrix. Another overarching dimension called *Themes* was also added.



| | | Fields of Science | |
|---|---|---|---|
| **Knowing And Doing** | **Earth** | **Physical** | **Life** |
| Conceptual Understanding | | | |
| Scientific Investigation | | | |
| Practical Reasoning | | | |
| | **Nature of Science** | | |
| | **Themes**<br>Systems, Models, Patterns of Change | | |

Figure 1: The 1996 – 2005 NAEP Science Framework.

The framework had commonalities with IEA developments occurring at the time, but with some key differences. Firstly, as with Klopfer (1971) and SISS (Rosier, 1987), NAEP extended the behaviour dimension by allowing it to include both (Bloomian) cognitive behaviour and scientific inquiry. This caused a similar re-interpretation of behaviour from what students should do “*to knowledge*” towards a general statement of what they should do “*in science*” (i.e. making it an objectives dimension rather than a classification of cognitive demand). While IEA, however, returned to a Tyler-Bloom interpretation, NAEP in the 1996-2005 framework kept the objectives interpretation. This, it seems, was influenced largely by US curriculum development project representation in NAEP (Ina V.S. Mullis, 1992). The framework committee "reviewed key blue-ribbon reports, examined exemplary practices, studied local and state-based innovations in science curricula, reviewed science education literature, and noted innovations emerging in other countries" (NAGB, 2004: 9). Among projects reviewed, for example, Mullis (1992) lists *Project 2061* (American Association for the Advancement of Science, 1989), by the American Association for the Advancement of Science, and *Scope and Sequence* (Aldridge, 1989), by the National Science Teachers Association. Both projects demanded widening the science curriculum from traditional teaching of scientific concepts and theories. In other words, there was a great pressure on NAEP to include "the full range of student behaviours" and not, like TIMSS, place scientific inquiry and nature of science (which do not fit into the Tyler-Bloom matrix) in the background. Secondly, NAEP expressed awareness of Millar and Driver (1987) and others who claimed that science behaviour is knowledge-dependent. Hence, statements such as "control of variables", became something students should *understand* rather than a skill they should *do*. The *Knowing and Doing* term was thus used to express that behaviour means *knowing* and *doing* science (as distinct curriculum aims), *and* that the behaviour *includes* knowledge. These changes had the effect of making the two NAEP matrix dimensions more independent, combining them becoming an *ideal* rather than a *psychologically inextricability* as it had been in the Tyler-Bloom rationale.

By abandoning the Tyler-Bloom interpretation of behaviour, NAEP was left with the same problem of describing levels of achievement as IEA had experienced (i.e. Klopfer's second problem). The solution came in terms of the "Angoff principle" and took place against a background of general debate about US academic achievement, which claimed unacceptably low levels, masked by norm-referenced reporting (Koretz, 2008: 182). Angoff (1971) suggested using panels of judges to evaluate item difficulty, coupled with alignment with *cut-scores* for achievement levels on assessment scales. Subsequently, three levels were introduced across all NAEP frameworks (NAGB, 2004). These were: *Basic*, denoting "partial mastery of prerequisite knowledge and skills that are fundamental for proficient work" (p. 36); *Proficient*, representing "competency over challenging subject matter, including subject matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter solid academic performance" (p. 36); and *Advanced*, meaning that students could "integrate, interpolate, and extrapolate information embedded in data to draw well-formulated explanations and conclusions" and "use complex reasoning skills to apply scientific knowledge to make predictions" (p. 36). Placing this principle onto the two-dimensional matrix (content and behaviour) created a "third dimension" for achievement level. Interestingly, this new dimension became similar to the hierarchical ordering of the TIMSS cognitive demand dimension, i.e. still had resemblance with Bloom's taxonomy.  For example, the *basic*, *proficient* and *advanced* levels matching *knowing*, *applying* and *reasoning, and* with many similar cognitive processes at each level. A difference, however, was NAEP including the complexity of knowledge and not *just* reasoning in their definition of cognitive demand.

One observation from the NAEP 1996-2005 framework document is the struggle in explaining the changes being made. Labelling the behaviour dimension *Knowing and Doing*, for example, illustrated a fundamental problem accounting for the knowledge-dependency of the behaviour dimension.  The combination of *content* (Fields of Science) and *behaviour* (Knowing and Doing) into the third achievement level dimension was also explained hesitantly.

The next, current, version (See Figure 2) for 2009 corrected some of this uncertainty, making the principles from the previous version explicit and theoretically coherent. The achievement level dimension, for example, was named *Performance Expectations* and explained:

> … science content statements can be combined (crossed) with science [behaviour] to generate performance expectations (i.e., descriptions of students' expected and observable performances on the NAEP Science Assessment). Based on these

> performance expectations, assessment items can be developed and then inferences can be derived from student responses about what students know and can do in science.
>
> (NAGB, 2008: 63)

This version of the framework emerged from another comprehensive development process involving many experts from different academic areas and an extensive hearing among science educators. NAEP, however, this time has moved away from the “curriculum influence” expressing more interest in including “new advances from science and cognitive research” and “to learn from experiences in TIMSS and PISA” (NAGB, 2008: 2). This has resulted in new principles, but also new ambiguities in the conceptualisation; discussed next.

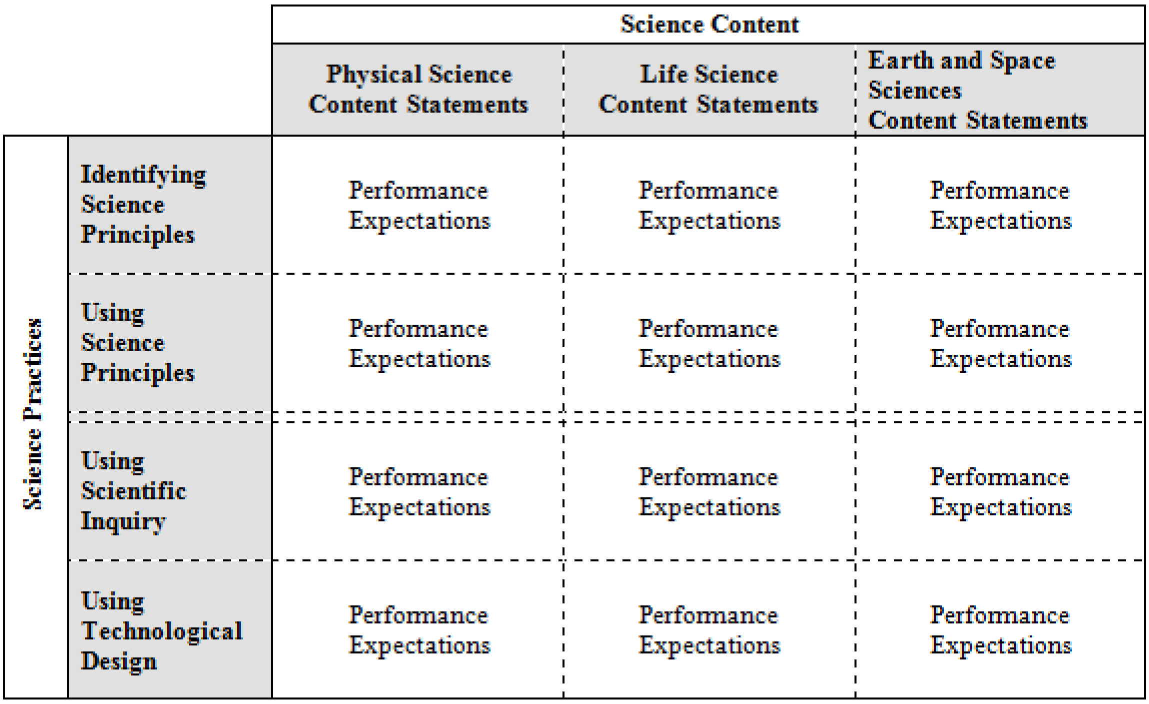
The framework has moved away the overarching dimensions (Nature of Science and Themes), and by using two dimensions only appears more similar to the traditional content-by-behaviour matrix. The behaviour dimension, however, has become *Science Practices*, demonstrating an interest to adapt to the “practice turn” in learning sciences and science studies. One implication arising is that the *nature of science* dimension is embedded in behaviour and linked to students’ cognition. The framework document explains this using Li and Shavelson’s (2001) distinction between declarative-, procedural- , schematic- and strategic knowledge, presented as “knowing that”, “knowing how”, “knowing why”, and “knowing when and where to apply knowledge” (NAGB, 2008: 65). For example, the practice “Using scientific principles” is explained as drawing on *schematic* and *declarative knowledge* to predict observations and evaluate alternative explanations (p. 68). Other *practices* in Figure 2 are explained similarly.

The framework, however, implies uncertainty about what *procedural*, *schematic* and *strategic* knowledge actually are. Firstly, the knowledge is concealed in the framework, and not listed explicitly, and secondly, Li and Shavelson (2001) link these concepts to psychology rather than science philosophy, making it unclear how they can replace *nature of science*.

The lack of rationale for outlining and choosing categories in the science practice dimension is also problematic. These are presented as a “natural outcome” of the fact that “science knowledge is used to reason about the natural world and improve the quality of scientific thought and action” (NAGB, 2008: 66), a statement which gives poor guidelines for knowing what are sufficient or necessary categories. The actual practice categories included have many similarities to the *Knowing and Doing* categories in the previous version of the framework, suggesting these have been influential in what is regarded as “natural”.

The overall impression is, therefore, that NAEP’s attempt to be at the cutting edge of science education has produced a framework which support current perspectives in learning science and science studies, but which fail to operationalise these at a detailed level. The commitment to bring in “hundreds of individuals across the country” (NAGB, 2008: vi) seems further to have forced compromises to the labelling and organising principles of the framework.

In summary, the NAEP science framework offers an alternative to that of the IEA. Both use two-dimensional content-by-behaviour matrixes, but with different dimensions and underlying principles. TIMSS retains a “cognitive matrix”, describing behaviours as what students should “do to the knowledge”. NAEP, in contrast, first established a “curriculum matrix”, treating the behaviour dimension as a fuller list of “objectives” of the science curriculum. This required a third dimension to define achievement levels. The conceptualisation has later been modified in the current version of the framework by redefining the behaviour dimension as *scientific practices*. It is, however, unclear how this actually should be interpreted and the framework document fails somewhat to explain the difference between science practice and science process. NAEP’s framework has been influenced by US curriculum changes and the intention to implement educational research findings, but these act as double-edge swords, causing uncertainties about understanding of concepts and principles. Current challenges include explaining the meaning of *embedding* nature of science into science practices and establishing a rationale for selecting science practices.



| | | Science Content | | |
|---|---|---|---|---|
| | | Physical Science Content Statements | Life Science Content Statements | Earth and Space Sciences Content Statements |
| Science Practices | Identifying Science Principles | Performance Expectations | Performance Expectations | Performance Expectations |
| | Using Science Principles | Performance Expectations | Performance Expectations | Performance Expectations |
| | Using Scientific Inquiry | Performance Expectations | Performance Expectations | Performance Expectations |
| | Using Technological Design | Performance Expectations | Performance Expectations | Performance Expectations |

Figure 2: The NAEP 2009 Science Framework

<u>The PISA Science Framework</u>

The PISA project was established in 1997 and started triennial surveys from 2000. The surveys focus on literacy in reading, mathematics *and* science, alternating between each domain as its main focus. Thus, the first PISA survey with science as main focus took place in 2006, with the next due in 2015. The OECD Secretariat is responsible for PISA survey design and implementation is through an international consortium led by the Australian Council for Educational Research (ACER).

Starting in the late 1990s, PISA was in a different position to NAEP and TIMSS, as no previous version guided the choice of assessment framework. Hence, a new model *could* have been created. However, Wynne Harlen, the first framework committee chair, had experience from two assessment projects, namely the *Techniques for the Assessment of Practical Skills* (TAPS) (Bryce, McCall, MacGregor, Robertson, & Weston, 1988) and the *Assessment for Performance Unit* (APU) (Johnson, 1989). Both proved influential to the PISA development process. The first PISA framework (OECD, 1999) was similar to the APU framework (Murphy & Gott, 1984), adopting a three dimensional framework with s*cientific* *processes*, *scientific concepts* and s*ituations* as dimensions. As in TAPS and APU, PISA announced *scientific process*es as the main target. Harlen (1999) continued a debate about knowledge-dependency of scientific processes using arguments similar to those emerging from TAPS and APU (Gott & Murphy, 1987). The framework document (OECD, 1999), for example, argued "there is no meaning in content-free processes" (p. 60) and scientific knowledge and process "are bound together" (p. 60). Operationalising these arguments proved to be as difficult as NAEP found. Authors resorted to phrases such as "processes, because they are scientific, involve knowledge" (p. 60) and "priority is given to processes *about* science compared to processes *within* science" (p. 61, their emphasis).

As the knowledge-dependency problem remained unresolved, the PISA framework's originality relative to TAPS and APU became the *scientific literacy* focus. This meant emphasising the processes of evaluating scientific evidence and claims in socio-scientific contexts, giving less attention to experiments and data gathering in a laboratory context. Five categories developed on the scientific process dimension were (OECD, 1999: 62):

1. Recognising scientifically investigable questions.
2. Identifying evidence needed in a scientific investigation.
3. Drawing or evaluating conclusions.
4. Communicating valid conclusions.
5. Demonstrating understanding of scientific concepts.

The second PISA framework retained the process-oriented focus, but rearranged the process dimension into three categories (OECD, 2003b: 137):

- Process 1: Describing, explaining and predicting scientific phenomena;
- Process 2: Understanding scientific investigation; and
- Process 3: Interpreting scientific evidence and conclusions

As with many previous framework reviews, limited explanation for this is found, but the new categories became similar to Klahr and Li's (2005) three main "phases of the scientific discovery process" (p. 218). A move is therefore observed away from the stepwise approach to the scientific method seen in TAPS and APU, towards describing more how science works *in principle*. This was a small but important step towards the same "practice turn" as observed in NAEP. The conceptual problem, however, about defining science processes as "knowledge-based" remained unsolved in the 2003 framework.

The *scientific concepts* and *situation* dimensions played inferior roles in both the first two PISA frameworks. The situation dimension, however, added an important difference from TIMSS and NAEP by describing characteristics of the context rather than what students should learn. This will be discussed later as an extension of the conceptualisation of the science domain.

The next development, when scientific literacy became main focus in PISA 2006, offered a new start and a new committee chaired by Rodger Bybee from the US national standards for science education committee. His background together with developments in OECD's DeSeCo project to define *key competencies for the future*'s (OECD, 2003a) made ground for a revised framework with *scientific competency* instead of *science processes* as main focus and using a new organising principle (see Figure 3).

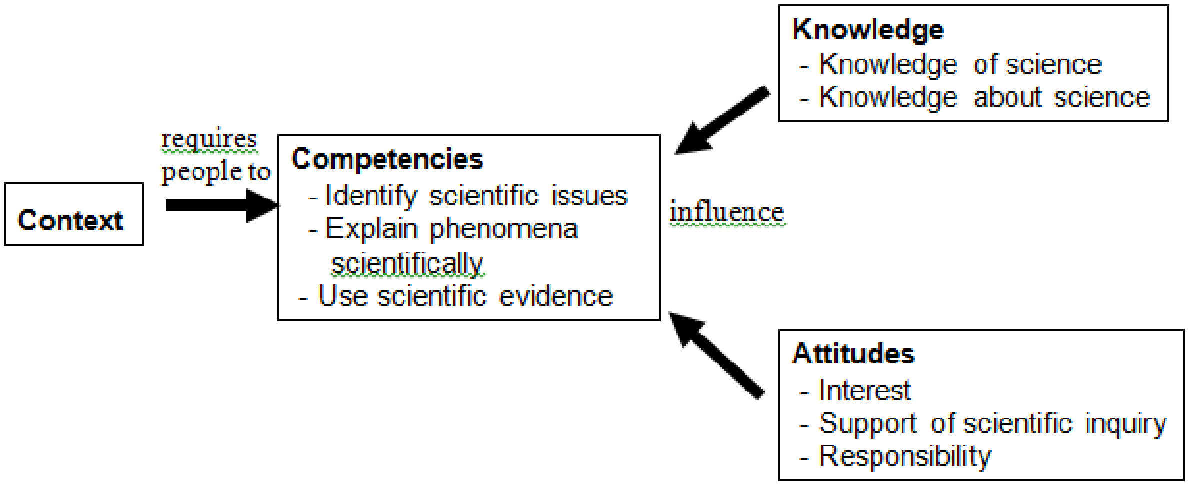

Context
requires people to
Competencies
- Identify scientific issues
- Explain phenomena scientifically
- Use scientific evidence
influence
Knowledge
- Knowledge of science
- Knowledge about science
Attitudes
- Interest
- Support of scientific inquiry
- Responsibility

Figure 3: PISA 2006 Science Framework (OECD, 2006)

The most notable characteristic on the new frameworks is omitting the traditional "matrix model", representing the domain instead as a *concept map*. Compared to the matrix model, this format allows additional dimensions to be involved and it explains more explicitly relationships between the dimensions. Accordingly, the PISA framework in Figure 3 suggests students' competency to "do science" *is influenced by* their knowledge and attitudes. This principle resolved Harlen's earlier conceptual problem of ascribing meaning to a science process being knowledge-based. It also provided an alternative to NAEP's problem discussed earlier about explaining nature of science as "embedded" in science behaviour. In the PISA framework, knowledge *about* science is placed alongside knowledge *of* science, that is, as a similar "cause" for scientific behaviour. The framework has stayed unchanged since 2006, and gradually become familiar to many science educators as PISA has become a more important source for international comparison and for defining scientific literacy (DeBoer, 2011).

Understanding the changes made to the new PISA framework, it is necessary to look towards the tradition of *competency modelling* that developed from the 1990s (Shippman et al., 2000) and was stimulated in particular by Prahalad and Hamel's (1990) demand for core competencies to prepare for increased global competition. In this context, competency is firstly a "managerial" concept, used, for example, by *assessment centres* conducting "job analysis" (Anastasi & Urbina, 1997). The concept merges two aspects; the activity or task someone should be able to master and the knowledge, attitudes, skills and other characteristics (so-called KASOs) that a person needs to learn in order to solve the task successfully. Kane (1992) uses these aspects in a competence definition:

> .. an individual's level of competence in an area of practice can be defined as the degree to which the individual can use the knowledge, skills, and judgment associated with the profession to perform effectively in the domain of possible encounters defining the scope of professional practice (166).

In this perspective, PISA made two major changes to framework development. Firstly, it defined scientific behaviour by turning towards "task analysis" for citizenship. The guiding question put forward was "What is it important for citizens to know, value, and be able to do *in situations involving science and technology*?" (OECD, 2006: 20, emphasis added). This is significant because behaviour is defined through the situations and tasks students should be able to handle rather than scientific principles. The PISA framework mentions briefly that competencies "rest on [their] importance for scientific investigation" (p. 29), but elaborating this is deemed unnecessary as "tasks define behaviour". Secondly, PISA made it more obvious and explicit that developing a framework means *modelling* and not just *categorising* the subject domain. The NAEP 2009 framework also had gone further than previous frameworks in trying to provide a rationale *explaining* the domain. In both frameworks, *organising principle* becomes a key to understand the domain.

In summary, the PISA framework's initial conceptualisation was similar to the UK-based APU and TAPS assessments, focusing on *science processes* and using a matrix as the organising principle. Inspired by competency modelling in the managerial sector, from 2006 this changed to *science competencies*. PISA then replaced the matrix-model with a concept-map, explaining that science behaviour is *influenced by* knowledge and attitudes. Competency modelling made the conceptualisation become "task-oriented", which means PISA has moved away from explaining science principles towards identifying task students should be able to handle in everyday life. PISA has become a recognised conceptualisation among science educators, not at least for its support to scientific literacy, but also because many agree to *competency* as an appropriate concept for scientific behaviour.

**Conceptualisations of the science domain**

The outlines so far have revealed how all three assessments have undergone a series of revisions to arrive at their current conceptualisations of the science domain. Once a framework is established, any later version is a negotiated combination of the original and new ideas introduced by incoming committees. Leaving old ideas, it seems, has proved as difficult as taking on new ones, exacerbated by democratic intentions to include many experts and stakeholders' opinions. All three criteria listed by Bloom et al. (1956) (i.e. support in educational theory, logical consistency and relevance to practice) have caused challenges. The first attempt to summarise the development is a listing of the types of dimensions that have been included:

1. *Conceptual knowledge*. This has been a main dimension in all frameworks and understood similarly as science theories, laws and concepts.
2. *Behaviour.* This is a second main dimension in all frameworks, but showing much greater variation. The labels have been *Behaviour Dimension* (Klopfer in 1971), *Objective Dimension* (SISS 1984), *Performance Expectation* (TIMSS 1995-1999), *Science Cognitive Domain* (TIMSS 2003-), *Process* (NAEP 1978-1982), *Thinking Skills* (NAEP 1990-94), *Knowing and Doing* (NAEP 1996-2005), *Scientific Process* (PISA 2000-2003), *Science Practices* (NAEP 2009-) and *Competencies* (PISA 2006-). The meanings of these will be discussed below.
3. *Knowledge about science*. This has been included in all frameworks but with variation in both organisation and meaning. Klopfer (1971) used *Orientation* as subcategory within the content domain and *General knowledge* (including *Nature of science* and *Nature of scientific inquiry*) in the behaviour domain. SISS 1984 copied this, but TIMSS 1995 (the next framework) made a change and placed *Knowledge about science* in a separate dimension, *Perspectives*, alongside *Safety in science performance* and *Attitudes*. The current TIMSS framework has stopped using knowledge about science as a separate category and makes a short mentioning only within *Scientific Inquiry*. NAEP started by having *Science and society* as a main dimension. *Nature of science* was later made an "overarching" dimension in NAEP 1996-2005, before the current NAEP framework "embedded" *procedural*, *schematic* and *strategic* knowledge in *Science practice*. PISA, from a literacy perspective, started in 2000 by mentioning nature of science as a key area, but struggled to

operationalise this into categories. From 2006, however, *knowledge about science* has been a sub-category alongside *knowledge of science* in the *content* dimension.

4. *Attitudes*. In one form or the other, this has been mentioned at some stage in the frameworks of all the projects. Klopfer 1971 placed it as sub-category in the *behaviour* dimension, TIMSS 1995 as a category in *perspectives* and NAEP 1981/82 as a separate main dimension. Currently, however, PISA is the only framework including attitudes in the science learning domain. It measures both *attitudes towards science* (interest in science) and *scientific attitudes* (support for scientific enquiry). The other projects measures attitudes towards school science as a "background variable" in a student questionnaire.
5. *Context/Situation.* PISA is the only framework to list context as a dimension. The dimension is principally different from any of the above, because it is not something students should *acquire*, and bringing it into the framework extends the meaning of the science domain.

On one hand, these dimensions suggest a shared conceptualisation of the science domain. They are fundamental dimensions science educators use when describing the science curriculum and what students should learn. On the other hand, the variation within them reveals discrepancies. Grasping the differences is not straightforward, but a place to start is the *behaviour dimension*, which has shown most variation and is central to understanding the domain. Table 1 presents the dimension with sub-categories across the frameworks. Similarities now appear between categories, but these are "false friends" if the dimension has different definitions. Similarities and differences among the frameworks, therefore, have to be traced in a combination of organising principles, definition of dimensions and choice of categories. From this perspective, Table 2 suggests five main types of conceptualisations appearing among the frameworks. Frameworks *not* mentioned in the table are found to use a combination of the listed conceptualisations.

The first row presents the simplest conceptualisation, which defines the behaviour dimension as *objectives* (demonstrated by SISS 1984 in Table 1). Although this conceptualisation uses two dimensions (content knowledge and objectives), it is a prolonging of the same *listing* of learning objectives that is commonly found in standards documents. Parallels, for example, can be drawn between the early NAEP frameworks, which listed four independent dimensions, similar to many standards documents, and SISS including all these dimensions in the behaviour dimension. The advantage of the conceptualisation is an unlimited possibility to include any learning objective found important. The disadvantage is limiting the explanation of the domain and inviting fragmentation of the science domain, both caused by a simple organising principle.

A second, and more elaborated, conceptualisation is the two-dimensional content-behaviour rationale underpinning the Tyler-Bloom matrix and used in TIMSS from 2003. This is a common conceptualisation used in assessments across subjects and relating behaviour to *cognitive demand*, defined as the thinking students "do to the material" (Bloom, Hastings and Madaus, 1971: 28). It means organising behaviour hierarchically from lower to higher order thinking, using *domain-general* levels adjusted to subject-related thinking. The conceptualisation is more powerful than the previous because of (a) linking the dimensions (conceptual understanding and cognitive demand are *psychological inextricable*), (b) offering a clear definition of the behaviour dimension and criteria for necessary and sufficient categories (all cognitive levels from recall to higher order reasoning should be included), and (c) identifying progression (as lower to higher order cognition).

One weakness of the conceptualisation, however, is narrowing the science curriculum. There is no obvious place for *attitudes*, *knowledge about science* or even *scientific inquiry* in the content-behaviour grid. It is telling, for example, how Klopfer's attempt to include "the full range of science achievement" failed. The only way to include scientific inquiry into the behaviour dimension is by taking a strict "nothing special view" (Simon, 1966), suggesting scientific method is accounted for by general cognitive behaviour. This, however, reveals a second and more serious problem with this conceptualisation: by linking general and scientific thinking it supports a logical-empiricist view of science (Finley, 1983;

Koslowski, 1996). TIMSS has attempted to compensate for both problems by having a separate classification of scientific inquiry, however, then creating a fragmented framework.

The third conceptualisation relates *knowledge* and *behaviour* to *product* and *process* in science, i.e. to science knowledge and science method. At first, this conceptualisation appeared as a way of making the previous conceptualisation more domain-specific by replacing or merging *cognitive behaviour* with *science process* in the Tyler-Bloom matrix. This modification, however, builds on the "nothing special view" mentioned above. Framework committees in PISA 2000 and 2003 and NAEP 1996 adapting to the critique of the process approach (e.g. Millar and Driver, 1987), therefore, tried explaining how *science process* is different from *general cognition*. The outcome was interpreting *science process* as a more independent concept, that is, as something to "know and do" rather than "thinking students do to knowledge". This turned *behaviour* into a "commodity" that students can learn parallel to science conceptual knowledge. Table 1 demonstrated this difference when, for example, comparing PISA 2000 with TIMSS 2007. The latter lists levels of cognition students apply when expressing their understanding of science knowledge, while the former lists processes as something students may acquire and that can be measured with a separate scale (Adams & Wu, 2002).

Later, however, *the product-process* conceptualisation has met more resistance than just the false alliance between science method and general cognition, and attention has been drawn towards denying the existence of a general science method (Duschl & Grandy, 2008). Jenkins (2007)describes how *science method* originated as a means for scientists and science educators to promote science in the 19th century, and has dominated much of the 20th century thinking about science, *despite* philosophical, conceptual, and methodological differences between scientific disciplines and *against* developments in science philosophy. Denying then both *science process* and *science method*, has meant framework committees over the last years have had to come up with an alternative approach to account for *scientific behaviour*.

Interestingly, PISA 2006 and NAEP 2009, in the last two rows in Table 2 and in Table 1, have provided very different answers to this challenge. PISA has replaced *science process* with *scientific competencies*, which has background in “managerial job-analysis” and suggests students, like scientists, should develop knowledge and attitudes to solve particular tasks or activities. The concept has a pragmatic use and is defined by listing “professional encounters” (Kane, 1992), which PISA has exchanged with “encounters citizens meet in everyday life”. NAEP uses *scientific practice*, which emerges more directly from the “practice turn” in science studies and learning sciences. Pickering (1992), for example, defines science practice as an activity driven by the interest of the individual and the community, and with resulting products evaluated against community standards. Lave and Wenger’s (1991) describe practice is a “joint enterprise” based on “mutual engagement” and using a “shared repertoire of communal resources”. Both this and Pickering’s account make clear that science behaviour can be understood only through *social and epistemic* aspects of the science community. NAEP reflects this view, but by holding on to a two-dimensional matrix struggles to establish a sufficient organising principle. PISA, on the other hand, offers a solution to the organising problem through the concept-map. The science domain is then not forced into a two or three-dimensional matrix but presented in a model explaining relationships between dimensions. The model offers a potential also to bring in more dimensions and scientific behaviour is explained as influenced by knowledge of science, knowledge about science and conceptual understanding. A further development of the framework conceptualisation, as will be suggested in the next section, therefore invites combining ideas from these two assessments.

As a conclusion, the conceptualisation of the science domain in the large-scale assessments has been dominated very much by the struggle to define science behaviour. From other developments in science education over the last forty years, this is maybe no big surprise. The puzzling outcome, however, is how much the struggle has impacted on inclusions or exclusion of other dimensions in the domain.

| **SISS 1984**<br>**Objectives Dimension** | **TIMSS 1995**<br>**Performance Expectation** | **TIMSS 2003**<br>**Science Cognitive Domain** | **TIMSS 2007**<br>**Science Cognitive Domain** |
|---|---|---|---|
| Knowledge and comprehension<br>Process<br>Applications<br>Manual skills<br>Attitudes<br>Orientations | Understanding<br>Theorizing, analyzing, solving problems<br>Using tools, routine procedures and science processes<br>Investigation the natural world<br>Communicating | Factual knowledge<br>Conceptual understanding<br>Reasoning and analysis | Knowing<br>Applying<br>Reasoning |
| **NAEP 1990**<br>**Thinking Skills** | **NAEP 1996-2005**<br>**Knowing and Doing** | | **NAEP 2009**<br>**Science Practices** |
| Knowing Science<br>Solving Problems<br>Conducting Inquiries. | Conceptual Understanding<br>Scientific Investigation<br>Practical Reasoning | | Identifying Science Principles<br>Using Scientific Principles<br>Using Scientific Inquiry<br>Using Technological Design |
| | **PISA 2000**<br>**Processes** | **PISA 2003**<br>**Processes** | **PISA 2006**<br>**Competencies** |
| | Recognising scientifically investigable questions<br>Identifying evidence needed in a scientific investigation<br>Drawing or evaluating conclusions<br>Communicating valid conclusions<br>Demonstrating understanding of scientific concepts | Describing, explaining and predicting scientific phenomena<br>Understanding scientific investigation<br>Interpreting scientific evidence and conclusions | Identifying scientific issues<br>Explaining phenomena scientifically<br>Using scientific evidence |

Table 1: Labelling of and main categories in the ‘behaviour dimension’ of the frameworks

## The way ahead?

Table 2 suggests a development towards more advanced *explanations* of the science domain. The fact that all assessments have put much effort and resources into this development suggests they find “explanatory” conceptualisations important and useful. The reason, we may think, is supporting the *interpretive argument* when comparing the *intended* and *actual* constructs (Kane, 2006; Wiley, 2001). In other words, that a framework offering a *rationale* for the science domain helps assessors maintain construct validity when developing items and interpreting results in a better way than one just *listing dimensions*. The same argument would apply to teachers operationalising the intended curriculum into teaching. PISA and NAEP, currently, have frameworks with the most elaborated explanations, but neither of them accounts sufficiently for the “practice turn” in the science studies and the learning sciences. A way forward, as suggested above, is developing elements from these two frameworks to provide better explanations of the science domain.

What particularly needs explaining is the meaning of scientific behaviour being a “community practice”. This influences the definition of dimensions involved and the relationship between them, but also the choice of sub-categories. PISA relates scientific behaviour to “science competency” and explains this as *influenced by* knowledge of science, knowledge about science and attitudes. NAEP uses “scientific practice” and explains that conceptual, procedural, schematic and strategic knowledge is *embedded in* scientific behaviour. The alternative that will be suggested is explaining scientific behaviour as *defined by* the knowledge and attitudes. As a “community practice”, behaving scientifically means using shared understanding and following norms and standards of the science community (Ford and Foreman, 2006), which are defined through conceptual, procedural and epistemic knowledge. This makes scientific knowledge and behaviour sociologically and psychologically inextricable.

| **Conceptualisation** | **Behaviour Dimension** | **Organising Principle** | **Background** | **Comments** | **Framework** |
|---|---|---|---|---|---|
| Objectives listing | 'Objectives'<br>Anything students do *in* science | Listing content and learning objectives | Curriculum analysis to identify aims for Science | Inclusive of many objectives, but limited explanatory power | Klopfer 1971<br>SISS 1984 |
| Content-Behaviour | 'Cognitive demand'<br>Cognition students do *to* science knowledge | Combining content and cognition to form hierarchically categories from lower to higher-order thinking | Psychological conceptualisation of human performance | Explains *cognitive demand* in a way that is incongruent with domain-specific views of science | TIMSS 2003 and current |
| Product-Process | 'Science Process'<br>Thinking and/or methods used by scientists | Combining science knowledge and science processes | Promotion of science methods as a useful tool for problem solving | Explains *science*, but rests on a false premise that scientists use a specific method | APU<br>PISA 2000/2003 |
| Science competency | 'Science Competency'<br>Successful performance of scientific task or activity | Modelling science competency in concept map | Managerial 'job analysis' to identify core competencies in scientific literacy | Takes a 'task-focus' that reduces the emphasis on explaining science principles | PISA 2006 and current |
| Science practice | 'Science Practice'<br>Behaviour within the science community, regulated by social and epistemic criteria | Combining content and practice to explain performance expectations | 'Practice turn', redefining science as social and naturalistic (rather than logical-normative) phenomenon | Adapts conceptualisation to the practice turn, but in need of better organising principle | NAEP 2009 and current |

Table 2: Key types of conceptualisations identified in the assessment frameworks

Kind, Osborne and Szu (in review) has proposed a rationale elaborating this alternative. It relates *science practices* to *purposes* of doing science. Philosophical (Giere, Bickle, &Mauldin, 2006) and psychological (Klahr, Fay, & Dunbar, 1993) accounts of science suggest there are three main purposes: developing scientific ideas and explaining phenomena (theorising), gathering data and testing theories (experimentation) and evaluating scientific claims and evidence and coordinating these in scientific argumentation (evidence evaluation). These purposes and practices are necessary to make a complete scientific argument. Two more purposes are technological problem solving, which leads to *engineering* as a practice, and *communication*. These practices, however, have a different status, because they exceed the science principles of explaining nature. The rationale further suggests parallels exist between how the *purposes* define *practices* in science research, society and science classroom. Scientists, for example, do "theorising" to develop new understanding of scientific phenomena, but students engage in a similar practice when learning science knowledge at school. This is important, because it makes the five science practices meaningful categories in a conceptualisation of the behaviour dimension in a school curriculum, not just in "real science". It also makes a common ground for defining categories in the knowledge dimension: students should understand *science conceptual knowledge*, which describes the shared understanding of science phenomena; *procedural knowledge*, which describes norms and criteria for gathering data and testing theories; and *epistemic knowledge*, which describes norms and criteria for evaluating and coordinating scientific claims and evidence. Behaving "scientifically" at school, when engaging in a socio-scientific debate or when doing scientific research, all mean applying and committing to this knowledge. Figure 4 expresses this conceptualisation using a similar concept map as in PISA 2006, but redefining dimensions and their relationship.

Two important elements are missing from the conceptualisation in Figure 4. The first is *cognitive demand*, which TIMSS has related to categories in the revised version of Bloom's taxonomy (Anderson et al., 2001). PISA and NAEP, however, have both made the point that demand relates also to the *knowledge* involved, and further, not just to science conceptual knowledge but all three types of knowledge in Figure 4. The current conceptualisation takes a similar view when suggesting that science practice is *defined* by knowledge. Higher cognitive demand then means using conceptual, procedural and/or epistemic ideas that are more complex and elaborated. Importantly, this complexity is seen from a students' point of view and aligns therefore with "developmental progression" (Berland & McNeill, 2010). When identifying cognitive demand, it is therefore necessary to look, first, to the science community to identify what ideas students should learn (i.e. what are the accepted knowledge and norms in the science community), and thereafter, to how the understanding of these ideas typically develops in students' learning. Kind (2013) has demonstrated that a rich source of research literature is available to identify progression in students' understanding of epistemic and procedural knowledge, in the same way as learning progression research already exist in understanding many areas of conceptual knowledge, and that this knowledge is useful to develop assessment scales. Progression in complexity of knowledge, of course, combines with, rather than replace, the cognitive demand identified in Bloom's taxonomy.

The second element missing in Figure 4 is *attitudes*. On one hand, this is as a natural dimension to include. Siegel (1989), for example, argues convincingly that *evidence evaluation* is based upon a willingness to conform to criteria and not simply understanding them. From his view, *attitude* could be place in Figure 4 parallel to knowledge as a dimension that defines science practice, which would mean following a similar principle as the current PISA framework (Figure 3). This attitude dimension would link also to knowledge, because knowledge makes the *attitude object* (Ajzen, 2001). This means, for example, that commitment to scientific behaviour requires having a positive attitude towards scientific explanations, scientific data and scientific criteria. Hodson (1998), however, in contrast to Siegel, denies this and places scientific attitudes as a "myth of science education". His argument is that good scientific practice exists independently of the persons' attitude. Besides, it is impossible to tell if a person behave scientifically due to scientific attitudes or not. Clarifying the meaning and role of scientific attitudes is therefore necessary *before* this variable can be included meaningfully in framework conceptualisation.

Even without being complete, the conceptualisation in Figure 4 suggests important implications for assessment. Test items, for example, have to include both dimensions. Making an item testing a scientific practice but not knowledge is impossible, because engaging in a scientific practice *means* applying science conceptual, procedural and/or epistemic knowledge. Testing knowledge without engagement in a scientific practice may be possible in principle, but would mean having an item not including a scientific problem and therefore has little value. For these reasons, the conceptualisation supports Allchin's (2011) claim that knowledge about science, that is procedural and epistemic knowledge, should be tested implicitly through science practices, but suggests the same applies also to science conceptual knowledge. Besides, the conceptualisation adds another perspective that even if knowledge is tested implicitly through engagement in practices, explicit definition of the knowledge categories is necessary for meaningful operationalisation into assessment items. A more open question is how the three types of knowledge combined with the five types of practices allow development of subscales. This is a matter about what type of knowledge category is included in the different practices and the degree to which each category can be separated out. Although not answering this, the conceptualisation makes a starting point for closer examination of practices and their related need for use of knowledge.

## Concluding remarks

Au (2007) warns that high-stake testing undermines education by narrowing and fragmenting the curriculum. From reviewing TIMSS, NAEP and PISA, this study suggests these problems are caused in part by assessment frameworks' conceptualisation of the domain. A framework identifying dimensions without sufficiently explaining their inter-relationships may support a wide curriculum, but invites fragmentation because assessors make items and scales measuring dimensions separately. Establishing a rationale and organising principle explaining the relationship between domains may prevent this fragmentation, but also contribute to narrowing the curriculum if the principle is too simple. The requirement for a functional conceptual framework is therefore to establish a rationale that accounts for the whole science curriculum. The study has demonstrated that this is not a straightforward task. The large-scale assessments have been working on their frameworks for more than forty years and have yet to reach accomplished solutions. The study, however, suggests improvement is being made and that frameworks have become better rationales for teaching and assessment in science education.

## References

Adams, R., & Wu, M. (2002). *PISA 2000 Technical Report*. Paris: OECD.

Ajzen, I. (2001). Nature and operation of attitudes. *Annual review of psychology, 52*, 27-58.

Aldridge, B. G. (1989). *Essential changes in secondary school science: Scope, sequence and coordination:* Washington DC: National Science Teachers Association.

Allchin, D. (2011). Evaluating Knowledge of the Nature of (Whole) Science. *Science Education, 95*, 518-542.

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology*. Washington DC: Author.

Anastasi, A., & Urbina, A. (1997). *Psychological testing (7th ed.)*. Upper Saddle River, NJ: Prentice Hall.

Anderson, L. W., Krathwohle, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revison of Bloom's Taxonomy of Educational Objectives*. New York: Longman.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.* (pp. 508-600). Washington, DC: American Council on Education.

Au, W. (2007). High Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher, 36*(5).

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays (C. Emerson & M. Holquist, Trans.).* Austin, TX: University of Texas Press.

Berland, L., & McNeill, K. (2010). A Learning Progression for Scientific Argumentation: Understanding Student Work and Designing Supportive Instructional Contexts. *Science Education, 94*, 765-793.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: the classification of educational goals; Handbook I: Cognitive Domain*. New York, : David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G., E. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York: McGraw-Hill Book Company.

Brown, G., & Desforges, C. (1977). Piagetian Psychology and Education: Time for revision. *British Journal of Educational Psychology, 47*, 7-17.

Bryce, T. G., McCall, J., MacGregor, J., Robertson, I. J., & Weston, R. J. (1988). *Techniques for Assessing Process Skills in Practical Science (TAPS 2)*. London: Heinman.

Champagne, A., Bergin, K., Bybee, R., Duschl, R. A., & Gallagher, J. (2004). *NAEP 2009 Science Framework Development: Issues and Recommendations. Paper prepared for the National Assessment Governing Board Washington*. Washington, DC.

Comber, L. C., & Keeves, J. P. (1973). Science Education in Nineteen Countries.

DeBoer, G. E. (2011). The Globalization of Science Education. *Journal of Research in Science Teaching, 48*(6), 567-591.

Donaldson, M. (1984). *Children's Minds* London Fontana

Driver, R., & Easley, J. (1978). Pupils and paradigms: a review of literature related to concept development in adolescent science students. *Studies in Science Education, 5*, 61-84.

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312.

Duschl, R. A., & Grandy, R. E. (2008). Reconsidering the character and role of inquiry in school science: framing the debates. In R. A. Duschl & R. E. Grandy (Eds.), *Teaching Scientific Inquiry: Recommendations for Research and Implementation*. Rotterdam, NL: Sense Publisher.

Finley, F. (1983). Science Processes. *Journal of Research in Science Teaching, 20*, 47-54.

Ford, M., & Forman, E. A. (2006). Redefining Disciplinary Learning in Classroom Contexts. *Review of Research in Education, 30*, 1-32.

Gagne, R. M. (1965). The psychological basis of science - a process approach. *AAAS miscellaneous publication*, 65-68.

Giere, R. N., Bickle, J., & Mauldin, R. F. (2006). *Understanding Scientific Reasoning*. CA: Belmont: Thomson Wadsworth.

Gott, R., & Murphy, P. (1987). *Assessing Investigation at Ages 13 and 15. Assessment of Performance Unit Science Report for Teachers: 9*. London: Department of Education and Science, Welsh Office, Department of Education for Northern Ireland.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*, 199-220.

Haertel, E., & Calfee, R. (1983). School Achievement: Thinking about What to Test. *Journal of Educational Measurement, 20*(2), 119-132.

Harlen, W. (1999). Purposes and Procedures for Assessing Science Process Skills. *Assessment in Education: Principles, Policy & Practice, 6*(1), 129-144.

Hewson, P. W. (1981). A conceptual change approach to learning science. *European Journal of Science Education 3*(4), 383-396.

Hodson, D. (1998). It this really what scientists do? Seeking a more authentic science in and beyond the school laboratory. In J. Wellington (Ed.), *Practical work in school science. Which way now?* (pp. 93-108). London: Routledge.

Hueftle, S. J., Rakow, S. J., & Welch, W. W. (1983). *Images of science: A summary of results from the 1981-1982 National Association in Science*. Minneapolis, M.M: Minnesota Research and Evaluation Centre.

Hughes, J., Jewson, N., & Unwin, L. (2007). Introduction: Communities of practice: a contested concept in flux. In J. Hughes, N. Jewson & L. Unwin (Eds.), *Communities of practice: Critical perspectives* (pp. 1-16). Abingdon . Routledge.

Husen, T., & Postlewaite, T. N. (1996). A Brief History of the International Association for the Evaluation of Educational Achievement (IEA). *Assessment in Education: Principles, Policy & Practice, 3*(2), 129-141.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking*. London: Routledge Kegan Paul.

Jenkins, E. (2007). School science: a questionable construct? *International Journal of Science Education, 39*(3), 265-282.

Johnson, S. (1989). *National Assessment: the APU science approach*. London: HMSO.

Kane, M. T. (1992). The Assessment of Professional Competence. *Evaluation & the Health Profession, 15*(2), 163-182.

Kane, M. T. (2006). Validation. In L. Brennan (Ed.), *Educational measurement, 4th ed.* Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kind, P.M. (2013). Establishing assessment scales using a novel knowledge-based rationale for scientific reasoning *Journal of Research in Science Teaching*, 50 (5), 530-560.

Kind, P.M., Osborne, J. & Szu, E. (In review). Towards a Model of Scientific Reasoning for Science Education. *Science Education*.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. Cognitive Psychology, 24(1),111-146.

Klahr, D., & Li, J. (2005). Cognitive Research and Elementary Science Instruction: From the Laboratory, to the Classroom, and Back. *Journal of Science Education and Technology, 14*(2), 217-238.

Klopfer, L. E. (1971). Evaluation of learning in science. In B.S. Bloom, J. T. Hastings & G. F. Madeus (Eds.), *Handbook on formative and summative evaluation of student learning*. New York: McGraw Hill.

Koretz, D. (2008). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, MA: Harvard University Press.

Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning*. Cambridge, MA: MIT Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: Univ. of Chicago Press.

Latour, B. (1987). *Science in action*. Milton Keynes, England: Open University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lehrer, R., & Schauble, L. (2006). Scientific thinking and science literacy: Supporting development in learning in contexts. In W. Damon, R. M. Lerner, K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology, 6th ed. (Vol. 4)*. Hoboken, NJ: John Wiley and Sons.

Li, M., & Shavelson, R. J. (2001). *Examining the links between science achievement and assessment*. Paper presented at the American Educational Research Association.

Linn, R. L., & Baker, E. L. (1995). What do International Assessments imply for world-class standards? *Educational Evaluation and Policy Analysis, 17*(4), 405-418.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. NJ: Hillsdale: Lawrence Erlbaum.

Millar, R., & Driver, R. (1987). Beyond Process. *Studies in Science Education, 14*, 33-62.

Miller, G., A. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences, 7*(3), 141-144.

Moseley, D., Baumfield, V., Elliot, J., Gregson, M., Higgins, S., Miller, J., et al. (2005). *Frameworks for Thinking. A handbook for Teaching and Learning*. Cambridge: Cambridge University Press.

Mullis, I. V. S. (1992). Developing the NAEP Content-Area Frameworks and Innovative Assessment Methods in the 1992 Assessments of Mathematics, Reading, and Writing. *Journal of Educational Measurement, 29*(2), 111-131.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment Frameworks*: International Association for the Evaluation of Educational Achievement.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. MA: Boston: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., et al. (2003). *TIMSS Assessment Frameworks and Specifications 2003, 2nd Edition*. Boston: International Study Center, Boston College.

Murphy, P., & Gott, R. (1984). *The Assessment Framework for Science at Age 13 and 15. APU Science report for teachers: 2*. UK: DES.

NAEP. (1979). *Three assessments of science, 1969-77: technical summary. Report No. 08-S-21.* Washington, DC: Education Commission of the States.

NAGB. (2004). *Science Framework for the 2005 National Assessment of Educational Progress*. Washington, DC: U.S. Department of Education.

NAGB. (2008). *Science Framework for the 2009 National Assessment of Educational Progress*. DC: Washington: Author.

Neidorf, T. S., Binkley, M., & Stephens, M. (2006). *Comparing Science Content in the National Assessment of Educational Progress (NAEP) 2000 and Trends in International Mathematics and Science Study (TIMSS) 2003 Assessments. Technical Report*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Nohara, D., & Goldstein, A. A. (2001). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA). Working Paper No. 2001-07* U.S. Department of Education, National Center for Education Statistics.

OECD (1999). *Measuring student knowledge and skills. A New Framework for Assessment*. Paris: OECD Publisher.

OECD (2003a). *Definition and Selection of Competencies: Theoretical and Conceptual Foundations (DeSeCo), Summary of the final report "Key competencies for a Successful life and a Well-Functioning Society"*. Paris: OECD.

OECD (2003b). The PISA 2003 Assessment Framework- Mathematics, Reading, Science and Problem Solving Knowledge and Skills. Paris: OECD Publisher.

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy A Framework for PISA 2006*. Paris: OECD Publisher.

Osborne, R. J. (1982). Conceptual change - for pupils and teachers. *Research in Science Education, 12*, 25-31.

Osborne, R. J., & Wittrock, M. C. (1985). The generative learning model and its implications for science education. *Studies in Science Education, 12*, 59-87.

Pickering, A. (Ed.). (1992). *Science as Practice and Culture*. Chicago, IL: The University of Chicago Press.

Prahalad, C. K., & Hamel, G. (1990). The Core  Competence of the Corporation. *Harvard Business Review, 68*(3), 79-91.

Quine, W. V. (1969). Natural kinds. In W.V. Quine (Ed.), *Ontological relativity and other essays*. New York: Columbia University Press.

Robitaille, D. F., Schmidt, W.H., Raizen, S., Mc Knight, C., Britton, E., & Nicol, C. (1993). *Curriculum Frameworks for Mathematics and Science. TIMSS Monograph No.1*. Vancouver: Pacific Educational Press.

Rosier, M. J. (1987). The Secod International Science Study. *Comparative Education Review, 31*(1), 106-128.

Shippman, J. S., Ash, R. A., Battista, M., Carr, L., Eyde, L. D., Hesketh, B., et al. (2000). The practice of competency modeling. *Personnel Psychology, 53*, 703-740.

Siegel, H. (1989). The Rationality of Science, Critical Thinking, and Science Education. *Synthese, 80*, 9-41.

Simon, H. A. (1966). Scientific discovery and the psychology of problem solving. In R. Colodny (Ed.), *Mind and cosmos* (pp. 22-40). Pittsburgh, PA: University of Pittsburgh Press.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction.* IL, Chicago: Chicago University Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Wiley, D. E. (2001). Validity of constructs versus construct validity. In H. Braun, D. N. Jackson & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 207-227). Mahwah, NJ: Lawrence Erlbaum.

Wu, M. (2010). *Comparing the Similarities and Differences of PISA 2003 and TIMSS, OECD Education Working Papers, No. 32*. Paris: OECD Publishing.