Towards a Framework for Assessing 21st Century Science Skills

Maria Araceli Ruiz-Primo

University of Colorado Denver

Commissioned paper for The National Academies

February, 2009

Towards a Framework for Assessing 21st Century Science Skills

Maria Araceli Ruiz-Primo

University of Colorado Denver

This paper offers a framework for designing technically-sound assessments for measuring 21st century skills in the context of science education. It is organized in five sections. In the first section, I examine the challenges of defining and understanding 21st century skills. In the second section, I discuss 21st century skills in the context of science education and make an attempt to define the constructs that are relevant to their assessment. In the third section, I map out these constructs according to four well-known large-scale science assessments frameworks: the National Assessment for Educational Progress (NAEP), Trends in Mathematics and Science Study (TIMSS), Programme for International Students Assessment (PISA), and Collegiate Learning Assessment (CLA). In the fourth section, I propose a framework for developing and evaluating assessment tasks for measuring 21st century skills. Finally, in the fifth section I examine issues relevant to evaluating the technical quality of these assessments.

Understanding the 21st Century Skills

In this section, I provide some background necessary for understanding current views of 21st century skills. First, I discuss the framework that resulted from a workshop organized by The National Academies with the purpose of identifying basic 21st century skills. Then, I compare these skills with other frameworks and conceptions of 21st century skills.

In 2007, The National Academies held the *Workshop on Research Evidence Related to Future Skill Demands*. The research discussed at that workshop highlighted five broad skills that appear valuable across a range of jobs, from low-wage service work to professional work: *adaptability*, *complex communications/social skills*, *non-routine problem solving*, *self-management/self-development*, and *systems thinking*. These five 21st century skills were adapted from a set of six broad competencies initially proposed by Janis Houston (2007), of which two (self-management and self-development) were collapsed into one. Hilton (2008) defined the five competencies from the workshop as follows:

1. **Adaptability:** The ability and willingness to cope with uncertain, new, and rapidly-changing conditions on the job, including responding effectively to emergencies or situations of crisis and learning new tasks, technologies, and procedures. Adaptability also includes handling work stress; adapting to different personalities, communication styles, and cultures; and physical adaptability to various indoor or outdoor work environments.

2. **Complex communications/social skills:** Skills in processing and interpreting both verbal and non-verbal information from others in order to respond appropriately. A skilled communicator is able to select key pieces of a complex idea to express in words, sounds, and images, in order to build shared understanding. Skilled communicators negotiate positive outcomes with customers, subordinates, and superiors through social perceptiveness, persuasion, negotiation, instructing, and service orientation.

3. **Non-routine problem solving:** A skilled problem-solver uses expert thinking to examine a broad span of information, recognize patterns, and narrow the information to reach a diagnosis of the problem. Moving beyond diagnosis to a solution requires knowledge of how the information is linked conceptually and involves metacognition—the ability to reflect on whether a problem-solving strategy is working and to switch to another strategy if the current strategy isn't working. It includes creativity to generate new and innovative solutions, integrating seemingly unrelated information, and entertaining possibilities others may miss.

4. **Self-management/Self-development:** Self-management skills include the ability to work remotely, in virtual teams; to work autonomously; and to be self motivating and self monitoring. One aspect of self-management is the willingness and ability to acquire new information and skills related to work.

5. **Systems thinking:** The ability to understand how an entire system works, how an action, change, or malfunction in one part of the system affects the rest of the system; adopting a "big picture" perspective on work. It includes judgment and decision-making; systems analysis; and systems evaluation as well as abstract reasoning about how the different elements of a work process interact.

My comparison of frameworks on 21st century skills is supported by a literature review which started with the critical citations in the *Committee of Research Evidence Related to Future Skills Demands'* report (Hilton, 2008). Although there are many ways in which the skills could be defined, the information provided by Hilton (2008) guided the selection of the dimensions defined. Following the strategy taken by researchers who study work and the job place, I focused on identifying dimensions within the skills. Table 1 provides the dimensions identified for each of the skills.

Table 1. *Dimensions by Skill*

Skill	Dimension	General Definition	Source
Adaptability	Dealing with uncertain and unpredictable work situations	Taking effective actions when necessary without having to know the whole picture; effectively adjusting plans, goals, actions, or priorities to deal with changing situations; changing gears to unpredictable or unexpected events; not being paralyzed by uncertainty or ambiguity; imposing a structure when necessary; not needing things to be black and white.	Hilton, 2008; Houston, 2007; Pulakos et al., 2000, 2002
	Learning work tasks, technologies, and procedures	Demonstrating enthusiasm for learning new approaches and technologies for conducting work; doing what is necessary to keep knowledge and skills current; quickly and proficiently adjusting to new processes and procedures; anticipating changes to work demands; participating in training for changes; taking action to improve deficiencies.	Hilton, 2008; Houston, 2007; Pulakos et al., 2000, 2002
	Demonstrating interpersonal adaptability	Being flexible and open-minded when dealing with others' viewpoints, listening others' viewpoints; accepting negative feedback regarding work, developing effective relationships with diverse personalities; tailoring behavior to persuade, influence, or work more efficiently with others.	Hilton, 2008; Houston, 2007; Mumford, Peterson, & Childs, 1999; Pulakos et al., 2000, 2002
	Demonstrating cultural adaptability	Taking action to learn about and understand climate orientation, needs, and values of other groups, organizations, or cultures; adjusting behavior as needed, to show respect for other values and customs; understanding the implications of own actions; adjusting to maintain positive relations.	Houston, 2007; Hilton, 2008; Pulakos et al., 2000, 2002
	Demonstrating physically oriented adaptability	Adjusting to challenging environmental states such as extreme heat, humidity, cold, or dirtiness; adjusting weight and muscular strength or becoming proficient in performing physical tasks as necessary for the job.	Houston, 2007; Hilton, 2008; Pulakos et al., 2002, 2002
	Handling emergencies or crisis situations	Reacting with appropriate and proper urgency to emergency situations (e.g., life threatening); quickly analyzing options; making split-second decisions based on clear and focused thinking; maintaining emotional control and objectivity while keeping focused.	Houston, 2007; Hilton, 2008; Pulakos et al., 2000, 2002
	Handling work stress	Remaining composed and cool when faced with difficult circumstances or a highly demanding workload or schedule; not overreacting to unexpected news or situations; managing frustration by directing efforts to constructive solutions rather than blaming others.	Hilton, 2008; Houston, 2007; Pulakos et al., 2000, 2002
Complex Communication/ Social Skills [b]	Demonstrating custom fitting communication	Explaining (to provide an understandable or intelligible communication that conveys what information means and how it can be understood or used), listening (to pay attention to what others are saying or asking as appropriate), persuading (to induce people to approach things differently), and negotiating (to reconcile differences) when necessary. Prompting proper questions to clarify information.	Hilton, 2008; Houston, 2007; Jeanneret, Borman, Kubisiak, & Hanson, 1999; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
	Processing verbal and non-verbal information	Interpreting verbal language, corporal language, signs and responding properly; being aware of other's reactions and understanding why they react the way they do; negotiating positive outcomes through social perceptiveness, persuasion, negotiation, instruction, and service orientation.	Hilton, 2008; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
	Distilling information	Selecting key pieces of a complex idea to express in words, sounds, and images in order to build shared understanding.	Hilton, 2008; Levy & Murnane, 2004
	Filling-in information	Filling in information not contained in the information received.	Levy & Murnane, 2004

[a] For Borman, Kubisiak, & Schneider (1999), this dimension is a low-order work style descriptor.

[b] Jeanneret, Borman, Kubisiak, & Hanson (1999) considered Communicating as a second-order *generalized work activity* that is part of the work requirements.

Table 1. (*Continued)*

Skill	Dimension	General Definition	Source
Non-routine Problem Solving	Recognizing or perceiving meaningful patterns	Recalling and perceiving clusters of information has been replicated in different domains (e.g., reading circuit diagrams, architecture, and medicine).	Hilton, 2008; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
	Diagnosing the problem	Identifying the nature of the problem; examining a broad span of information and narrowing it to reach a diagnosis of the problem.	Hilton, 2008; Levy & Murnane, 2004
	Organizing information in schemas	Reflecting deep understanding by linking facts, concepts, and small pieces of information in cause-and-effect sequences that connect the goals and sub-goals needed to solve a problem.	Chi, Glaser & Farr, 1988; Levy & Murnane, 2004
	Monitoring problem-solving activities	Reflecting on whether a problem solution strategy is working, and if it is necessary change to another strategy.	Hilton, 2008; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
	Gathering relevant information	Knowing how to find information and identifying essential.	Chi, Glaser & Farr, 1988; Hilton, 2008; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
	Creating new and innovative solutions	Generating a number of different approaches.	Hilton, 2008; Mumford, Peterson, & Childs, 1999; Levy & Murnane, 2004
Self Management/ Self Development	Working autonomously	Developing own ways of doing things, guiding oneself with little or no supervision, and depending mainly on one-self to get things done. [a]	Borman, Kubisiak, & Schneider, 1999; Hilton, 2008
	Being self-motivated	Setting personal goals, trying to succeed at those goals, and striving to be competent in own work. [a]	Borman, Kubisiak, & Schneider, 1999; Hilton, 2008; Houston, 2009
	Being self-monitoring	Managing one's time and the time of others; assessing how well one is doing when doing or learning something; [d] persistently seeking feedback from others as well as opportunities to update and master job-relevant knowledge and skills.	Hilton, 2008; Mumford, Peterson, & Childs, 1999
	Willing to acquire new information and skills related to work	Volunteering for training and development opportunities; changing behaviors in response to feedback and experience. [d]	Hilton, 2008; Houston, 2009
Systems Thinking	Understating how a system works	Conceptualizing how individual parts fit together into a whole system; seeking and adopting a "big picture" perspective on work and organizational issues; reasoning about how the different elements of a system interact.	Houston, 2007; Hilton, 2008
	Judging and decision making	Appreciating the consequences of individual actions for all parts of the system; weighing the relative cost and benefits of potential actions for the system.	Houston, 2007; Hilton, 2008; Mumford, Peterson, & Childs, 1999
	System analysis and evaluation	Looking at different indicators of system performance; taking into account their accuracy.	Hilton, 2008; Mumford, Peterson, & Childs, 1999

[a] For Borman, Kubisiak, & Schneider (1999), this dimension is a low-order work style descriptor.
[d] For Mumford, Peterson, & Childs (1999), self-monitoring is considered a process skill rather.

Several facts need to be noted regarding these definitions of 21st century skills. First, they focus more on general reasoning and problem-solving abilities, rather than domain-specific knowledge (see Shavelson & Huang, 2003). However, since learning is highly contextualized and situated, it is important to recognize that knowledge can become increasingly decontextualized only through extensive experience, practice, and feedback in a given domain. It is not until this decontextualization takes place that knowledge cannot be transferred to enhance general reasoning, problem solving and decision making in a broad domain (Bransford, Brown, & Cocking, 1999; Klein, Kuh, Chun, Hamilton, & Shavelson, 2005).

Second, the definitions appear to be grounded on different theoretical perspectives. At least three perspectives can be identified. One comes from the knowledge gathered over years about the nature of jobs and work (e.g., Peterson et al., 2001); another comes from the field of organizational psychology (as is the case of adaptability; Pulakos et al., 2000, 2002); and the third one comes from the field of cognitive psychology (e.g., expert thinking; see Levy & Murnane, 2004). These perspectives cannot always be aligned. Take as an example, non-routine problem solving. Should its focus be regarded as a matter of expert thinking only, should processes and procedures proposed in problem-solving be included (see Polya, 1945), or should it be regarded as a combination of both?[1] Should the name be non-routine or ill-structured or ill-defined?

Third, some of the skill dimensions identified by the committee or in this paper have received more attention than others. For example, there is empirical evidence in support of the existence of distinct dimensions of adaptability (Pulakos et al., 2000, 2002). Furthermore, there is evidence that different jobs (e.g., attorney, managers) involve different adaptability dimensions (Pulakos et al., 2000), and that the different dimensions are linked to different personality traits (e.g., Borman & Motowidlo, 1997; Pulakos et al., 2000). However, less is known about the dimensions involved, for example, in systems thinking. What criteria should be considered in deciding which dimensions to include?

Fourth, both the set of general definitions provided in the workshop and the definitions provided by Hilton (2008) are silent about technology-related skills, despite that it is clear that the development of skills is a result of the impact that technological change has on the demands within the global competition for jobs (Hilton, 2008). How, then, should technology be incorporated in the definition of the skills?

Thus, it seems that there is a need for a discussion of what exactly constitutes 21st century skills. Among the issues that need to be addressed in this discussion are: Is the term, *skills* the most appropriate to use in this context? Would it better to use the term, abilities, knowledge *and* skills? Reasoning, critical

---

[1] Processes and procedures proposed in problem-solving have been researched in work performance research (e.g., information gathering, generation of ideas; see Mumford & Peterson, 1995; Peterson et al, 1999).

thinking, problem-solving, decision making, and communicating, terms included in the definitions of the 21st century skills, have been referred to as abilities in other frameworks (Shavelson & Huang, 2003). It is worth mentioning that "skills" is only one of the multiple descriptors (e.g., abilities, knowledge, work styles) used in the Occupational Information Network (O*NET) to portrait the nature of a job (see Peterson, Mumford, Broman, Jeanneret, Fleishman, 1999; Peterson et al., 2001). To complicate matters, the O*NET document uses terms other than *skills* (e.g., work styles) to refer to some of the 21st century skills. As it can be seen, there are basic issues concerning labeling that need to be addressed to clarify what is meant with skills.

Comparing Perspectives

Just as the *Committee of Research Evidence Related to Future Skills Demands* defined five broad skills, other committees and organizations have defined their conceptions of 21st century skills. I compare the five broad skills, which I have named *NRC's 21st century skills*, with those proposed by other committees and organizations. I focus on three well-known national documents, *Partnership for the 21st Century Skills*, *The American Association of School Librarians (AASL) Standards for the 21st-Century Learner*, and *enGauge 21st Century Skills for 21st Century Learners*.

This comparison across frameworks focuses on mapping those skills and dimensions considered by *NRC 21st century skills*, which are not necessarily considered by other frameworks. The comparison is based on the descriptions of the dimensions provided in the various documents. When the descriptions were similar or equivalent, they were considered as being aligned across frameworks (i.e., *NRC's 21st century skills* and the framework at hand). The lack of specificity in the skills or dimensions mentioned, sometimes made it difficult to clearly determine how they are aligned. Therefore, when the descriptions were not clearly similar or equivalent, they were classified as "not aligned." This comparison is intended to allow identification of commonalities and differences across the frameworks and to allow identification of dimensions across skills that can help better conceptualize them. Appendix A shows the mapping of *NRC's 21st century skills* with those proposed by other frameworks. I first describe each organization briefly.[2]

Partnership for the 21st Century Skills. The Partnership for 21st Century Skills has emerged as the leading advocacy organization focused on infusing 21st century skills into education (http://www.21stcenturyskills.org/). The organization brings together the business community, education leaders, and policymakers to define a powerful vision of education for the 21st century and to ensure that students emerge from our schools with the skills needed to be effective citizens, workers, and leaders in

---

[2] It is important to mention that the mapping of the skills was done only by the paper's author; that is, there is no information about the consistency of the mapping with another person. It is up to the reader to judge the accuracy of the classification.

the 21st century. The mission of this partnership is to serve as a catalyst to position 21st century skills at the center of US K-12 education by building collaborative partnerships among education, business, community, and government leaders. Some members of this organization are Adobe Systems, Inc., ASCD, American Association of School Librarians, Dell, Inc., Education Testing Service, and Hewlett Packard, to mention only few of its members. The partnership proposes three general skills: (I) Learning and Innovation skills (i.e., creativity and innovation, critical thinking and problem solving, and communication and collaboration); (II) Information, Media, and Technology Skills (i.e., information literacy, media literacy, information, communication, and technology-ICT-literacy), and (III) Life and Career Skills (e.g., flexibility and adaptability, initiative and self direction, socials and cross-cultural skills, productivity and accountability, leadership and responsibility).

The American Association of School Librarians (AASL) Standards for the 21st-Century Learner. AASL is an organization that addresses issues, anticipates trends, and sets the future agenda for the profession. The AASL mission is to advocate excellence, facilitate change, and develop leaders in the school library media field (http://www.ala.org/ala/mgrps/divs/aasl/index.cfm). AASL proposes four general skills (http://www.ala.org/ala/mgrps/divs/aasl/aaslproftools/learningstandards/AASL_LearningStandards.pdf). Learners should use skills, resources, and tool to: (I) Inquire, think critically, and gain knowledge (e.g., follow an inquiry-based process in seeking knowledge); (II) Draw conclusions, make informed decisions, apply knowledge to new situations, and create new knowledge (e.g., organize knowledge so that it is useful); (III) Share knowledge and participate ethically and productively as members of our democratic society (e.g., participate and collaborate as members of a social and intellectual network of learners); and (IV) Pursue personal and aesthetic growth (e.g., seek information for personal learning in a variety of formats and experience). Standards are organized in four aspects: skills, dispositions in action, responsibilities and self-assessment strategies.

EnGauge 21st Century Skills for 21st Century Learners. The North Central Regional Educational Laboratory and the Metiri Group (2003) identified four general skills and developed profiles of students who have achieved the skills. Skills are described through different dimensions: (I) Digital-Age Literacy (i.e., basic, scientific, economic, technological, visual, information, multicultural, and global); (II) Inventive Thinking (e.g., adaptability and managing complexity, curiosity, creativity and risk taking); (III) Effective Communication (e.g., teaming, collaboration, and interpersonal skills, interactive communication); and (IV) High Productivity (e.g., prioritizing, planning, and managing for results, ability to produce relevant high-quality products).

The mapping shows that the different frameworks are tackling very similar skills despite the use of different categories and dimension names. The *systems thinking skill* was hardly tapped by the frameworks. This is probably one of the skills that would require more work to reach a better conceptualization. In contrast, *communication* and *non-routine problem-solving* were heavily tapped across frameworks. Thus, a closer analysis across frameworks should help to develop a more complete definition of these skills and to determine the relative standing of these different skills. Also, in order to pave the way for a clearer, future conceptualization, the status of "not aligned" given in this paper to certain skills and dimensions should be carefully assessed.

Overall, the Partnership for the 21st Century Skills framework seems to be the one that best aligns with *NRC's 21st century skills*. The AASL's Standards for the 21st-Century Learner does not tap adaptability at all, but aligns well with communication and non-routine problem-solving. Although the enGauge 21st century skills could be tapped with *NRC's 21st century skills,* there were fewer dimensions in common than with the Partnership for the 21st Century Skills framework.

Hopefully, the comparison among frameworks can help to better define NRC's 21st century skills not only from the perspective of the workforce—the competencies with the highest demands in the world of work—but also based on how these competencies can be linked to standards for teaching, learning, and assessment—how the standards-based education can support these skills).

NRC's 21st Century Skills and the Technology Standards

Whereas NRC's *document* is silent about technology, it is clear that technology should play an important role in the definition, instruction, and assessment of 21st century skills (see for example, the Secretary's Commission on Achieving Necessary Skills, *What Work Requires of School*, 1991; Levy & Murnane, 2004). Two of the national technology frameworks are also compared with *NRC's 21st century skills*: ISTE's National Educational Technology Standards and NRC's Technological Literacy Dimensions. [3] The comparison procedure was the same as the one described before. Appendix B presents the comparison. As before, I provide first a short description of the organizations proposing such standards.

National Educational Technology Standards. These standards are proposed by the International Society for Technology Education (ISTE), a non-profit membership organization whose mission is to provide leadership and service to improve teaching, learning, and school leadership by advancing the effective use of technology in PK–12 and teacher education (http://www.iste.org/AM/Template.cfm?Section=NETS). ISTE represents more than 85,000 professionals worldwide. The organization proposes six general standards (http://www.iste.org/Content/NavigationMenu/NETS/ForStudents/2007Standards/NETS_for_Students_2007_Standards.pdf): (I) Creativity and Innovation (e.g., apply existing knowledge

[3] ITEA's Standards for Technological Literacy were not mapped due to the large number of subcategories within each of the standards. The mapping of the standards only would not provide the picture needed.

to generate new ideas, products, or processes); (II) Communication and Collaboration (e.g., communicate information and ideas effectively to multiple audiences using a variety of media and formats); (III) Research and Information Fluency (e.g., plan strategies to guide inquiry); (IV) Critical Thinking, Problem Solving, and Decision Making (e.g., identity and define authentic problems and significant questions for investigation); (V) Digital Citizenship (e.g., advocate and practice safe, legal, and responsible use of information and technology); and (VI) Technology Operations and Concepts (e.g., understand and use technology systems).

Technological Literacy Dimensions. The Committee on Assessing Technological Literacy (see Garmire & Pearson, 2006) defined technological literacy in its most fundamental sense, as a general understanding of technology. The goal of technological literacy "is to provide with the tools [people] need to participate intelligently and thoughtfully in the world around them" (Garmire & Pearson, 2006, p. 33). The Committee used basically the same dimensions of technological literacy described in *Technically Speaking: Why All Americans Need to Know More About Technology* (NAE & NERC, 2002) but renamed the second dimension: (I) Knowledge (e.g., recognizing the pervasiveness of technology in everyday life); (II) Critical thinking and decision making – formerly, ways of thinking and acting (e.g., asking pertinent questions, of self and others, regarding the benefits and risks of technologies); and (III) Capabilities (e.g., can identify and fix simple mechanical or technological problems at home or at work).

Standards for Technological Literacy. These standards were proposed by the International Technology Education Association (ITEA) and its Technology for All American Projects (TfAAP). ITEA is a professional organization for technology, innovation, design, and engineering educators. Its mission is to promote technological literacy for all by supporting the teaching of technology and promoting the professionalism of those engaged in this pursuit. The standards proposed by this organization define what students should know and be able to do in order to be technologically literate. ITEA appointed teams, committees, and various groups of educators, engineers, technologies, and others for the standard developing process, a process that took about three years. The standards were reviewed in more than 60 schools nationwide.

Mapping the National Educational Technology Standards and the Technological Literacy Dimensions led to similar conclusions as before—complex communication and non-routine problem solving are the skills most frequently tapped. Also, whether technological literacy should be embedded in the NRC's 21st century skills is an issue. While technology is a critical part of the skills in the other frameworks (Partnership for the 21st century skills, AASL's Standards for the 21st-Century Learner, and enGauge 21st Century Skills for 21st Century Learners), it is emphasized in different ways. How should NRC's 21st century skills incorporate technology, if at all?

NRC's 21st Century Skills in the Context of Science Education

What is the intersection between the development of *NRC's 21st century skills* and the goals pursued in science education? Is there any alignment between these two sets of learning targets? What needs to be measured from the 21st century skills in the context of science education? Answers to these questions are the initial point in defining the constructs to be measured. I start by explaining science education goals and looking for critical dimensions that can help to bridge both contexts the 21st century skills and science education. By mapping the science dimensions to the 21st century skills, I aim to contextualize science education goals within the *NRC's 21st century skill.*[4]

Strands of Scientific Proficiency

The broad learning goals of science education have been named recently as *strands of scientific proficiency* (Duschl, Schweingruber, & Shouse, 2007). These strands "…address the knowledge and reasoning skills that students must eventually acquire to be considered fully proficient in science" (p. 36). Students who understand science are expected to:

I. Know, use, and interpret scientific explanations of the natural world

This educational goal focuses on the conceptual structures or schemes proposed by Schwab (1962) and Brandwein (1962). The strand is about building an organized and meaningful "society of ideas" that is used to interpret, construct, refine explanations, arguments, and models (Duschl, Schweingruber, & Shouse, 2007). "This strand includes acquiring facts and the conceptual structures that incorporate those facts and using these ideas productively to understand many phenomena in the natural world. This includes using those ideas to construct and refine explanations, arguments, or models of particular phenomena" (Duschl, Schweingruber, & Shouse, 2007, p. 37).

As proposed by Brandwein in 1962, conceptual schemes are not to be learned at once. They should be learned over a long period of time. Concepts are combined with other concepts into even larger patterns and put to the test when students are confronted with contradictions. Major changes in students' conceptual frameworks (or schemes) are often difficult because they require students to break their old frameworks and reorganized them (conceptual change).

The development of conceptual knowledge – not isolated definitions, but systems of concepts that are linked into the kind of rich, interconnected knowledge structures—should be a critical science educational goal (Lehrer & Schauble, 2006). Proficiency in this goal or strand is reflected on changes in the organization of knowledge, not just the accretion of more pieces of knowledge. Developing a coherent

[4] Whether the reader would agree with the dimensions identified or its categorization is an open question. However, the reader should consider that this was as a trial rather than a final set of dimensions.

organization of knowledge increases the likelihood of applying that knowledge properly and learn new, related materials more effectively.

II. Generate and evaluate evidence and explanations

This strand of scientific proficiency focuses on "the knowledge and skills needed to build and refine models based on evidence" (Duschl, Schweingruber, & Shouse, p. 37). It focuses on those activities related with the practices of scientific inquiry processes and places a heavy emphasis on posing questions and hypothesis, designing experiments, gathering and analyzing data, and constructing evidence-based arguments (e.g., Chinn & Malhotra, 2002; Kuhn, Black, Keselman, & Kaplan, 2000; Khishfe & Abd-El-Khalick, 2002). Heuristics and skills such as what to measure, developing measures, documenting outcomes, organizing and interpreting data, and developing and refining models are considered important targets for science education because they are assumed to be widely applicable and transferable. Within these targets, controlling variables in experimentation to test and rule out competing hypotheses appears to be a basic learning strategy in this strand.

In the context of scientific inquiry, reasoning processes and conceptual knowledge are interdependent and in fact facilitate each other. I argue that in this strand the triad of problem, method, and theory becomes important; "… the effective teaching of the physical sciences as inquiry becomes possible in a particular and important sense once we understand that the conclusions of science are closely linked with the inquiry which produced them, and, conversely, that the nature of a given inquiry depends on upon the topic of investigation" (Rutherford, 1964, p. 80). Two aspects are important to consider in this strand:

Goals and its Relation with Hypothesis and Experimental Designs. The goal(s) that students set for themselves in a science task has a large effect on their ability to discover new concepts (Dunbar, 1993). If students do not have clarity of the goal(s) they are pursuing, they are unable to formulate adequate hypothesis (explanatory mechanisms) needed to design an investigation and to account for their investigation findings, which then affects their understanding of the concept to be learned during the investigation activity (or a series of activities).

Data and Models. The lack of a goal and a hypothesis that guide the students' activities may in turn affect how they interpret the data (as confirming or disconfirming evidence of their hypotheses) and how this evidence is treated (e.g., ignored). Research has concluded that students (although it also applies to scientists) have a number of biases (Alexander, 2006; Dunbar, 1993): (a) conduct experiments that will confirm rather than disconfirm their current hypothesis, (b) ignore evidence inconsistent with their current hypothesis, (c) distort evidence to fit their current hypothesis, and (d) fail to consider alternative explanations.

Data interpretation (what to do after data collection) is another critical strategy that students need to learn. They need to evaluate what raw data become evidence, how evidence can be represented to show patterns, how these patterns can be converted into models, and how these patterns and models fit or do not fit the scientific theory and explanations (Duschl, 2000, 2003).

It is also important for students to recognize that *data* entails a separation between the world and a representation of that world (Lehrer & Romberg, 1996). The creation of a symbolic structure in correspondence with the world, but not in that world, is an essential characteristic of models (see Lehrer & Romberg, 1996). For example, recording the incidence of a form of teacher questioning in the classroom entails detaching and representing an element of the continuous activity in the classroom. The separation between elements of the model and elements of the world begins the process of *modeling*, but the concept of data goes beyond these elements and requires a structure (i.e., data requires construction of a structure to represent the world; different structures entail different conceptions of the world). Many practicing social scientists have had the experience of structuring data in one form only to find that different structures would have been better for answering their question (e.g., considering other attributes of the outcomes/variables observed/measured, or presenting a different form of data structure that could be more revealing; Lehrer & Romberg, 1996).

III. Understand the nature and development of scientific knowledge

This strand focuses on what Brandwein (1962) named "*the scientist's way*" of knowing which involves: *a challenge* – the universe can and should be investigated; *a faith* – the universe is in a state of change, but still it has its uniformities, its continuities, its probabilities; *a way of life as a process* – inquiry; *an aim* – construction and dissemination of a meaningful world; *a posture* – individual liberty, or idiosyncrasy, in mode and method of investigation; *and humility* – the capacity to self-correct one's own conclusions, "the knowledge that the only certainty is uncertainty" (p. 113). In terms of the National Research Council, "Students who understand scientific knowledge recognize that predictions or explanations can be revised on the basis of seeing new evidence or developing a new model" (Duschl, Schweingruber, & Shouse, p. 37).

The nature of science is concerned with the values and epistemological assumptions underlying the activities involved in scientific processes (Khishfe & Abd-El-Khalick, 2002). For example, observing and hypothesizing are scientific processes. However, understandings that observations are constrained by our perceptual apparatus, that hypothesis generation is necessarily imaginative and creative, and that both activities are inherently affected by the theory selected involves notions of the nature of science. Although there is an overlap and an interaction between science processes and the nature of science, it is nevertheless important to tell one from the other. Following Kuhn, (1989), Driver, Asoko, Leash, Mortimer, & Scott (1994), and Duschl (2000, 2003), I postulate five principles of what has been named, the nature and development of scientific knowledge (NDSK).

a. Scientific knowledge *is a social and cultural construction* – something that people do and create; it involves the creation and invention of facts, theories, explanations, models, etc. in which social interactions and argumentation are involved (Giere, 1988; Latour & Woolgar, 1979; Khishfe & Abd-El-Khalick, 2002; Sandoval as cited by Duschl, Schweingruber, & Shouse, 2007)
b. Scientific knowledge *is empirical* – based on and/or derived from observations of the world (Khishfe & Abd-El-Khalick, 2002). Scientific thinking begins with careful observation of the world around us and gathering pertinent data in a systematic and accurate way (Alexander, 2006).
c. Scientific knowledge *is tentative and uncertain* – it is subject to change based on new evidence (see Brandweing, 1962; Schwab, 1962)
d. Scientific knowledge *results from diverse forms of scientific practice* – there is no scientific method that applied to all scientific inquiries (Lehrer & Schauble, 2006; Sandoval, as cited in Duschl, Schweingruber, & Shouse, 2007).
e. Scientific knowledge *varies in its explanatory and predictive power* (e.g., theories, laws, hypotheses; Sandoval, as cited in Duschl, Schweingruber, & Shouse, 2007)

IV. Participate productively in scientific practices and discourse

"This strand includes students' understanding of the norms of participating in science as well as their motivation and attitudes toward science. Students who see science as valuable and interesting tend to be good learners and participants in science. To engage productively in science, however, students need to understand how to participate in scientific debates, adopt a critical stance, and be willing to ask questions" (Duschl, Schweingruber, & Shouse, 2007, p. 37).

This strand focuses on two critical aspects, the power of talk and argumentation and the culture of science. Scientists talk frequently with their colleagues, formally and informally to be engaged in discussions and share ideas and observations in a myriad of ways, all with the purpose of validating existing ideas and generating new ideas (Michaels, Shouse, & Schweingruber, 2008). Argumentation, the art of discussion, in Schwab's (1962) terms, is critical to this strand. Furthermore, students should be encouraged to learn and use the precise vocabulary (e.g., observe, predict, and check) that makes inquiry processes more visible to them and, hence, more open to inspection and self-evaluation (Lehrer &

Schauble, 2006). Also, they should communicate in certain ways that are specific to the field (see Bazerman, 1998; Halliday & Martin, 1993; Martin, 1993, 1998).

Mapping the Strands of Scientific Proficiency to the 21st century skills

Measuring *NRC's 21st century skills* in the context of science education, necessarily trespasses one of the characteristics of the 21st century skills; it becomes domain-specific rather than focusing on general abilities. Somehow, it consists of moving the target from abstract oriented to concrete-content oriented processes (Klein et al., 2005). It should be assumed, then, that performance on the assessment tasks will be influenced more by domain specific knowledge and skills, than general reasoning abilities. The skills, then, are reflected somehow in a different form and level.

To map the science education goals to *The NRC 21st century skills,* I identify dimensions based on the general descriptions provided in the previous section. Dimensions are mapped to each of the five skills. It is important to mention that the dimensions were defined considering a level of specificity that was expected to facilitate the development of the assessments but considering the critical aspects identified in the general descriptions.[5] For example, it should be expected that students who have a coherent organization of knowledge, a system of concepts linked into interconnected knowledge, should be able to develop explanations through scientific reasoning with principles or explanatory models (Li & Tsai, 2007). Then, the dimension, "explaining or justifying why something happens," will entail constructing explanations using scientific evidence, and understanding scientific theories, models, or principles would require of a coherent knowledge structure.

Table 2 presents the mapping of the science dimensions to the 21st century skills. The dimensions presented are not intended to be exhaustive; rather, they are just examples of the type of analysis that can be conducted to fully map these two frameworks. I acknowledge that the identification and the mapping are the product of a sole person. Different persons would arrive to different dimensions and different mapping. Whether the reader would agree with this mapping and the dimensions used is an open question.

[5] It goes beyond of the usual "what students *know* and *are be able to do*" in a particular domain after instruction (see Messick, 1984).

Table 2. *Mapping Science Education Goals to NRC's 21st century skills* *

The NRC 21st Century Skills					Not Aligned (Not Considered Equivalent or Difficult to Align)
Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	
IV. Participate productively in scientific practices and discourse - Being open to inspection and self-evaluation - Being open to communicate ideas in certain ways	**II. Generate and evaluate evidence and explanations** - Selecting appropriate ways to organize and represent data - Selecting appropriate ways to represent patterns - Selecting appropriate ways to communicate explanations - Being fluent in multiple modes of representation, visual or verbal formal or informal (verbal and non-verbal; e.g., words, pictures, graphs, formulas, tables, figures) **III. Understand the nature and development of scientific knowledge** - Collaborating with others to shape knowledge, to define what and how should be communicated, represented, argued, and debated - Writing and talking using the language of science as well as its forms of representation **IV. Participate productively in scientific practices and discourse** - Sharing ideas to validate them, to make them open to evaluation and inspection - Sharing ideas, procedures, observations, results in numerous ways to make them more visible to a network of participants - Participating in scientific debates, adopting a critical stance, and asking clarifying questions when necessary - Making and evaluating arguments based on evidence - Persuading and negotiating based on evidence	**I. Know, use and interpret scientific explanations** - Demonstrate relationships among different representation of principles - Explaining or justifying why something happens - Developing explanations through scientific reasoning with principles or explanatory models - Applying principles to solve problems **II. Generate and evaluate evidence and explanations** - Identifying and posing questions - Generating and/or testing hypotheses according to a goal pursued and the theory - Designing and conducting studies that can provide evidence that has the potential to confirm and disconfirm formulated hypotheses - Gathering pertinent data, systematically and accurately - Representing data in multiple forms and modes - Interpreting data in different forms and modes - Identifying patterns - Analyzing and synthesizing evidence - Formulating reasonable explanations from patterns observed, drawing conclusions **III. Understand the nature and development of scientific knowledge** - Predicting using evidence and/or models - Distinguishing patterns of evidence that do and do not support conclusions - Evaluating the quality of the evidence based on the methodology used - Revising explanation based on new evidence - Acknowledging alternative explanations	**III. Understand the nature and development of scientific knowledge** - Having an inquiry mind or a desire to discover the whys or hows of events or objects - Asking questions derived from curiosity about everyday experiences		

* Roman numbers correspond to each strands of scientific proficiency.

Few dimensions were mapped into adaptability and self-management/self-development. It was also difficult to define dimensions that could be mapped to the systems thinking skills. In contrast, the mapping of dimensions in complex communications and non-routine problem-solving was easier. A closer look at Table 2 makes it clear that the dimensions or *learning performances*--if we use the term Wilson and Bertenthal's terms (2006)—can be considered as a way to specify the construct, the working definition of what is to be measured. They specify, in learning performances terms, what students should be able to do, which in turn can help to define what will be considered as evidence of the proficiencies in the students' responses.

Finally, it is important to notice that the scientific proficiencies are not "crossed" with science content since this topic is out of the scope of this paper. Science content domains (e.g., physics, biology) are to be tapped in the content dimension. However, considering that the purpose is to tap *NRC's 21st century skills* and that the content in which these skills will be embedded can be selected from different domains (e.g., physics, chemistry, biology), two recommendations come handy. The first recommendation is, define clear criteria that can guide the selection of the content knowledge. Content knowledge selected should: (1) be relevant to real life situations – some science knowledge is more useful in the everyday life than other (OECD, 2006; Shavelson, 2007), (2) represent core scientific concepts with lifelong utility (OECD, 2006); and (3) be appropriate to the developmental level of the students to be tested.

The second recommendation is based on the reflections that resulted from the analysis of the *NRC's 21st century skills:* Include explicitly, in the content dimension, the *domain of technology* (e.g., the role of technology design in solving problems; Garmire & Pearson, 2006; WestEd & CCSSO, 2005; OECD, 2006). As human beings, we live in interconnected worlds—natural,social, and technological (Garmire & Pearson, 2006). Therefore, we use the artifacts created through technology for adaptability (e.g., learning how to use e-mail or other technologies), complex communications (e.g., using e-mail; Levy & Murnane, 2004), or self-management/self-development (e.g., for gathering information or self-study). We can use technology to solve a problem (e.g., designing and developing tools or equipment to solve a scientific problem; Garmire & Pearson, 2006; Mumford, Peterson, & Childs, 1999). Therefore, it seems appropriate to include the technology domain.

21st Century *Science* Skills in the Context of Large-Scale Assessment Frameworks

In this section, I map the 21st century *science* skills to well known large-sale assessment. The purpose of this mapping is to identify possible strategies for measuring the skills. I focus on the National

Assessment for Educational Progress (NAEP), Trends in Mathematics and Science Study (TIMSS), Programme for International Students Assessment (PISA), and Collegiate Learning Assessment (CLA).[6]

Large-scale assessment frameworks consider at least two dimensions: a *content dimension* – the subject matter dealt with in a field of study, and a *cognitive dimension* **-** the intellectual (cognitive) processes that we would like students to engage in when dealing with a given content. Whereas the content dimension is usually organized in domains (e.g., life science, chemistry, physics, earth science, and environmental science), the cognitive dimension is organized around domains that represent cognitive processes. Cognitive domains also vary by content domain. For example, in 2003, TIMSS proposed three cognitive domains in science (i.e., factual knowledge, conceptual understanding, reasoning and analysis) and four in mathematics (e.g., knowing facts and procedures, using concepts, solving routines problems and reasoning). For the purposes of this paper, the most critical dimension to consider is the cognitive dimension, a name which may not be the same across frameworks. In what follows, I discuss the dimensions of each of these large-scale assessment programs. Appendix C shows the mapping.

National Assessment for Educational Progress. National Assessment of Educational Progress (NAEP) is a congressional mandated project of the U.S. Department of Education's National Center for Education Statistics. Since 1969, NAEP has surveyed student achievement. Initially, NAEP tested students of ages 9, 13 and 17; since 1980, NAEP tests students in grades 4, 8, and 12.

The framework for developing assessment tasks has evolved over the years with attempts to maintain some conceptual continuity. For example, in the 2005 Framework for the Science Assessment (Council of Chief State School Officers [CCSSO], National Center for Improving Science Education [NCISE], & American Institutes for Research [AIR], 2004), the content dimension included three *Fields of Science*: Earth, Physical, and Life Sciences. The framework merged technology within each of the content areas and included *The Nature of Science* (also part of the 1991 NAEP Science Framework) and Themes (i.e., systems, models, patterns of change) as categories that should be integrated within the content domains. The cognitive dimensions (named *Knowing and Doing*) considered three aspects: (I) Conceptual understanding, (II) Scientific investigation, and (III) Practical reasoning (CCSSO, NCISE, AIR, 2004). The 2009 NAEP Science Framework (WestEd & Council of Chief State School Officers, 2005) content domain taps essentially the same three *broad science areas*: physical science, life science,

[6] Unfortunately, no concrete information could be found about the dimensions WorkKeys measures. Access to the information required a pay. Still, in what follows I provide minimum information about this assessment. WorkKeys Assessment System was developed by ACT, and it is one four components of a comprehensive system for measuring, communicating, and improving the common skills required for success in the workplace (i.e., job profiling, training, and research). WorkKeys measures skills that employers believe are critical to job success and that are valuable for any type of occupation, and at any level of education (ACT, 2008). WorkKeys groups basic workplace skills into eight skill areas: reading for information, applied mathematics, locating information, applied technology, writing, listening, observation, and teamwork. Items are mainly multiple-choice.

and earth and space science. However, in this framework, the cognitive demands are presented as four *science practices*: (I) Identifying science principles, (II) Using science practices, (III) Conducting scientific inquiry, and (IV) Employing technological design. It also describe the cognitive demands posed to students by assessment tasks.

Trends in International Mathematics and Science Study (TIMSS). The International Association for the Evaluation of Educational Achievement (IEA) has conducted comparative research studies for the last 45 years focusing on educational policies, practices, and outcomes (Mullis, Martin, Ruddock, O'Sullivan, Arora, & Erberber, 2005). In the 2007 framework, the content domain varies across the grades tested. At Grade 4, three major domains were covered: life science, physical science, and earth science; at Grade 8, the domains covered are: biology, chemistry, physics, and earth science. The cognitive dimension is divided into three domains, based on "what students have to know and do when confronting the various items" (Mullis et al., 2005, p. 68): (I) Knowing (facts, procedures, and concepts in a domain), (II) Applying (the ability to apply knowledge and conceptual understanding in a problem situation), and (III) Reasoning (which includes problem solving in unfamiliar situations, complex contexts, and multi-step problems). Although the cognitive domains are the same across grades, the percentage of items across domains varies between grades.

Programme for International Students Assessment (PISA). The Programme for International Students Assessment (PISA) was created by the Organisation for Economic Co-Operation and Development (OECD) in 1997. In 2006, scientific literacy was the major domain assessed, and therefore, the conceptualization of the assessed construct was more elaborated, compared to PISA 2000 and 2003. Unlike other previous assessments programs, PISA's framework is guided by what citizens require. It focuses on the following questions, "As citizens, what knowledge is more appropriate for 15-year-old? What is important for citizens to value about science and technology? What is important for individuals to be able to do that is science related? Can citizens distinguish claims that are scientifically sound from those that are not?" (OECD, 2006). Therefore, the development of the assessment was around the idea that "understanding of science and technology contributes significantly to the personal, social, professional and cultural lives of all people" (OECD, 2006). The content domains in the PISA 2006 are approached through two components, *knowledge **of** science* and *knowledge **about** science*. Specifically, four content domains tapped the knowledge of science: physical systems, living systems, earth and space systems, and technology systems. Two content domains tapped the knowledge about science: scientific inquiry (how scientists collect data) and scientific explanations (how scientists use data). The cognitive domains (named *competencies*) assessed are broad and include aspects that relate to personal utility, social responsibility, and the intrinsic and extrinsic value of scientific knowledge. PISA 2006 focused on three competencies that clarify what 15-year-old students should know, value, and be able to do within reasonable and appropriate personal, social and global contexts" (OECD, 2006, p. 21): (I) Identify scientific issues, (II) Explain phenomena scientifically, and (III) Use scientific evidence. Finally, it is important to mention that the PISA 2006 assessment of science also assessed student *attitudes* (interest in science, support for scientific inquiry, and responsibility towards resources and environments). It was assumed that the competencies required students to demonstrate knowledge and cognitive abilities, but attitudes required students to demonstrate values and motivations. In my analysis, I focus only on the competencies and one aspect of attitudes, support for scientific inquiry, which becomes the fourth competency (IV) mapped.

Collegiate Learning Assessment (CLA). The Council for Aid to Education (CAE) and RAND Corporation developed the Collegiate Learning Assessment (CLA). One of CLA's goals is to measure "value added" by educational programs in colleges and universities—the contribution made by these programs to the students learning in the areas tested. There are three programs: (1) CLA, designed for four-year institution, the Community College Learning Assessment (CCLA) designed for community colleges, and the College and Work Readiness Assessment (CWRA) designed for secondary schools. CLA taps abilities that are considered to be applicable to a wide range of academic majors and are also valued by employers (Klein, Freedman, Shavelson, & Bolus, in press). Therefore, its framework is based only on cognitive domains, but CLA uses direct measures of ability in which students actually perform cognitively demanding tasks based on which quality of performance is scored. CLA measures the following skills: critical thinking, analytic reasoning, problem-solving, and written communication (Klein et al, in press). These skills are assumed to be intertwined, and therefore are measured holistically. They have been named *collective outcomes* (Hersh, nd); they require students to use them in combination to respond to the assessment tasks, which are drawn from a real-world domain of activities (see Shavelson, 2007). For scoring purposes, the skills are divided in two categories: (I) critical thinking, analytic reasoning, and problem solving (i.e., evaluation of evidence, analysis and synthesis of evidence, drawing conclusions, and acknowledging alternative explanation/viewpoint); and (II) written communication (i.e., presentation, development, persuasiveness, mechanics, and interest).

The mapping (Appendix C) across the different large-scale assessment framework shows that one of the NRC's 21st century skills–non-routine problem solving–is tapped by all the programs analyzed, at least based on the descriptions. What this analysis is missing is the information about the quality of the items for each of these assessment programs (an issue that I discuss in the next section). It is important to point out that most of the assessment programs do not focus on complex communication; only CLA emphasizes this skill. Not all the programs focus on strands 3 and 4 of scientific proficiency, Understanding the nature and development of scientific knowledge and Participating productively in scientific practices and discourse; NAEP 2006, PISA 2006, and CLA (Klein et al, 2005) do. It is evident that problem-solving is a critical skill to be measured across programs. In the next section, I propose a framework for assessing 21st century skills in the context of science education.

Towards a Framework for Developing Assessments for Measuring 21st Century Skills in the Context of Science Education

Defining the Construct: A Most Likely Controversial Perspective

A necessary first step in the development of any assessment, whether we deal with classroom or large-scale assessments, is defining the construct. A construct is a working definition of what is to be measured. It is the concept or the characteristics that an assessment is designed to measure (AERA, APA, NCME, 1999). A definition of a construct can be broad (e.g., measuring scientific literacy) or can be specific (e.g., measuring students understanding of relative density) (Wilson & Bertenthal, 2006). It can be domain-specific (e.g., NAEP-Science) or it can focus on general abilities such as verbal, quantitative and visual-spatial reasoning (e.g., SAT-I). The construct of interest in this paper is the 21st century skills and more specifically, the 21st century skills in a science education context.

Critical to defining a construct is the question, "What knowledge, skills or attributes should be assessed?" In the case at hand, "What does it mean to achieve 21st century skills in the context of science?" Based on the information collected through the different analyses and comparisons across assessment frameworks thus far, I propose to consider three domains to describe the skills in the context of science education: *dispositional factors* (personal characteristics that affect performance), *cross-functional skills* (cognitive activities likely to occur in any domain), and *scientific proficiencies* (knowledge of specific subject-matter, science-based knowledge). Table 3 provides a graphical representation of the domains. The construct domains have to be thought of as a hierarchy. That is, each of the domains of the construct (e.g., dispositions) has dimensions or sub-constructs (e.g., adaptability), and each of them has their definitional characteristics. I explain each of the domains and dimensions.

Table 3. *Construct Domains*

Dispositional Factors	Cross-Functional Skills	Content-Specific Knowledge
Adaptability Self-Management/Self-Development	{ Problem solving Complex Communication }	**Science Content and Practices** Knowledge, use, and interpretation of scientific explanations Generate and evaluate evidence and explanations Understand the nature and development of scientific knowledge Participate productively in scientific practices and discourse

Dispositional Factors

Dispositional factors focus on recurrent behaviors of affective tendency that distinguishes one person from another one (VandenBos, 2007). These dispositions are expected to affect student performance. In the context of work research, this category has been named *work styles* (Borman, Kubisiak, & Schneider, 1999). Following the 21st century skills proposed by NRC, two constructs were considered under this dimension: adaptability and self-management/self development. Both skills are actually personality constructs (Borman, Kubisiak, & Schneider, 1999). Adaptability is basically about flexibility - coping with frequently changing work environments (e.g., new technology), and tolerance of diverse kinds of stress. In the context of science education, adaptation may refer to interpersonal adaptability (see Table 1), for example, when working in diverse small groups with different students on different occasions. Self-management/self-development involves striving for competence in one's work, expending effort at getting better, persisting in the face of obstacles, setting high standards, and wanting to improve. In the job realm, these dispositional factors are measured mainly with 7-point scale questionnaires (e.g., Pulakos et al., 2000, 2002; Borman et al., 1999).

Cross-Functional Skills

Some of the skills proposed by NRC, can be conceived as being broad enough that can be applied across subjects domains (e.g., mathematics or history; Klein et al., 2005; Klein, Benjamin, Shavelson, & Bolus, 2007), as well as across jobs (Mumford, Peterson, & Childs, 1999). That is, they are cross-functional skills or broad abilities (Shavelson & Huang, 2003). Due to the role they play across dimensions, I considered two of the skills proposed by NRC as cross-functional: problem solving and complex communication.

Problems solving skills. This term refers to the cognitive processes and procedures that people apply when confronted with problems and lack a clear way to solve them. For example, we can say that we may confront a problem when we try to adapt to an unpredictable situation.

Problem solving is cognitive processing directed at achieving a goal (solving a problem) when there is no solution that is obvious to the problem solver (Mayer & Wittrock, 2006). Terms as thinking, reasoning, creative thinking, and critical thinking are considered synonyms here, although fine distinctions among them can be made.[7] It is assumed that problem solving involves different cognitive

[7] For example, thinking is a broader concept that includes both, directed and undirected cognitive processing (e.g., day dreaming). Reasoning, "in the strictest sense, refers to directed cognitive processing applied to certain class of tasks – that is, reasoning tasks in which there are premises and the goal is to derive a conclusion using logical rules – and requiring a certain class of cognitive process" (Mayer & Wittrock, 2006, p. 287). Finally, creative thinking and critical thinking are considered two aspects of problem solving (Mayer & Wittrock, 2006). Creative thinking involves generating ideas that can be used to solve a

processes (Alexander, 2006; Mayer & Wittrick, 2006): (1) *representing a problem* or building the problem space (the distance between someone's current or initial state and the desired state); (2) *planning*, which involves devising a method for solving a problem (e.g., breaking a problem into small parts); (3) *monitoring*, which involves evaluating the appropriateness and effectiveness of the solution; (4) *executing* or carrying out the plan; and (5) *self-regulation*, that involves modifying or sustaining cognitive activities towards the achievement of the goal – solving the problem.

Complex Communication. This skill should be considered part of a higher order thinking skill, social skills (Mumford, Peterson, & Childs, 1999). Social skills are a set of learned abilities that enable individuals to interact completely and appropriately in any given social context (VandenBos, 2007), be it teamwork in the classroom, costumer service, or working at a distance. This skill, as a higher-order thinking skill, entails another subset of skills, such as social perceptiveness or coordination (see Mumford, Peterson, & Childs, 1999). However, given the importance provided to communication in NRC's 21st century skills and other frameworks (e.g., Partnership for the 21st Century Skills, CLA), it seems to be appropriate to have it as a stand-alone skill. Complex communication is a cross-functional skill because it is involved in both the domain-specific knowledge and dispositional factors dimensions. For example, complex communication is important within the strands of scientific proficiencies, and it is also important in adaptability (e.g., demonstrating interpersonal adaptability by listening others' viewpoints; Pulakos, et al., 2000, 2002).

Content-specific knowledge. To be able to develop 21st century skills in the context of science education, students should develop science knowledge and skills that are structured in ways that can contribute to their ability to think and reason with what they know and can do. They need to know important facts within the science domain, they should be able to identify objects and phenomena, apply routine and non-routine and simple and complex procedures, and carry out and design investigations (Li & Tsai, 2007). All this knowledge must be more than isolated elements, or "beads on a chain" (Shavelson & Ruiz-Primo, 1999). This knowledge needs to be organized in structures that allow students to develop schemas (e.g., problem schemata and solution schemata; Alexander, 2006) that make it possible for students to use, interpret, judge, explain, and generate strategies to solve novel problems. That is, they require different types of knowledge that can be integrated into a meaningful whole (Li, 2001).

We (Li, 2001; Li, Ruiz-Primo, & Shavelson, 2006; Shavelson & Ruiz-Primo, 1999; Ruiz-Primo, 1997, 1998, 2002) have proposed to approach achievement within the science domain based on the notion

---

problem, whereas critical thinking involves evaluating ideas that could be used to solve a problem. This conceptualization of problem solving applies, then, to any domain.

of types of knowledge.[8] This approach, then, conceptualizes achievement in science as a multi-faceted construct. An underlying assumption in this conception is that the types of knowledge proposed reflect to certain degree the nature of subject-matter expertise. It rests on the idea that expertise is necessarily constrained to a subject matter or content domain. Evidence shows that expert's knowledge and skills are not easily transferable across domains (Chi, Feltovich, & Glaser, 1981; Chi, Glaser, & Farr, 1988). That is, types of knowledge lie in a continuum from concrete to abstract, from bits of information to high levels of organized knowledge. Therefore, it should be assumed that higher levels of achievement are linked to certain types of knowledge (Li, Ruiz-Primo, & Shavelson, 2006).

Our notion identifies four types of knowledge for assessment purposes: declarative knowledge or *knowing that*, procedural knowledge or *knowing how*, schematic knowledge or *knowing why*, and strategic knowledge or *knowing when, where, and how knowledge applies*. The framework has been empirically tested in the context of science assessments with confirming results around the categories proposed (Li, 2001; Li, Ruiz-Primo, & Shavelson, 2006; Li & Shavelson, 2001; Li & Tsai, 2007). This four-type knowledge notion was considered as a framework in the 2009 Science NAEP Framework (WestEd & CCSSO, 2005).[9] Still, the complete type-of-knowledge notion (Ruiz-Primo, 1997), considers a fifth type of knowledge linked less to domain-specific knowledge and more to general cognitive processes, metacognitive knowledge or *knowing about one's cognition*.

Declarative knowledge - Knowing that. This type includes knowledge that ranges from discrete and isolated content elements such as terminology, facts, or specific details to a more organized knowledge forms such as statements, definitions, knowledge of classifications, and categories (e.g., vocabulary such as mass or density, or mass is a property of an object, or density is expressed as the number of grams in 1 cubic centimeter of the object).

Procedural knowledge - Knowing how. This type involves knowledge of skills, algorithms, techniques, and methods. Usually, it takes the form of if-then production rules or a sequence of steps (e.g., measuring temperature using a thermometer, applying an algorithm to balance chemical equations. add, subtract, multiply, and divide whole numbers). It ranges from motor procedures (e.g., folding a filter paper and placing it in a funnel or classify objects by size), to simple application of a well-practiced algorithm (e.g., subtracting two numbers), to a complex procedure (e.g., implementing a procedure to find out which substance is the most soluble or applying strategies to estimate the results of rational-number computations). Procedural knowledge involves the knowledge of techniques, and methods that are the result of consensus, agreement, or disciplinary norms (e.g., Anderson, L. et al., 2001). It involves how to

---

[8] The framework, as it has evolved, has been described in diverse papers (e.g., Li, 2001; Li, Ruiz-Primo, & Shavelson, 2006; Ruiz-Primo, 1997, 1998, 2002; Ruiz-Primo, Shavelson, Hamilton & Klein, 2002; Shavelson & Ruiz-Primo, 1999;).

[9] A more detailed and complete description of each knowledge types of provided in Li and Tsai (2007).

complete a task and it has been viewed as a skill (e.g., Roger, Ciscero, & Carlo, 1993). Procedural knowledge can be automatized over many trials (practice) allowing retrieval and execution without deliberate attention. Automaticity is considered as one of the key characteristics of expertise (e.g., Anderson, J. R. 1983).

Schematic knowledge - Knowing why. This type involves more organized bodies of knowledge, such as schemas, mental models, or "theories" (implicit or explicit) that are used to organize information in an interconnected and systematic manner. This form of organization allows individuals to apply principles or explanatory models to approach a problem (troubleshooting), provide an explanation, or predict an outcome (e.g., explaining why we have seasons; De Kleer & Brown, 1983; Gentner & Stevens, 1983). For example, combined with procedural knowledge, schematic knowledge is involved, in the process of reasoning from several theories to design experiments.

Strategic knowledge - Knowing when, where, and how to apply knowledge. "The application of strategic knowledge involves navigating the problem, planning, monitoring, trouble-shooting, and synchronizing other types of knowledge. Typically, strategic knowledge is used when one encounters ill-defined tasks" (Li & Tsai, p. 14). It includes domain-specific strategies, such as ways of representing a problem or strategies to deal with certain types of tasks. It also entails such general monitoring performance or planning strategies as dividing a task into subtasks, reflecting on the process to explore alternative solutions, knowing where to use a particular piece of schematic knowledge, or integrating the three other types of knowledge in an efficient manner. It is important to mention that strategic knowledge, a higher-order knowledge, is based on the other three forms of knowledge (e.g., Anderson, L. et al. 2001; Li, 2001). An attempt to focus only on strategic knowledge without a strong base for the other forms of knowledge does not support transfer to new situations (Mayer, 1997; Pellegrino, 2002). Unlike the other three types of knowledge, "strategic knowledge … is applicable to a wider variety of types of problems within a domain" (de Jong & Ferguson-Hessler, 1996, p. 107).

Metacognitive knowledge – Knowing about one's cognition. This type involves knowledge about one's cognition and the regulation of one's own cognition (Alexander, Schallert, & Hare, 1191). It includes awareness and control of one's own cognitive processes; reflecting on and evaluating one's own thoughts and learning. There are two aspects of metacognition: knowledge about our own cognition and knowledge about how to regulate one's cognition (with metacognitive strategies; Alexander, 2006).

Figure 1 shows science items that tap different types of knowledge within the same domain. It provides examples of science multiple-choice items used in a study on formative assessment (Shavelson & Young, 2000). All the items shown focus on density and sinking/floating. In what follows, I describe the item classifications by referring to four types of item characteristics (task demand, cognitive, demand, item openness, and item complexity; see a more detailed description in next section).

Example (a) can be thought as an item tapping declarative knowledge: First, the response is expected to be in the form of a definition of a term (e.g., what is density). This item asks for a very specific content question (i.e., a definition), leaving students little opportunity to provide relations between concepts or to apply principles. Second, the cognition elicited is likely to consist of recognizing information. Note that the item is similar to school-type problems and, consequently, when answering it, students may identify exactly what they have been taught. The cognitive process involved in answering the item consists of directly retrieving information or performing a minimum of scientific reasoning to sort out relevant information. Third, in terms of item openness, the item is restricted in the sense that it is a multiple-choice item and its unfinished stem forces students to select options instead of responding to a question prior to reading any options. Being restricted, in turn, reinforces the task and cognitive demands placed on students. Finally, the coding of complexity does not add new information for the classification.

Example (b) was classified as tapping procedural knowledge. It provides students with a table with information about the mass and volume of different objects. First, the item requires students to use an algorithm to calculate density and compare the different densities to arrive at the solution. These two pieces of knowledge fall into the category of procedural knowledge as defined in this paper. Second, the cognitive process in which students are likely to engage consists of applying a calculation algorithm in which mass is divided by volume to calculate density. Although this item allows students to generate their own responses before reading the options, students can arrive at the correct answer by working backwards from the options or even merely guessing. The analysis of complexity does not provide additional information.

Example (c) taps schematic knowledge. First, the item asks students to provide a prediction be based on an understanding of the concept of density. Second, the dominant cognitive process is reasoning with theories or a mental model. It goes beyond the formula or how to apply it. An individual who can calculate density correctly every time, may not respond this item correctly if a deeper understanding has not been reached. Finally, the item does not involve unnecessarily heavy reading or irrelevant information. The low complexity strengthens the posited link to schematic knowledge by reducing construct-irrelevant variances.

Density equals

A. buoyancy divided by mass.
B. buoyancy divided by volume.
C. volume divided by mass.
D. mass divided by volume.

(a)

Which object listed in the table has the greatest density?

Object	Mass of Object	Volume of Object
W	11.0 grams	24 cubic centimeters
X	11.0 grams	12 cubic centimeters
Y	5.5 grams	4 cubic centimeters
Z	5.5 grams	11 cubic centimeters

A. W
B. X
C. Y
D. Z

(b)

Ball A and Ball B have the **SAME** mass and volume. Ball A is solid; Ball B has a hollow space in the center, which accounts as half of the volume of the ball (see the pictures below).

Ball A subsurface floats in water. When placed in water, ball B will__________

A. sink.
B. float.
C. subsurface float.
D. not sure.

(c)

Figure 1. Examples of items focusing on different types of knowledge: (a) declarative, (b) procedural, and (c) schematic.

Unfortunately, because it is difficult to track strategies used by students when solving problems, it is very difficult to find items that assess strategic knowledge. So much so, that the lack of items tapping this type of knowledge can be considered as an indicator of what is taught and what is tested. These examples made it clear that the type of item does not necessarily reflect the complexity of the cognitive process involved. Clearly, linking types of assessment to types of knowledge is not straightforward. Testing method alone does not determine the type of knowledge measured by an item (Bennett, 1993; Martinez, 1999; Snow, 1993).

Some Implications of Considering Types of Knowledge.

What would be the importance of defining types of knowledge in assessing 21st century skills? As discussed above, certain types of assessment tap better than others certain types of knowledge. The notion of type of knowledge is helpful in the development or selection of assessments. It is important to mention that this notion is linked to the construct of problem solving. Problem-solving processes are dependant of types of knowledge (Mayer & Wittrock, 2006). Representing and framing a problem depends largely on facts and concepts (declarative knowledge), schemas (schematic knowledge), planning on strategic knowledge, executing on procedural knowledge, and monitoring and self-regulation on metacognitive knowledge.

The notion of type of knowledge has at least three implications. It helps to: (1) profile assessments (e.g., What is being measured by a particular test?); (2) make meaningful interpretations of students' scores (e.g., What exactly does a student's score represent?); (3) and design or select assessment tasks that are aligned with instruction (e.g., What types of assessment tasks can lead to know whether students understand the concept of density? What are the cognitive demands that need to be imposed on students?).

An example can help understand these implications. We (Li, Ruiz-Primo, & Shavelson, 2006) analyzed the characteristics of TIMSS 1999 items, Science Booklet 8, and classified them according to the types of knowledge (89% inter-coder agreement). Our results indicated that the TIMSS-R Booklet 8 science test was heavily loaded on declarative knowledge (approximately 50 percent) with balance of procedural and schematic knowledge questions equally distributed. Unfortunately, items were identified as assessing students' strategic knowledge. The pre-classification of the items was supported by a confirmatory factor analysis. Most factor loadings were statistically significant and generally high. The fit and the factor loadings supported the feasibility of using types of knowledge in analyzing the TIMSS science items. These findings support the notion that analyzing large-scale or classroom assessments with the notion of types of knowledge helps to identify what is being assessed.

These results lead to second important implication. Think about students who have exactly the same total score in an assessment, but the pattern of their responses is different. One student may arrive at the total score by answering correctly most of the items tapping declarative knowledge, whereas another one may respond correctly to those items tapping schematic knowledge. Therefore, using a single score to infer students' understanding may lead to invalid interpretations, thus ineffective instruction (Li, Ruiz-Primo, & Shavelson, 2006).

The third implication, designing and evaluating assessments, is explained in the next section, in which I propose an assessment development and evaluation model.

An Approach for Developing and Evaluating Assessments

The proposed approach is based on a *modified* version of the *assessment square* proposed by the Stanford Education Assessment Laboratory (Ruiz-Primo, 2003, 2007; Shavelson, Ruiz-Primo, Li, & Ayala, 2002). The assessment square builds on the cognition-observation-interpretation assessment triangle proposed in the seminal report by Pellegrino, Chudowsky, and Glaser (2001). The assessment square consists of four components, one at each corner (1. construct; 2. observation model; 3. assessment;

and 4. interpretation), and involves three types of analysis (conceptual, logical, and empirical). The arrows in Figure 2 illustrate that the development and evaluation of assessments involves an iterative process in which the corners of the square loop back to earlier corners. This framework has been successfully used to develop and validate science assessments (e.g., Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Yin, 2005). In what follows, I explain and illustrate each of the components. I focus on the two cross-functional skills, problem solving and complex communications..

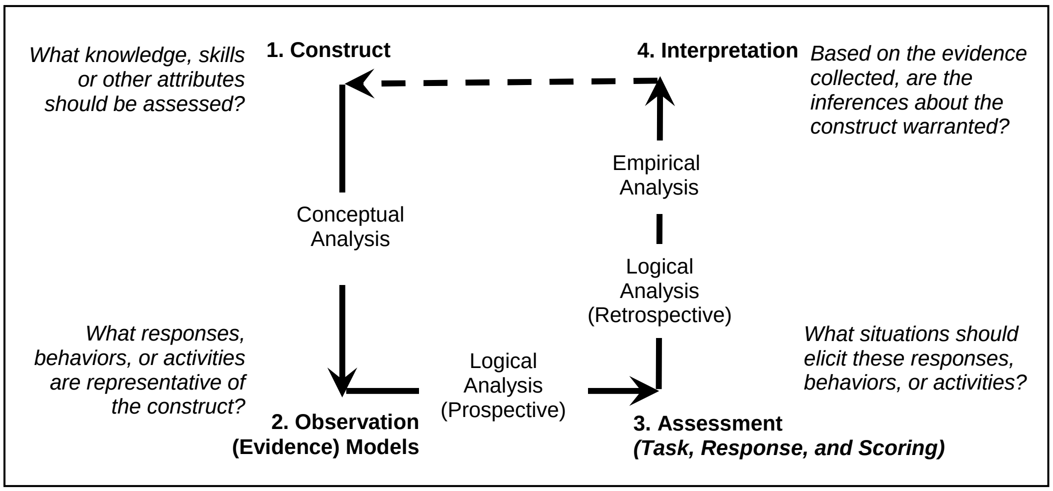


Figure 2. The assessment square (After Ruiz-Primo, 2003, 2007; Shavelson et al., 2002).

Construct. The assessment square begins with a working definition of what is being measured, the construct. The critical question to ask is, *What knowledge, skills or other attributes should be assessed?* Assuming that the cross-functional skills (problem solving and communication) are at the center of the assessment of the 21st century skills, I am proposing to focus on assessments that require both problem solving and communication skills. To tap the cross-functional skills to the content domain, I propose as the construct to be measured, the second strands of the scientific proficiencies *Generate and evaluate evidence and explanations* - the knowledge and skills needed to build and refine explanations and models based on evidence (Duschl, Schweingruber, & Shouse, 2007). I discuss this strand across the different components of the assessment square.

Observation (Evidence) Models. Observation models specify what, in a student's response, we will value as evidence of the construct. Observation models are evidence-based models which delineate, based on *conceptual analyses*, what constitute evidence of a construct (Mislevy & Haertel, 2006; Mislevy, Steinberg, Breyer, Almosnd, & Johnson, 1998). The conceptual analysis of Strand II was presented in a previous section (see Tale 2). Observation models provide information about the types of responses, behaviors, or activities that will be representative of the construct. They may specify different levels of the quality of responses. They help to delineate the tasks and response demands that the student will confront. Observation models are guided by the question, *What responses, behaviors, or activities are representative of the construct?* What do students need to do or respond allowing us to infer that they can generate and evaluate evidence and explanations? It can be argued that this strand of scientific proficiency focuses on the ability of students to apply what they know to:

- design an investigation (pose a problem, formulate an hypothesis, analyze the elements of the problem, select investigation methods) to conduct an investigation;
- criticize the diverse aspects of an investigation;
- conduct scientific investigations using appropriate tools and techniques;
- identify patterns in data by developing or selecting appropriate tools to interpret complex data sets and/or relate patterns in data to theoretical models; and
- use empirical evidence to validate or criticize conclusions about explanations and predictions

Logical Analysis. Logical analysis focuses on the coherent link between the observation models and the nature of the assessment (Ruiz-Primo et al., 2001). Assessment methods differ on what they ask students to do. Therefore, it is assumed that some methods are better suited to tap factual information (declarative knowledge), while some others are better suited to tap algorithms (procedural), and still others are better suited to test other aspects of understanding (schematic and strategic knowledge). The logical analysis helps to describe the features of the tasks and thus link type of assessment and type of knowledge. A critical question to ask when conducting a logical analysis is, *What are possible tasks that can* ***elicit*** *the behaviors, responses, and activities that were defined as* ***evidence*** *of the construct in the observation model? What features should the assessment tasks need to have in order for them to elicit the expected performance?* Good assessment tasks will be those that elicit relevant evidence about the measured construct.

Prospective logical analysis helps to describe the features of the situations to be used for eliciting the expected performance (e.g., responses, behaviors, skills) in the development of assessments. Prospective logical analysis leads to specifications intended to characterize and construct situations with which a student will interact to provide evidence about targeted aspects of knowledge. Once the logical analysis is completed, the tasks, response format, and the scoring systems have to be delineated to arrive at the third component of the square, the assessment. Figure 3 presents some types of assessment linked to the types of knowledge. Various assessment methods (e.g., multiple-choice, constructed-response, concept maps, performance assessments) can be linked to various types of knowledge through logical, cognitive analysis, and statistical model fitting. It has to be noted that the process is not straightforward; there is no perfect match between types of assessments and types of knowledge. Still, when considering the cognitive demands involved in certain assessments, it seems possible to conclude that some types of knowledge may be better tapped by certain assessment tasks based on the affordances and constraints provided by those tasks (Li, 2001; Li & Tsai, 2007).

Proficiency/Expertise → *Low* *High*	**Declarative Knowledge** *Knowing "that"*	**Procedural Knowledge** *Knowing "how"*	**Schematic Knowledge** *Knowing "why"*	**Strategic Knowledge** *Knowing "when, where and how"*
Extent *(How much?)*	• Multiple-Choice • Short-Answer • Constructed response	• Performance Assessments • Multiple-Choice	• Multiple-Choice • Performance Assessments • Constructed response • Predict-Observe-Explain (POE)	• Performance Assessments
Structure *(How is it organized?)*	• Concept Maps	• Procedural Maps?	• Maps, Diagram?	• Computer Simulations
Others *(How efficient and automatic?)*	• ?	• ?	• ?	• Computer Simulations

Figure 3. Linking types of assessments to achievement components.

Retrospective logical analysis, on the other hand, is used to analyze assessment tasks that have already been developed (e.g., analyze a TIMSS 2003 item). Once an assessment has been developed or selected for its use, its tasks and response demands can be analyzed logically, but in retrospect, to see if it falls within the construct domain, and if it is likely to elicit the expected behaviors from a student. The retrospective logical analysis involves reviewing how the task elicits the targeted knowledge and influences students' thinking and responding. This analysis posits cognitive activities that a task might evoke by examining the opportunities and constraints that the assessment task provides to students (Li, 2001; Li & Tsai, 2007; Ruiz-Primo et al., 2001). Based on our research work and that by other researchers (e.g., Anderson, L. et al., 2001; Ayala, Ayala, & Shavelson, 2000; Baxter & Glaser, 1998; Li, 2001; Li, Ruiz-Primo, & Shavelson, 2006; Quellmalz, 2002; Ruiz-Primo, Schultz, Li, & Shavelson, 1999; Ruiz-Primo, Shavelson, Li, & Schultz, 2001) we have proposed four aspects that can be used in retrospective logical analysis (Appendix D): (1) *task demands* - what students are asked to perform (e.g., define a concept or provide an explanation), (2) *cognitive demands* - inferred cognitive processes that students likely act upon to provide responses (e.g., recall a fact or reasoning with a model);[10] (3) *item openness* - gradation in the constraints exerted in the nature and extent of the response (e.g., selecting vs. generating responses or requiring information only found in task vs. information that can be learned from the task), and (4) *complexity of the item* – the diverse characteristics of an item such as familiarity to students, reading difficulty, and the extent to which it reflects experiences that are common to all students (e.g., ancillary skills, inclusion of relevant and irrelevant information, language demands).[11] A more complete and elaborated set of aspects to conduct logical analysis has been recently proposed by Li & Tsai (2007).

In the context of the 21st century skills and giving the relevance of the cross-functional skills to be assessed, I propose to focus here on assessment tasks that allow for solving problems with more complex characteristics. How can this complexity be manipulated? I propose to focus on the nature of the problem to be solved, the context, and the environments.

Nature of the Problem to be Solved. Table 4 presents examples of problem dimensions to be considered in the logical analysis in defining assessment tasks. Appendix E provides a more complete definition of the dimensions. A clear advantage of considering the proposed dimensions is to better conceptualize the assessment task problem space (Newell & Simon, 1972). They allow determining a problem *configuration or profile* based on the dimensions, which in turn will reflect the *complexity* of the problem, and this complexity will be associated with the nature of the solution paths. It should be assumed that problems with profiles on the right of the continuum should be more complex than those whose profiles are towards the left end of the continuum. Also, it should be expected that assessment tasks that are conceived at the right end will require higher order types of knowledge (e.g., schematic and strategic) than the merely applying procedures or identifying facts and concepts. It is important to remind the reader here that, in the use of schematic and strategic knowledge, procedural and declarative knowledge may be involved too.

---

[10] If assessment tasks are to tap higher-order cognitive processes, they must require that students cannot answer them correctly by relying on memory alone (Anderson, L. et al, 2001).

[11] "Any [assessment] task requires skills beyond those it is intended to measure" (Haertel & Linn, 1996, p. 63). Capabilities not explicitly part of what is to be measured, but nonetheless necessary for a successful performance are called "*ancillary or enabling skills*" (Haertel & Linn, 1996). Ancillary skills range from students' understanding of the assessment task requirements to their understanding that they should show their best work, to their level of reading, to the language used in the assessment tasks. Ancillary skills have an impact on the characteristics of both the task and the format used by the student to respond. Both affect the level of difficulty of the assessment task. Therefore, ancillary skills have an impact on the validity of the assessment task. Observed variations in the scores of equally capable students due to differences in their ancillary skills give raise to construct-irrelevant variance in assessment scores.

Table 4. *Problem Complexity Dimensions*

Dispositions	Cross-Functional Skills		General Description
	Nature of a Problem		
	Well-Structured ↔	Ill-Structured	***Structure***. Level of problem definition. Defined by the developer
	Routine ↔	Non-Routine	***Routine***. Solution procedures already learned? Depend on the examinee.
	Lean ↔	Rich	***Richness***. Number of activities & assessment subtasks involved
Adaptability & Self-Management	Schooled ↔	Unschooled	***Schoolness***. Obligatory academic exposure to solve the problem?
	Independent ↔ **OR**	Collaborative	***Collaboration***. Solution approached individually or with others?
	Un-timed ↔	Timed	***Time***. Is it a time constraint to solve the problem?
	Complex Communication		
	Extent of Communication		
	Selected or Short-Answer ↔	Constructed Response	Intensity of amount of writing

Types of Assessment Contexts and Environments. Assessment tasks can be embedded in diverse contexts and environments. To assess NRC's 21st century skills, selecting contexts considering the students' interests and lives may provide a useful approach for developing diverse scenarios. For example, PISA 2006 used three contexts to embed the assessment tasks: *Personal* – self, family, and peer groups; *Social* – the community; and *Global* – life across the world. When contexts are crossed with content domains (e.g., health, natural resources, environment), the possibilities for the development of assessments tasks become clear (see Table 5).

Table 5. *Contexts and a Sample of Content Domains for the PISA 2006 (OECD, 2006)*

Content Domain	**Context**		
	Personal (Self, family, and peer groups)	**Social** (The community)	**Global** (Life across the world)
Health	Maintenance of health, accidents, nutrition	Control of disease, social, transmission, food choices, community health	Epidemics, spread of infectious diseases
Natural Resources	Personal consumption of materials and energy	Maintenance of human population, quality of life, security, production and distribution of food, energy supply	Renewable and non-renewable, natural systems, population growth, sustainable use of species.

Compared to traditional paper-and-pencil delivery, technology offers more options, for presenting the diverse environments for assessments. Computer-based assessment tasks involve an environment with diverse tools for students to use to solve the problem at hand. Examples of these tools are databases, text editors, simulations, visual displays, interactive feedback, spreadsheets, and presentation and communication tools. Assessment developers can manipulate the number of tools need to use to solve the problem. Different problems may require different tools. This feature allows tapping skills involved in NRC's 21st century skills (e.g., examine a broad span of information and narrow; processing and interpreting both verbal and non-verbal information).

Tapping Dispositions and Content-Specific Knowledge. I submit that, by manipulating collaboration and time, it is possible to impose constraints on the assessment tasks which can create different situations to tap students' adaptability and self-management. It is important to consider that adaptability is a construct developed around the idea of the changing character of today's organizations. Therefore, it appears that, a condition for measuring adaptability is the capability to create a change in a situation; a change in the status of a given situation. Will the student adjust to a new situation, to new time constraints, or to work with others that s/he does not know? Will the student listen to others' viewpoints? Will the student manage her/his time properly?

Assessment. An assessment is a systematic procedure for eliciting, observing, and describing behavior(s), often with a numeric scale (cf. Cronbach, 1990). Since assessments are a manifestation of the working construct definition, they should be developed according to the observation models defined. An assessment is one of many possible manifestations of the construct in the form of a test that could have been produced. An assessment, then, can be thought as a "sample" from a universe of possible assessments that are consistent with the construct definition (Shavelson, Baxter, & Gao, 1993; Shavelson & Ruiz-Primo, 1999). We (Ruiz-Primo & Shavelson, 1996) have characterized an *assessment* as having three components: a *task* (eliciting), a *response format* (capturing or observing), and a *scoring system* (describing behavior, possibly with a numeric value). We have argued that, without these three components, an assessment is not properly defined.

Coming back to our example for assessing strand II, What assessment tasks elicit the responses, behavior or activities defined as evidence of generating and evaluating evidence and explanations? The development of an assessment should consider this question at every moment during the development process. An overall judgment of the observation model can lead to say that students can be required to *do* something or *critique examples* of the different aspects defined in the observation model. Based on the observation model and the logical analysis, it seems appropriate to focus on tasks that are towards the right end of the nature of the problem dimensions (see Table 4); that is, towards ill-structured or defined, non-routine, and rich problems. Schoolness, Collaboration, and Time are characteristics that can be manipulated in such a way that can tap, for example, adaptability and self-management.

Therefore, assessment methods such as essays and performance assessment tasks, seems to be good candidates to be considered in the measurement. Predict-Observe-Explain (POE) seems another possible type of assessment that combines written responses but without the burden of materials kits for each student. Still, the use of multiple-choice items of high quality is still desirable to tap declarative, procedural, and schematic knowledge. As mentioned, strategic knowledge items are difficult to find since it is hard to know the development and efficiency of the strategies used to solve the problem. However, thanks to the role that technology is playing now in testing (Bennett, Persky, Weiss, & Jenkins, 2007; Quellmalz & Haertel, 2004; Quellmalz & Pellegrino, 2009), the possibilities of having an assessment that can tap strategic knowledge and self management is a reality (Stevens, Johnson, & Soller, 2005).

In what follows, I provide some examples of items that can be used to measure what was defined in the observation model as evidence of generating and evaluating evidence and explanations. These items have been used in other large-scale assessments. The items are analyzed considering the different aspects defined in the logical analysis. I start with two examples that are not domain-content bounded, CLA assessments, followed by other examples that are domain-content bounded. The first three examples are assessments which focus "the criticizing" and the last two are assessments which focus on "the doing" of the behaviors, responses, or activities defined in the observation (evidence) models.

Assessment Examples

Example 1. Figure 4 provides an example of one of the two types of essay questions in the CLA assessment. This thirty-minute assessment task presents a real-life scenario. Students are asked to make an argument for or against the principal's decision.

Sample CLA Analytic Writing Task: Critique an Argument

A well-respected professional journal with a readership that includes elementary school principals recently published the results of a two-year study on childhood obesity. (Obese individuals are usually considered to be those who are 20 percent above their recommended weight for height and age.) This study sampled 50 schoolchildren, ages 5-11, from Smith Elementary School. A fast food restaurant opened near the school just before the study began. After two years, students who remained in the sample group were more likely to be overweight—relative to the national average. Based on this study, the principal of Jones Elementary School decided to confront her school's obesity problem by opposing any fast food restaurant openings near her school.

Figure 4. Analytic writing task: Critique an argument. (Source: Shavelson, 2008)

A retrospective logical analysis led to conclude that the assessment item is ill-structured. It is not possible for students to have a clear sense of what an acceptable answer would be or how to get to that answer. Whether it is a routine or non-routine problem is unclear. Examinees may or may not have a routine procedure to approach it. It is a lean problem, since it does not require conducting diverse activities to respond to this analytical writing task. However, the problem may require sub-tasks (e.g., to make a list of pros and cons first before writing the argument). This also seems to be an unschooled problem; that is, it does not require an academic procedure taught at the school. CLA administers this type of task individually, but it can also be a collaborative task. It is a-timed assessment; therefore, students need to self-monitor their time to finish it in 30 minutes. The context of the task can be considered "social," something that students can observe in their own community. According to the CLA framework (Klein et al, 2005, 2007; Shavelson & Huang, 2003) the assessment task presented in Figure 4 taps broad abilities not linked to domain-specific knowledge. Therefore, the types of knowledge proposed do not apply (Klein et al., 2005). It is argued that these abilities are developed well into adulthood through learning and transfer from non-school and school experiences. This prompt is expected to tap, then, the following complex cognitive processes: critical thinking, analytic reasoning, problem solving, and written communication, with an emphasis on the latter ability (other CLA prompts focus on more work-sample performance tasks). It is important to note that the students' responses are scored with the same dimensions defined for the strands of scientific proficiencies (e.g., evaluation of evidence, analysis and synthesis of evidence, drawing conclusions), as well as with some criteria related to the quality of the writing (e.g., development, presentation, development, persuasiveness). Similar tasks should be considered for assessing the 21st Century skills but using science-based scenarios.

Example 2. Figure 5 provides another example from CLA. This is a ninety-minute performance assessment task in which students are asked to pretend that they work for the company DynaTech and that they are asked by their boss to evaluate the pros and cons of purchasing a plane (called the "SwiftAir 235") for the company. Concern about this purchase has risen with the report of a recent SwiftAir 235 crash. Students respond in a real-life manner by writing a memorandum (the response format) to their boss analyzing the pros and cons of alternative solutions, anticipating possible problems and solutions to them, recommending what the company should do, and focusing on evidence to support their opinions and recommendations. In scoring performance, alternative justifiable solutions to the problem and alternative solution paths are recognized and evaluated.

A retrospective logical analysis lead to conclude that the assessment is ill-defined, since no clear information is given on the characteristics of the correct response (the right memo). Although routine procedures to solve the problem depend on the knowledge that the examinee brings to the situation, it is very likely that this assessment can be in the middle-right of the continuum since some sense on how to approach the problem is probably common across examinees (e.g., read sources of information provided). It is a rich problem, since examinees are given a library of information (in-basket-information) about the SwiftAir 235, in particular, and airplane accidents, in general, to evaluate the situation. Some of the information is relevant and sound, but some is not. Therefore, part of the problem is for students to tell relevant from irrelevant information. Students integrate these multiple sources of information to arrive at a problem solution, decision, or recommendation. The problem is an unschooled problem, it does not seem to require academic exposure to solve it and it requires to be approached individually rather than in a collaborative form. The problem is timed (giving the possibility to manipulate time for self-management). The context is job related.

You are the assistant to Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigational equipment. Sally Evans, a member of Dyna Tech's sales force, recommended that Dyna Tech buy a small private plane (a SwiftAir 235) that she and other member of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235. You are provided with the following documentation.

1: Newspaper articles about the accident
2: Federal Accident Report on in-flight breakups in single engine planes
3: Pat's e-mail to you & Sallys e-mail to Pat
4: Charts on SwiftAir's performance characteristics
5: Amateur Pilot article comparing SwiftAir 235 to similar planes
6: Pictures and description of SwiftAir Models 180 and 235

Please prepare a memo that addresses several questions, including what data support or refute the claim that the type of wing on the SwiftAir 235 leads to more in-flight breakups, what other factors might have contributed to the accident and should be taken into accont, and your overall recommendation about whether or not DynaTech should purchase the plane.

Figure 5. An example of a CLA performance task (Source: Klein et al., 2007)

Example 3. Figure 6 presents an example from PISA 2006, the *School Milk Study*. The first part of the item, Question 1 requires students to identify the possible purposes of the study, which can tap posing a scientific question to investigate, part of designing an investigation. Question 2 taps criticizing diverse aspects of the investigation. Both questions tap some of the behaviors, responses, and actions defined in the observation (evidence) models for gathering and evaluating evidence and explanations.

Based on a retrospective logical analysis, this problem can be defined as well-structured, since it is clear that the question has an acceptable answer or correct response. As with the other examples before, it is difficult to determine the routines dimension of the problem task. The routines may vary from student to student according to their experiences on identifying scientific problems and using evidence to support explanations. It should be expected, however, that this problem lays on the mid-left of routines if used with students who are under science education programs. The problem is lean, it does not require students to carry out different activities (such as judging diverse sources of information) or subdivide the problem in subtasks. This example can be classified as a schooled problem dealing with processes and procedures most likely experienced in an academic environment. It is an item to be responded individually, timed but in relation to other items. The context has an historical setting within a global context. There are no written communication demands.

School Milk Study

In 1930, a large-scale study was carried out in the schools in a region of Scotland. For four months, some students received free milk and some did not. The head teachers in each school chose which of their students received milk. Here is what happened:

- 5 000 school children received an amount of unpasteurised milk each school day
- Another 5 000 school children received the same amount of pasteurised milk
- 10 000 school children did not receive any milk at all

All 20 000 children were weighed and had their heights measured at the beginning and the end of the study.

Question 1: SCHOOL MILK STUDY

Is it likely that the following questions were research questions for the study?

Circle "Yes" or "No" for each question.

Is it likely that this was a research question for the study?	Yes or No?
What has to be done to pasteurise milk?	Yes / No
What effect does the drinking of additional milk have on school children?	Yes / No
What effect does milk pasteurisation have on school children's growth?	Yes / No
What effect does living in different regions of Scotland have on school children's health?	Yes / No

Question 2: SCHOOL MILK STUDY

On average, the children who received milk during the study gained more in height and weight than the children who did not receive milk.

One possible conclusion from the study, therefore, is that school children who drink a lot of milk grow faster than those who do not drink a lot of milk.

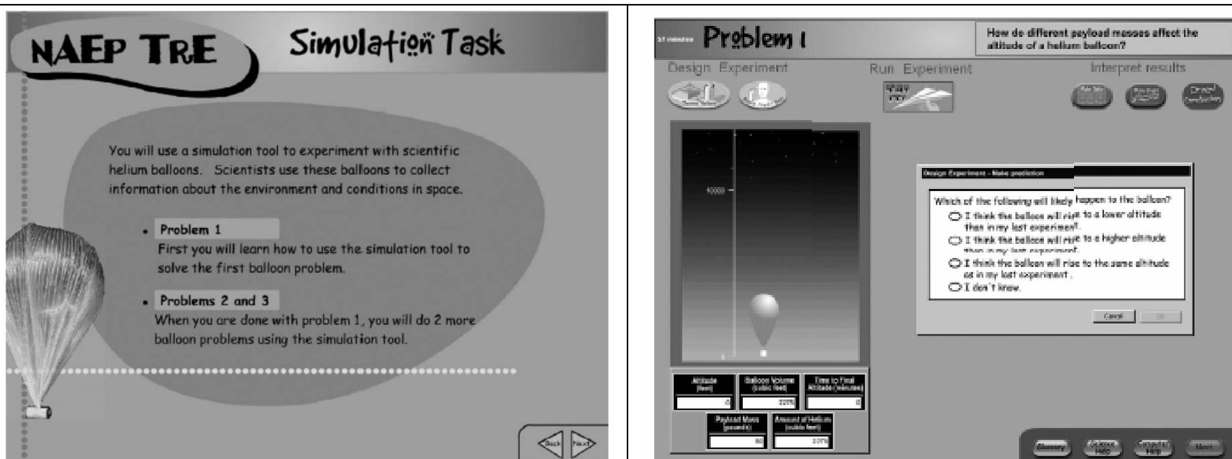
To have confidence in this conclusion, indicate one assumption that needs to be made about these two groups of students in the study.

Figure 6. School Milk Study PISA 2006 item. (Source: OECD, 2006.)

This item taps mainly declarative knowledge: First, students are asked to identify scientific questions. Samples of the research questions provided can be considered as instances of a class, scientific questions (Li & Tsai, 2007). Students then may recognize each instance as a member of that class. Second, the cognition evoked is likely to be recognizing information. Third, regarding item openness, the item is restricted in the sense that it is a selection type of item; therefore, it involves no complex written communication. Being restricted, in turn, reinforces the task and cognitive demands placed on students.

Example 4. Figure 7 provides an example of one of the Problem Solving in Rich Technology Environments (TRE), the last of three field Investigations in the NAEP Technology-Based Assessment

Project (Bennett, Persky, Weiss, & Jenkins, 2007). The context for the three problem-solving tasks was the domain of physical science: (1) to determine how the different payload masses affect the altitude of the balloon; (2) to determine the relationship between the amount of helium put in the balloon and the altitude that the balloon could reach; and (3) to determine the effect of different payloads in conjunction and helium in the altitude of the balloon. There are two scenarios for the problems. First, the TRE Search scenario requires students to locate and synthesize information about scientific helium balloons from a simulated World Wide Web environment. Second, the TRE Simulation scenario requires students to experiment to solve problems of increasing complexity about relationships among buoyancy, mass, and volume. Students can see animated displays after manipulating the mass carried by a scientific helium balloon and the amount of helium contained in the balloon.



(a) Introducing the problem to solve

(b) Computer screen with the prediction options in TRE Simulation scenario problem 1.

Figure 7. TRE assessment. (Source Bennett et al., 2005)

The analysis of this assessment is made considering the three problems students needed to respond. The problems are well-structured since there is clear sense that there is a correct answer. The routiness dimension of the problems are toward the left end of the dimension, although that may vary from student to student. How to gather information from the World Wide Web is more or less a routine procedure for students in schools. The simulation scenario is a fix procedure that seems to require changing the values of the different variables; what is important is which values to select on each trial The richness of the problems may also be similar and towards the right end of the dimensions, since in both cases there are several activities and sub-tasks to carry out (e,g., databases, text editors, simulation tools). It is possible that the problems can be located towards the middle left end of the schoolness dimension, rather than the right; despite the technicality of the terms used, the procedure itself may not require of a taught procedure. The problem is not from a real world context and it should be solved individually and with time constraints.

A retrospective logical analysis leads to conclude that this item taps mainly procedural knowledge: First, students are asked to carry out procedures to search information on the World Wide Web, or to carry out the simulation. Furthermore, students under the simulation scenario are trained to do so. Second, the cognition evoked is likely to be recognizing information; although data interpretation is required. Students do not have a choice on what forms of representation are the best to represent the data collected; it is given. Third, in terms of item openness, the item is restricted in the sense that students select from the tools available. Being restricted, in turn, reinforces the task and cognitive demands placed on students. Therefore, weighing the three characteristics, the item can be thought as tapping procedural knowledge. Some written communication is required

Measuring Strategic Knowledge. As mentioned above, it is difficult to find items that measure strategic knowledge. Is it then possible to measure strategic knowledge at all? With computer technology is now possible to track how students use information to solve problems and then to identify strategies used, the pattern followed. Therefore, it is possible to recognize whether the pattern followed by the student to solve the problem was a result or trial-and-error or a planned strategy that was carefully monitored. This would be an indicator of how students approach the solution of a problem based on the strategies known already by them, combining some of them to develop new and more efficient strategies.

An example of such technology is already available in the Interactive Multi-Media Exercises (IMMEX; Case, Stevens, & Cooper, 2007; Cooper, Stevens, & Holme, 2006). IMMEX is a system that presents students with real-world case-based complex problems that are solved in an online environment. The program tracks students' actions and data-mining strategies used to arrive to the solution. The system uses artificial neural networks and hidden Markov models to identify groups of strategies and pathways into specific strategy types into general problem solving states (Figure 8). IMMEX has proved to be a reliable and repeatable measures of students problem solving (Case, Stevens, & Cooper, 2007; Cooper, Sandi-Urena, & Stevens, 2008).

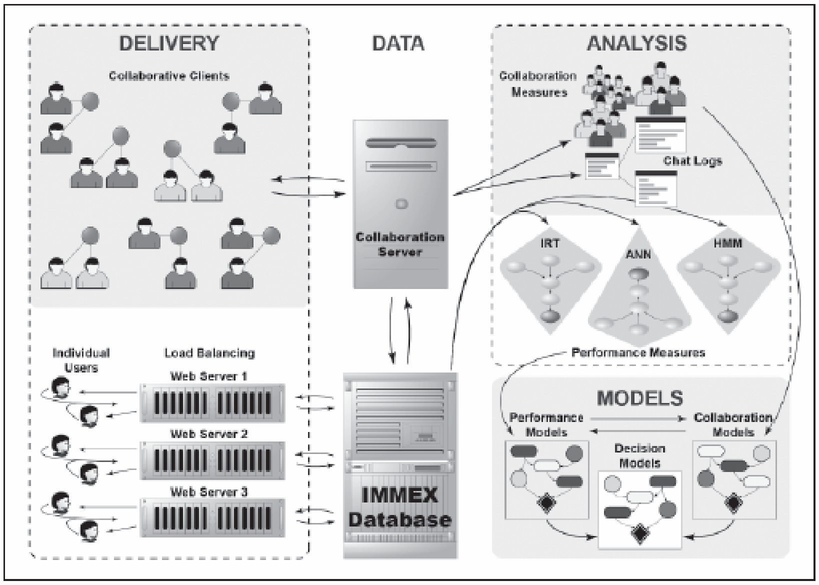


Figure 8. Overall architecture for IMMEX delivery and assessment modules. (Source: Cooper, Stevens, & Holme, 2006).

The program also offers the possibility for collaboration to solve a problem, which makes it possible to manipulate collaboration with colleagues in different rooms and requires effective communications skills, determine procedures to consider the opinion of others, and even easily tap adaptability and self-management. Figure 9 shows one of the problems posed to students.

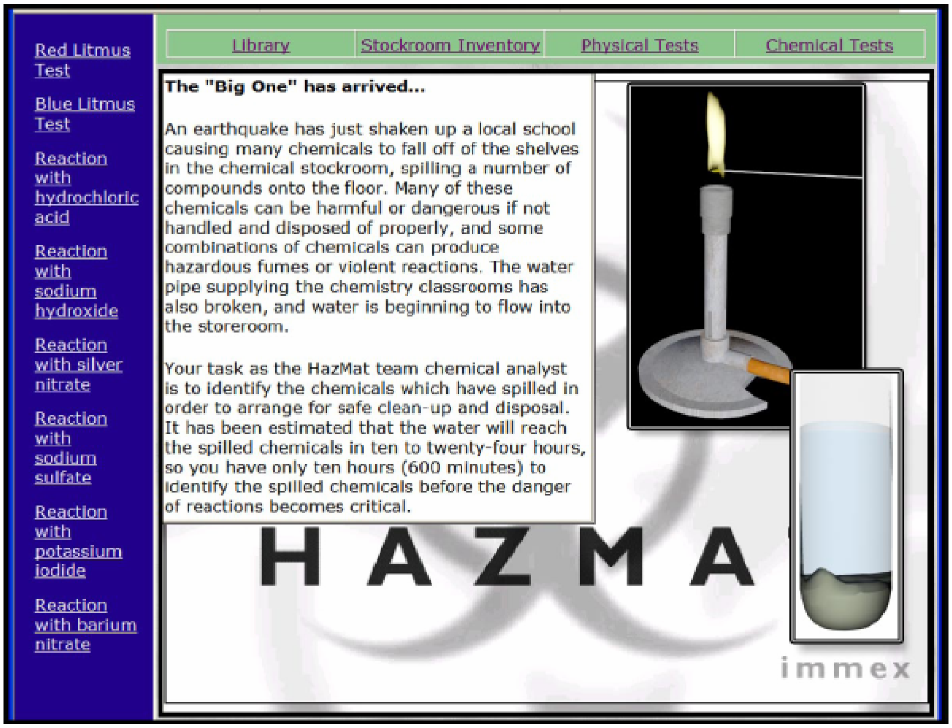


Figure 9. The prologue for Hazmat, an IMMEX qualitative inorganic analysis problem set.

Figure 9 shows an example of an IMMEX problem and Figure 10 shows an example of the solutions paths used to analyze the students' solving strategies.

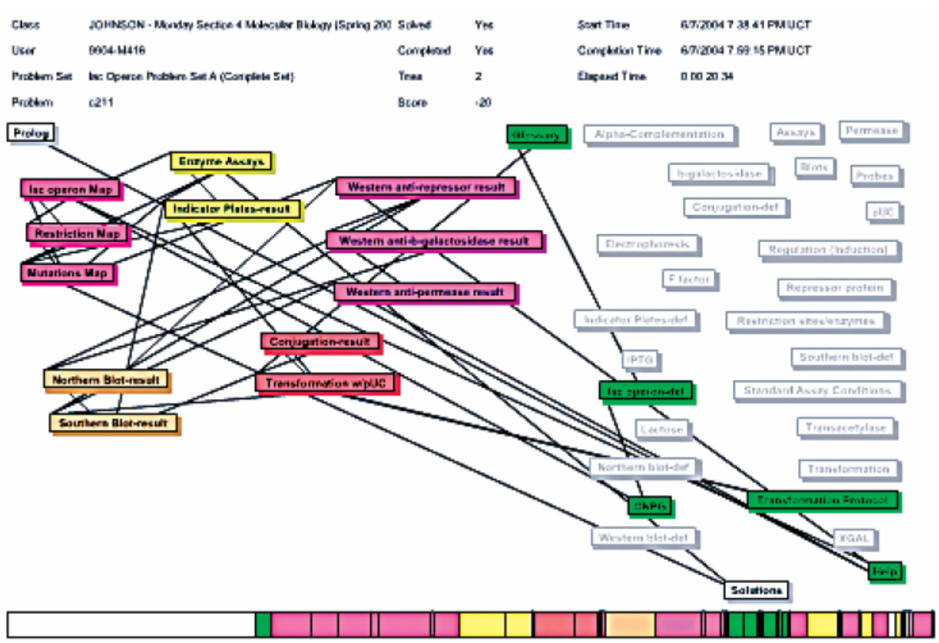


Figure 10. An IMMEX sample search path map.

The assessment tasks presented in the previous section show the options to be considered to assess critical 21st century skills. Although not all of them tap high-order knowledge, they are possible options to be considered in a large-scale context. Furthermore, some of these options make it possible to analyze the strategy path selected by the students, which can reflect how students put together new strategies from what they already know to approach novel problems. Furthermore, they can model the development of scientifically problem-solving strategies.

Empirical Analysis. This is the final type of analysis required in the assessment square. Empirical analysis involves collecting and summarizing students' responses to the assessment task. The analysis empirically examines both the assessment-evoked cognitive activities and the observed student performance. It includes the following evidence (Ruiz-Primo et al., 2001): (1) cognitive activities evoked as students perform on the assessment, (2) the relation between cognitive activities and performance levels (scores), and (3) statistical analyses of results (e.g., correlation among assessments measuring similar or different constructs, differences among groups). If we understand the link between assessments and cognitive processes, we expect to be able to design assessment tasks and responses to evoke different cognitive processes, different levels of performance, or both. The analysis brings the three types of evidence to bear on the link between the assessment and the construct definition.

We have proposed the collection of information about the cognitive processes elicited as a critical source of information in addition to the more traditional strategies used to evaluate the assessments. Cognitive analysis provides evidence on a student's cognitive activities that are evoked by the task as well as the level of performance (Ruiz-Primo et al, 2001). The analysis brings both to bear on the link between the assessment and the construct definition. We ask, does the assessment evoke the intended behaviors? Is there a correspondence between the intended behaviors and the performance scores? Cognitive validity can provide evidence on construct-relevant and -irrelevant sources of variance (e.g., Messick, 1995). For example, the characteristics of the assessment tasks may make students to respond in ways that are not relevant to the construct assessed (e.g., guessing), or may be too narrow that it fails to tap important aspects of the construct (e.g., the assessment task is too structured). Several methods can be used to examine cognitive processes. Messick (1989) recommends think-aloud protocols, retrospective reasons for answering, and errors made by examinees. Perhaps the method most widely used is the "think-aloud" or "talk-aloud" protocol method. The usual procedure is to ask examinees to think aloud while completing the assessment items. In most of the cognitive validity studies, after completing an item, students respond to interview questions. The combination of spontaneous think-aloud protocols and structured interview prompts allows students to respond to items without intervention at the same time that they provide information not given in the unstructured think-aloud format. Talk-aloud protocols and interviews are audiotaped and transcribed. In some studies (e.g., Kupermintz, Le, & Snow, 1999; Hamilton, Nussbaum, & Snow, 1997), interviewers use a structured observation sheet to record events that cannot be captured on audiotape, such as the use of gestures. This information is added to the session transcripts. Another method, less intrusive, is to listen to students working in dyads talk to one another as they tackle the assessment task. These interactions provide real-time verbalizations of students' thinking and corresponding actions.

Interpretation. This is the final corner of the assessment framework that brings together evidence from the logical and empirical analyses to assess the validity of the interpretations. For example, according to the purpose of the assessment, we could ask if the inferences about students' performance reflect differences in students' differential instructional experiences. During the development of an assessment, we iterate, somehow informally, through the assessment square until we have fine-tuned the assessment. In research and practice where learning is assessed, we formally evaluate the inferences.

Concluding Comments: Some Discussion Issues

This paper proposed a framework for developing suitable assessments to measure 21st century skills. The framework proposed considers an alternative way of conceptualizing 21st century skills, a notion of types of knowledge, and dimensions on the nature of the problems. The development process is based on an assessment square that has proved to be appropriate for developing reliable and valid assessments. The topic of this paper, the 21st century skills in the context of science education, is a complex one and require some discussion in order to develop suitable assessments. The development of the proposed approach led to some reflections around which future work should be organized:

1. Define the science context of the 21st Century Skills. Is it the case that a particular perspective of science instruction is better than another one to embed the 21st century skills? In this document, the scientific inquiry perspective was used. But was this the best approach?

2. Establish the criticality of the skills. Which of the 21st century skills are critical? Are there primary and secondary ones? Different frameworks tap different skills; some tap five, while other tap more than 20. The mapping of the different frameworks pointed to two critical ones: problem-solving and written communication. Is it the case that these two skills are the most fundamental? For large-scale assessment purposes, it seems that focusing on critical skills is better than trying to measure all.

3. Define assessment purposes. For what purposes should a large-scale assessment tapping the 21st century skills be developed? Is it to provide information for accountability, evaluation, or comparative purposes? Is it to focus on public and media attention on educational concerns? Is

it to change educational practice by influencing curriculum and instruction or by spurring greater effort on the part of school administrators, teachers, and students? (see Haertel, 1999). We know that different purposes lead to different sources of validity evidence. Thus, defining assessment purpose will help to better design the framework for developing quality assessments and for evaluating their technical quality.

4. Define an appropriate approach for Validity. Validity is the most fundamental consideration in developing and evaluating assessments (AERA, APS, NCME, 1999). Still, it seems that in practice, certain pieces of evidence are always expected and, therefore, these pieces are what is provided. The list, overall, is appropriate (e.g., documenting content coverage, reporting reliabilities and standard measurements errors, estimating correlations with relevant criteria). However, it is jus a list without attention to how the pieces come together to make the case for validity. Approaching validation as a coherent argument rather than a list of requirements should be encouraged (Haertel, 1999). Validation should be a process of constructing and evaluating arguments *for* and *against* proposed assessment interpretations and uses. Each argument involves assumptions which require support. Furthermore, massive evidence in support of one critical argument does not buy as much if there is no evidence for some other argument. It is important to understand that the "checklist approach" to validity has a powerful built-in bias (see Haertel, 1999) towards looking for supporting evidence, not for disconfirming evidence; what Cronbach (1988) named confirmationist bias; "the task of validation is not to uphold a test, practice, or theory. Ideally, validators will prepare as debaters do. Studying a topic from all angles, a debater grasps the arguments pro and con so well that he or she could speak for either side" (p. 3).

5. Define an appropriate approach for Reliability. In evaluating achievement assessment we have thought of assessments as a sample of student behavior (Ruiz-Primo & Shavelson, 1996; Shavelson & Ruiz-Primo, 2000; Shavelson, Baxter, Gao, 1993). Inferences are made from this sample to a "universe" of behavior of interest. From this perspective, a score assigned to a student is but one possible sample from a large domain of possible scores that a student might have received if a different sample of assessment tasks were included, if a different set of judges were included, and the like. A sampling framework is constructed by identifying the *facets* that characterize the measurement. Facets include the task presented, the occasion of measurement, the rates who judged the performance, and so forth. This

means that for a particular type of assessment, the relevant measurement facets may vary. For example, for multiple-choice tests the facet of raters is irrelevant, but task and occasion are relevant. For other assessments (e.g., performance assessments, predict-observe-explain) other combination of facets are relevant (see Ruiz-Primo & Shavelson, 1996). This means that for any particular type of assessment, a subset of facets may define the sample of behavior collected. Once a test score is conceived of as a sample of performance from a complex universe, statistical procedures can be brought to bear on the score's technical quality (e.g., classical reliability theory, item response theory, and generalizability theory; Cronbach, Gleser, Nanda & Rajaratnam, 1972; Shavelson & Webb, 1991).

6. Define an appropriate approach for Fairness. In the context of globalization, valid assessment of 21st century skills poses an additional challenge—linguistic, cultural, and socio-economic diversity. As experience is gained in international test comparisons such as TIMSS and PISA, efforts to improve procedures for translating and adapting instruments in international comparisons are continuously refined (Hambleton 2005). Most of the approaches created with the intent to address item bias are based on examining performance differences or differential item functioning between groups of students from a culture for which a test was originally created and a culture for which that test is adapted, or between students who are tested in a language in which the test was originally written and students who are tested in a language into which the test is translated. But these ex-post-facto approaches are costly and time consuming (see Allalouf, 2003). In addition, their proper implementation is often jeopardized by tight time lines in assessment development procedures, which limit the possibility of modifying and refining translated or adapted test items with the same detail as with their original versions. Procedures for examining test bias, linguistic, cultural, and socio-economic diversity should be addressed throughout the entire process of test development (Solano-Flores & Nelson-Barber, 2001)—a notion that is consistent with the notion, discussed above, that test validity is highly dependant of the integrity of the process of test development. Literature in the field indicates (Solano-Flores, 2009) three basic issues that need to be addressed, population specification (which involves, for example, accurate identification of students who belong to the different linguistic, cultural, or socioeconomic groups to be tested, and their appropriate representation in the samples of pilot students with which a test is tried out); task sampling (which, for example, requires that the process of selection of items includes procedures for examining to which extent the content and form of test items over-represent or under-represent the epistemologies of certain linguistic, cultural, or socio-economic groups or the situations with which these

groups are familiar); and generalization (which, for example, involves being caution about the extent to which test results for a given group of students are generalized to broad linguistic, cultural, or socio-economic groups.

The variety of tasks, types of technologies used for test administration, and formats of test administration that is inherent to assessing 21st century skills, underscores the limitations of current approaches for addressing diversity and speak to the need for research that examines the extent to which skills identified as 21st skills are relevant to multiple groups or nations, or the ways in which they are expressed in multiple languages and cultures and diverse socio-economic groups.

7. Issues on Practicality. It is clear that computer based technology makes the development, assessment implementation, and scoring more suitable for large-scale assessment. Assessing the 21st century skills seems to be doable for large-scale purposes if computer-based technology is used.

References

ACT (2008). WorkKeys. Preparing for the WorkKeys® Assessments. Iowa City, IA: ACT. Pearson Education, Inc.

Allalouf, A. (2003). Revising tanslated dfferential iem fnctioning iems as a tool for improving cross-lingual assessment. Applied Measurement in Education, *16*(1), 55-73.

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, *58*(4), 375-404

Alexander, P. A., Schallert, D. L. & Hare, V. C., (1991). Coming to terms: How researchers in learning and literacy talk about knowledge. *Review of Educational Research*, *61*(3), 315-343.

Ackerman, P. L. (1988). Determinants of individual differences during skills acquisitions: Cognitive abilities and information processing. *Journal of Experimental Psychology: General*, *117*, 288-318.

American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Association of School Librarians (AASL), Retrieved January 9, 2009, from *http://www.ala.org/ala/mgrps/divs/aasl/index.cfm*

American Association of School Librarians (AASL) Standards for the 21st-Century Learner. Retrieved January 9, 2009, from http://www.ala.org/ala/mgrps/divs/aasl/aaslproftools/learningstandards/AASL_LearningStandards.pdf .

Anderson, L. W., Krathwhol, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives*. New York, NY: Lonngman.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University.

Ayala, C. C., Ayala, M. A., & Shavelson, R. J. (2000, April). On the cognitive interpretation of performance assessment scores. Paper presented at the AREA annual meeting, New Orleans, LA.

Baxter, G. P. Elder, A. D.., & Glaser, R. (1996). Knowledge-base cognition and performance assessment in the science classroom. *Educational Psychologist*, *31*(2), 133-140.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. Educational Measurement: Issues and Practice, 7(3), 37-45.

Bazerman, C. (1998). Emerging perspectives on the many dimensions of scientific discourse. In J. R. Martin, & R. Veel (Eds.). *Reading science. Critical and functional perspectives on discourse of science* (pp. 15-33). London, England: Routledge.

Bennett, R. E. (1993). On the meaning of constructed response. In R. E., Bennett & W. Ward. (Eds.). Construct versus choice in cognitive measurement (pp. 1-27). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Bennett, R. E., Persky, H., Weiss, A. R., Jenkins, F. (2007). *Problem-solving in technology-rich environments. A Report from the NAEP Technology-Based Assessment Project*. Research and Development Series. Institute of Education Sciences, NCES 2007-466. Washington, DC: U.D. Department of Education.

Borman, W. C., Kubisiak, U. C., & Schneider, R. J. (1999). Work styles. In N. G., Peterson, M. D., Mumford, W., Borman, P. R., Jeanneret, & E. A., Fleishman, (1999). *An occupational information system for the 21st century: The development of O*NET* (pp. 213-226). Washington, DC: American Psychological Association.

Brandwein, P. F. (1962). *Elements in a strategy for teaching science in the elementary school*. New York, NY: Hartcourt, Brace & World, Inc.

Bransford, J. D., Brown, A. L., & Cocking, L. L. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: The National Academies Press.

Case, E., Stevens, R., & Cooper, M. (2007). Is collaborative grouping an effective instructional strategy? *Journal of College Science Teaching*, *36*(6), 42-47.

Chi, M. T. H. (2006). Two approaches to the study of experts' characteristics. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.). *The Cambridge handbook of expertise and expert performance* (pp. 21-30). Cambridge, NY: Cambridge University Press.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Chi, M. T. H., Glaser, R., & Farr, M. J., (1988). *The nature of expertise*. Hillsdale, NJ: Erlbaum.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86*, 175-218.

Cooper, M. M., Sandi-Urena, S., & Stevens, R. (2008). Reliable multi-method assessment of metacognition use in chemistry problem-solving. *Chemical Education Research and Practice*, *9*, 18-24.

Cooper, M. M., Stevens, R., & Holme, T. (2006). Assessing problem-solving in chemistry using the IMMEX system. Proceedings of the National STEM Assessment Conference 2006, pp.118-129.

Council of Chief State School Officers, National Center for Improving Science Education, & American Institutes for Research (2004). *Science Framework for the 2005 National Assessment of educational Progress*. Washington, DC: National Assessment Governing Board, U.S. Department of Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Darr, A. (2007). *The knowledge worker and the future skill demands of the United States workforce*. Paper prepared for the National Academies' Workshop on Research Evidence Related to Future Skill Demands, Washington, DC, May 31-June 1. Retrieved January 9, 2009, from http://www7.nationalacademies.org/cfe/Future_Skill_Demands_Commissioned_Papers.html.

de Jong, T., & Fergusson-Hessler, M. (1996). Types and qualities of knowledge. *Educational Psychologist, 31*(2), 105-113.

De Kleer, J., & Brown, J. S. (1983). Assumptions and ambiguities in mechanistic mental models. In D. Gentner & A. L. Stevens (Eds.) *Mental models* (pp. 155-190). Hillsdale, NJ: Erlbaum.

Donovan, M. S., & Bransford, J. D. (Eds.) (2005). *How student learn history, mathematics, and science in the classroom*. National Research Council. Washington, D.C.: The National Academies Press.

Driver, R., Asoko, H., Leach, J., Mortimer, E., & Scott, P. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, *23*(7), 5-12.

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17,* 397-434.

Duschl, R. A. (2000). Making the nature of science explicit. In R., Millar, J., Leach, & J. Osborne (Eds.) *Improving science education* (pp. 187-206). Buckingham, England: Open University Press

Duschl, R. A. (2003). Assessment of inquiry. In J.M. Atkin & J.E. Coffey (Eds.), *Everyday assessment in the science classroom* (pp. 41–59). Washington, DC: National Science Teachers Association Press.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8.* Washington, D.C.: The National Academies Press.

Garmire, E., & Pearson, G. (2006). *Tech tally. Approaches to assessing technological literacy*. Washington, D.C.: The National Academies Press.

Gelman, R., & Greeno, J. G. (1989). On the nature of competence: Principles for understanding in a domain. In L. B. Resnick (Ed.). *Knowing, learning, and instruction. Essays in Honor of Robert Glaser* (pp. 125-186). Hillsdale, NY: Lawrence Erlbaum Associate, Publishers.

Giere, R. N. (1988). *Explaining science. A cognitive approach*. Chicago, IL: The University of Chicago Press.

Gentner, D., & Stevens, A. L. (Eds.) (1983). *Mental models*. Hillsdale, NJ: Erlbaum.

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.). *Review of Research in Education* (Volume 20), (pp. 393-419. Washington, DC: American educational research Association.

Haertel, E. H. (1999). Validity argument for high-stake testing: In search of the evidence. Educational Measurement: Issues and Practices, pp. 5-9.

Haertel, E. H., & Linn, R. L. (1996). Comparability. In G. W. Phillips (Ed.) *Technical issues in large-scale performance assessments* (pp. 59-78). Washington, D.C. U.S. Department of Education. Office of educational Research and Improvement.

Hambleton, R.K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hersh, R. H. (nd). Assessing critical thinking, analytical reasoning, problem solving, and writing in high school. Retrieved January 9, 2009, from http://www.cae.org/content/pdf/CWRA%20_4_.pdf .

Hilton., M. (2008). Workshop: Exploring the Intersection of Science Education and the Development of 21st Century Skills. Unpublished document. Division of Behavioral and Social Sciences and Education. Center for Education. Washington, D.C. The National Academies.

Houston, J. (2007). *Future skill demands, from a corporate consultant perspective.* Presentation at the National Academies Workshop on Research Evidence Related to Future Skill Demands. Retrieved January 9, 2009, from http://www7.nationalacademies.org/cfe/Future_Skill_Demands_Presentations.html.

International Society for Technology Education. Retrieved January 9, 2009, from http://www.iste.org/AM/Template.cfm?Section=NETS

International Society for Technology Education. *National Educational Technology Standards*. Retrieved January 9, 2009, from http://www.iste.org/Content/NavigationMenu/NETS/ForStudents/2007Standards/NETS_for_Students_2007_Standards.pdf

Jeanneret, P. R., Borman, W. C., Kubisiak, U. C., & Hanson, M. A. (1999). Generalized work activities. In N. G., Peterson, M. D., Mumford, W., Borman, P. R., Jeanneret, & E. A., Fleishman, (1999). *An occupational information system for the 21st century: The development of O*NET* (pp. 105-125). Washington, DC: American Psychological Association.

Klein, S. P., Kuh, G., D., Chun, M., Hamilton, L., & Shavelson, R. J. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, *46*(3), 251-276.

Klein, S. P. Freedman, D., Shavelson, R. J., Bolus, R. (in press). Assessing school effectiveness. *Evaluation Review*.

Klein, S. P., Benjamin, R., Shavelson, R. J., & Bolus, R. (2007). The Collegiate Learning Assessment: Facts and fantasies. *Evaluation Review*, *31*, 415-439.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*(4), 674-689.

Kuhn, D., Black, J., Keselman, A., & Kaplan, D. (2000). The development of cognitive skills to support inquiry learning. *Cognition and Instruction*, *18*(4), 495-523.

Khishfe, R., & Abd-El-Khalick, F. (2002). Influence of explicit and reflective versus implicit inquiry-oriented instruction of sixth graders' views of nature of science. *Journal of Research in Science Teaching*, *39*(7), 551-578.

Latour, B., & Woolgar, S. (1979). *Laboratory life. The construction of scientific facts*. Princeton, NJ: Princeton University Press.

Lehrer, R., & Romberg, T. (1996). Exploring children's data modeling. *Cognition and Instruction*, *14*(1), 69-108.

Lehrer, R., & Schauble, L. (2006). Scientifi thinking and science literacy. In, W. Damon, R. Lerner, K. A. Renninger, and L. E. Sigel (Eds.), *Handbook of child psychology, 6th edition* (Vol. 4, pp. 153-196). Hoboken, NJ: Wiley.

Levy, F. & Murnane, R. J. (2004). *The new division of labor. How computers are creating the next job market*. New York, NY: Russell sage Foundation.

Li, M. (2001). *A framework for science achievement and its link to test items*. Unpublished dissertation. Stanford, CA: Stanford University.

Li, M., Ruiz-Primo, M. A. & Shavelson, R. J., (2006). Towards a science achievement framework: The case of TIMSS 1999. In S. Howie & T. Plomp (Eds.). *Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp. 291-311). London: Routledge.

Li, M., & Shavelson, R. J., (2001, April). *Using TIMSS items to examine the links between science achievement and assessment methods*. Paper presented at the American Educational Research Association annual meeting. Seattle, Washington

Li, M., & Tsai, S. (2007). Linking assessment to science achievement. NSF Technical report.

Martin, J. R., (1998). Discourse of science: Recontextualization, genesis, intertextuality and hegemony. In J. R. Martin, & R. Veel (Eds.). *Reading science. Critical and functional perspectives on discourse of science* (pp. 3-14). London, England: Routledge.

Martinez, M. E. (1999). Cognition and the questions of test item format. *Educational Psychology, 34*, 207-218.

Mayer, R. (1997). Incorporating problem-solving into secondary school curricula. In G. D. Phye (Ed.). *Handbook of academic learning. Construction of Knowledge* (pp. 473-492). San Diego, CA: Academic Press.

Mayer, R. E., & Wittrock, M. C. (2006). Problem solving. In P.A. Alexander, & P. H. Winne, P. H. (Eds.). *Handbook of educational psychology* (pp. 287-303). 2nd edition. Mahwah, NJ: Lawrence Erlbaum Associates, Pubishers.

Messick, S. (1984). Abilities and knowledge in educational achievement testing: The assessment of dynamic cognitive structures. In B. S. Plake (Ed.) *Social and technical issues in testing: Implication for test construction and usage* (pp. 155-172). Hillsdale, NJ: Lawrence Erlabaum.

Michaels, S., Shouse, A, W., Schweingruber, H. A. (2008). *Ready, Set, Science: Putting Research to Work in K-8 Science Classrooms.* Washington, D.C.: The National Academies Press.

Mislevy, R. J., & Haertel, G (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practices*, *25*(4), 6-20.

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., Johson, L. (1998). A cognitive task analysis, with implications for designing a simulation-based performance assessment. CSE Technical Report 487. Los Angeles, CA: UCLA/National Center for Research on Evaluation, Standards, and Student testing (CRESST).

Mullis, I. V. S., Martin, M. O., Ruddock, G J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 Assessment frameworks*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education.

Mumford, M. D., Peterson, N. G., & Childs, R. A. (1999). Basic and cross-functional skills. In N., Peterson, M. D., Mumford, W., Borman, P. R., Jeanneret, & E. A., Fleishman, (1999). *An occupational information system for the 21st century: The development of O*NET* (pp. 49-69). Washington, DC: American Psychological Association.

National Research Council (2000). *How People Learn: Brain, Mind, Experience, and School.* Washington, D.C.: National Academies Press.

National Research Council. (2008a). *Research on Future Skill Demands: A Workshop Summary.* Washington, D.C.: National Academies Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall, INC.

North Central Regional Educational laboratory and the Metiri Group (2003). *EnGauge 21st Century Skills for 21st Century Learners. Literacy in the digital age.* Naperville, IL: NCREL and the Metiri Group.

Organisation for Economic Co-Operation and Development (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris, France: OECD.

Partnership for the 21st Century Skills. (2009). Retrieved January 9, 2009, from http://www.21stcenturyskills.org/.

Pellegrino, J. W., (2002, February). *Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction and assessment*. Paper presented at the NSF instructional Materials and Development PI Meeting. Washington, DC

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

Polya, G. (1945). *How to solve it*. Princeton. NJ: University Press.

Proctor, R. W., & Dutta., A. (1995). *Skills acquisition and human performance*. Thousand Oaks, CA: Sage Publications.

Pulakos, E. D., Arad, S., Donovan, M.A., and Plamondon, K.E. (2000). Adaptability in the workplace: Development of a taxonomy of adaptive performance. *Journal of Applied Psychology,85*, 612-624.

Pulakos, E. D., Schmitt, N., Dorsey, D. W., Arad, S., Hedge, J., W., & Borman, W. C. (2002). Predicting Adaptive Performance: Further Tests of a Model of Adaptability. *Human Performance*, *15*(4), 299 – 323.

Quellmalz, E. (2002). Using cognitive analysis to study the validities of science inquiry assessments. Paper presented at the Annual AERA Meeting, New Orleans, LA

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, 323(5910), 75-79.

Royer, J. M., Ciscero, C. A., & Carlo, M. S. (1993). Techniques and procedures for assessing cognitive skills. *Review of Educational Research*, *63*, 201-243.

Ruiz-Primo, M. A. (1997). *Toward a framework of subject-matter achievement assessment*. Unpublished Manuscript. Stanford, CA: Stanford University.

Ruiz-Primo, M. A. (1998). *Models for measuring science achievement*. Invited talk. National CRESST Conference. University of California, Los Angeles.

Ruiz-Primo, M. A. (2003, April). *A framework to examine cognitive validity*. Paper presented at the meeting of the American Education Research Association, Chicago, IL.

Ruiz-Primo, M.A. (2007). Assessment in science and mathematics: Lessons learned. In M., Hoepfl & M. Lindstrom (Eds.), *Assessment of Technology Education, CTTE 56th Yearbook* (pp. 203-232). Woodland Hills, CA: Glencoe-McGraw Hill.

Ruiz-Primo, M. A. & Shavelson, R. J. (1996). Rhetoric and reality in science performance assessments: An update. *Journal of Research in Science Teaching*, *33*(10), 1045-1063.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative mapping techniques. *Educational Assessment*, *7*(2), 99-141.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L. & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*(5), 369-393.

Rutherford, F. J. (1964). The role of inquiry in science teaching. *Journal of Research in Science Teaching*, I, 80-84.

The Secretary's Commission on Achieving necessary skills (1991). What work requires of schools. A SCANS Report for America 2000. Washington, DC: U.S. Department of Labor.

Schwab, J. J. (1962). *The teaching of science as enquiry*. Cambridge, MA: Harvard University Press.

Shavelson, R. J. (2007). *A brief history of student learning assessment. How we got where we are and a proposal for where to go next*. Washington, DC: Association of American Colleges and Universities.

Shavelson, R.J., Baxter, G. P., Gao. X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, *30*, 215-232.

Shavelson, R. J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change*, *35*(1), 10-19.

Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the psychometrics of assessing science understanding. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.) *Assessing science understanding* (pp. 304-341). San Diego: Academic Press. Educational Psychology Series.

Shavelson, R. J., Ruiz-Primo, M. A., Li, M., & Ayala, C. C. (2002). *Evaluating new approaches to assessing learning*. CSE: Technical Report 604. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing/ University of California, Los Angeles.

Snow, R. E. (1993), Construct validity and constructed-response tests. In R. E. Bennett, & W. C. Ward (Eds.). *Construction versus choice in cognitive measurement: Issues in constructed response performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching, 38(*5*)*, 553-573.

Solano-Flores, G. (2009). The testing of English language learners as a stochastic process: Population misspecification, measurement error, and overgeneralization. To appear in K. Ercikan & W. M. Roth (Eds.), *Generalizing from Educational Research.*

Stevens, R., Johnson, D. F., & Soller, A. (2005). Probabilities and predictions: Modeling the development of scientifically problem-solving skills. *Cell Biology Education*, *4*, 42-57.

VandenBos, G. R. (2007) (Ed.). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

WestEd, & Council of Chief State School Officers (2005). *Science Framework for the 2009 National Assessment of educational Progress*. Washington, DC: National Assessment Governing Board, U.S. Department of Education.

Wiliam, D. (2007, September). *Sensitivity to instruction: The missing ingredient in large-scale assessment systems?* Paper presented at the Annual Meeting of the International Association for Educational Assessment. Baku, Azerbaijan.

Wilson, M., & Bertenthal, M. W. (2006). *Systems for state science assessments*. Washington, DC: The National Academies Press.

Yin, Y. (2005). *The Influence of Formative Assessments on Student Motivation, Achievement, and Conceptual Change.* Unpublished dissertation. Stanford, CA: Stanford University,

Appendix A

*Comparing and Contrasting Frameworks of 21st century skills**

THE NRC 21ST CENTURY SKILLS					
Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	**Not Aligned** (Not Considered Equivalent or Difficult to Align)
PARTNERSHIP FOR THE 21ST CENTURY SKILLS					
I. Creativity and Innovation - Being open and responsive to new and diverse perspectives **I. Communication & collaboration** - Demonstrating ability to work effectively with diverse teams - Exercising flexibility and willingness to be helpful in making necessary compromises to accomplish a common goal **III. Flexibility & adaptability** - Adapting to varied roles and responsibilities - Working effectively in a climate of ambiguity and changing priorities **III. Social & cross-cultural skills** - Working appropriately and productively with others - Leveraging the collective intelligence of groups when appropriate - Bridging cultural differences and using differing perspectives to increase innovation and the quality of work	**I. Creativity and Innovation** - Developing, implementing and communicating new ideas to others **I. Critical thinking and problem solving** - Identifying and asking significant questions that clarify various points of view **I. Communication & collaboration** - Articulating thoughts and ideas clearly and effectively through speaking and writing - Assuming shared responsibility for collaborative work **II. Media Literacy** - Understanding how media messages are constructed, for what purposes and using which tools, characteristics and conventions - Examining how individuals interpret messages differently, how values and points of view are included or excluded and how media can influence beliefs and behaviors **II. Information, communication, and technology** - Using digital technology, communication tools and/or networks appropriately	**I. Creativity and Innovation** - Demonstrating originality and inventiveness in work - Acting on creative ideas to make a tangible and useful contribution to the domain in which the innovation occurs **I. Critical thinking and problem solving** - Exercising sound reasoning in understanding - Framing, analyzing and synthesizing information in order to solve problems - Making complex choices and decisions **II. Information Literacy** - Accessing information efficiently and effectively, evaluating information critically and competently and using information accurately and creatively for the issue or problem at hand	**III. Initiative & self direction** - Monitoring one's own understanding and learning needs - Going beyond basic mastery of skills and/or curriculum to explore and expand one's own learning and opportunities to gain expertise - Demonstrating initiative to advance skill levels towards a professional level - Defining, prioritizing and completing tasks without direct oversight - Utilizing time efficiently and managing workload - Demonstrating commitment to learning as a lifelong process	**I. Critical thinking and problem solving** - Understanding the interconnections among systems **III. Leadership & responsibility** - Acting responsibly with the interests of the larger community in mind	**I. Communication & collaboration** - Assuming shared responsibility for collaborative work **II. Information Literacy** - Possessing a fundamental understanding of the ethical/legal issues surrounding the access and use of information **III. Productivity & accountability** - Setting and meeting high standards and goals for delivering quality work on time - Demonstrating diligence and a positive work ethic **III. Leadership & responsibility** - Using interpersonal and problem-solving skills to influence and guide others toward a goal - Leveraging strengths of others to accomplish a common goal - Demonstrating integrity and ethical behavior

* Roman numbers correspond to the general skills mentioned in the frameworks description.

Appendix A (*Continued*)*

THE NRC 21ST CENTURY SKILLS					
Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	**Not Aligned** (Not Considered Equivalent or Difficult to Align)
AASL STANDARDS FOR THE 21ST-CENTURY LEARNER [b]					
	I. Inquire, think critically, and gain knowledge - Read, view, and listen for information presented in any format in order to make inferences and gather meaning - Collaborate with others to broaden and deepen understanding - Make sense of information gathered from diverse sources - Demonstrate mastery of technology tool for accessing information and pursuing inquiry **II. Draw conclusions, make informed decisions, apply & create knowledge** - Use technology and other information tools to analyze and organize information - Collaborate with others to exchange ideas, develop understanding, make decision, and solve problems - Use the writing process, media and visual literacy, and technology skills to create products that express ideas. **III. Share knowledge and participate ethically & productively as members of a society** - Participate and collaborate as members of social & intellectual network - Share new understanding - Use writing and speaking to communicate new understanding effectively - Use technology and other information tools to organize and display knowledge & understanding	**I. Inquire, think critically, and gain knowledge** - Develop & refine a range of questions to frame research for new understanding. - Find, evaluate, and select appropriate sources to answer questions. - Evaluate information found in selected sources on the basis or accuracy, validity, & appropriateness **II. Draw conclusions, make informed decisions, apply & create knowledge** - Continue an inquiry-based research process by applying critical thinking to construct new understanding - Organize knowledge so that it is useful - Use strategies to draw conclusions from information and apply knowledge to curricular areas, real world situations, and further investigations. **III. Share knowledge and participate ethically & productively as members of a society** - Conclude an inquiry-based research process by sharing new understanding & reflecting on the knowledge **IV. Pursue personal and aesthetic growth** - Organize personal knowledge in a way that can be called upon easily	**I. Inquire, think critically, and gain knowledge** - Use prior and background knowledge as context for new learning **IV. Pursue personal and aesthetic growth** - Read, view and listen for personal pleasure and growth - Read widely and fluently to make connections with self, the world, and previous reading. - Seek information for personal learning in a variety of formats and genres. - Connect ideas to own interests and previous knowledge and experiences	**IV. Pursue personal and aesthetic growth** - Use social networks and information tools to gather and share information	**I. Inquire, think critically, and gain knowledge** - Follow an inquiry based process in seeking knowledge in curricular subjects, and make the real world connection for using this process in own life **III. Share knowledge and participate ethically & productively as members of a society** - Connect learning to community issues - Use information and technology ethically and responsibly **IV. Pursue personal and aesthetic growth** - Respond to literature and creative expressions in various forms and genres - Use creative and artistic formats to express personal learning

* Roman numbers correspond to the general skills mentioned in the frameworks description.

[b] The comparison focuses on one of the three aspects proposed by AASL, *skills*. Dispositions in action, responsibilities, and self assessment strategies were not considered in the comparison.

Appendix A (*Continued*)*

THE NRC 21ST CENTURY SKILLS					
Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management/ Self-Development**	**Systems Thinking**	**Not Aligned** (Not Considered Equivalent or Difficult to Align)
ENGAUGE 21ST CENTURY SKILLS FOR 21ST CENTURY LEARNERS [c]					
II. Inventive thinking - *Adaptability and managing complexity* – Modify one's thinking, attitude, or behavior, to be better suited to current or future environments; handle multiple goals, tasks, and inputs, while understanding and adhering to constraints of time, resources, and systems III. **Effective communication** - *Interpersonal skills* – Ability to read and manage the emotions, motivations, and behaviors of one-self and others during social interactions or in a social interactive context	**III. Effective communication** - *Teaming and collaboration* – Cooperative interaction between two or more individuals working together with a purpose **III. Effective communication** - *Interactive communication* – Generation of meaning through exchanges using a range of contemporary tools, transmissions, and processes. **IV. High Productivity** - *Effective use of real-world tools* – Ability to use real-world tools (e.g., software, hardware, or peripheral devices used by information technology) to communicate, collaborate, solve problems, and accomplish tasks	**II. Inventive thinking** - *Creativity* – Bringing something into existence that is genuinely new and original - *Risk Taking* – Willingness to tackle challenging problems without obvious solutions - *Higher order thinking and sound reasoning* – Cognitive processes of analysis, comparison, inference and interpretation, evaluation, and synthesis applied to a range of academic domains and problem-solving contexts.	**II. Inventive thinking** - *Self direction* – Ability to set goals related to learning, plan for the achievement of those goals, manage of time and effort, independently assessment of quality of learning and products - *Curiosity* – Desire to know or the speak of interest that leads to inquiry - *Risk Taking* – Willingness to make mistakes, advocate unconventional or popular positions, or tackle personal growth or accomplishments **IV. High Productivity** - Prioritizing, planning, and managing results – Ability to organize to achieve goals of specific projects or problems	**III. Effective communication** - *Social and civic responsibility* – Ability to manage technology and governs its use in a way that promotes public good and protects society; the environment, and democratic ideas.	**I. Digital-Age-Literacy** - All the different types of literacy mentioned. **III. Effective communication** - *Personal responsibility* – Deep knowledge about legal and ethical issues **IV. High Productivity** - Ability to produce relevant, high-quality products – Ability to produce intellectual, informational, or material products that serve authentic purposes

Appendix B

*Comparing and Contrasting Frameworks of 21st century skills with Some Technological Standards**

THE NRC 21ST CENTURY SKILLS					
Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management/ Self-Development**	**Systems Thinking**	**Not Aligned** (Not Considered Equivalent or Difficult to Align)
ISTE'S NATIONAL EDUCATIONAL TECHNOLOGY STANDARDS					
	II. Communication and collaboration - Use of digital media and environments to communicate and work collaboratively, including at a distance to support individual learning and contribute to the learning of others	**I. Creativity and Innovation** - Demonstrate creative thinking, construct knowledge, and develop innovative products and processes using technology **III. Research and information fluency** - Apply digital tools to gather, evaluate, and use information to, for example, plan strategies to guide inquiry **IV. Critical thinking, problem solving, and decision making** - Use critical thinking skills to plan and conduct research, mange projects, solve problems, and make informed decisions using appropriate digital tools and resources			**V. Digital citizenship** - Understand human, cultural, and societal issues related to technology and practice legal and ethical behavior **VI.** Technology operations and concepts - Demonstrate a sound understanding of technology concepts, systems, and operations
TECHNOLOGICAL LITERACY DIMENSIONS [c]					
		II Asks pertinent questions, of self and others, regarding the benefits and risks of technologies **II** Weights available information about benefits, risks, costs, and trade-offs of technology **III** Can use design-thinking process to solve a problem **III** Can obtain information about technological issues of concern from variety of sources			**I** Knowledge **II** Participates, when appropriate, in decisions about the development and uses of technology **III** Has a range of hands-on skills, such as operating a variety of home and office appliances and using computer for word-processing and surfing the internet.

* Roman numbers correspond to the general skills mentioned in the frameworks description.

[c] For the purposes on the paper, the comparison considers only the skills but not the students profiles (for student profiles see North Central Regional Educational laboratory & the Metiri Group, 2003).

Appendix C

*Mapping Large-Scale Assessment Framework, Science Education Goals, and NRC's 21st century skills **

Strands of Scientific Proficiencies	THE NRC 21ST CENTURY SKILLS					Not Aligned (Not Considered Equivalent or Difficult to Align)
	Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	
NAEP 2005						
I. Knowledge, use and interpretation of scientific explanations			**I. Conceptual Understanding** - Understanding basic concepts and tools in the process of scientific investigation			
II. Generate and evaluate evidence and explanations			**II. Scientific Investigation** - Using the appropriate tools and thinking processes in the application of science **III. Practical reasoning** - Engaging in practical reasoning by suggesting effective solutions to everyday problems by applying scientific knowledge and using skills			
III. Understand the nature and development of scientific knowledge						
IV. Participate productively in scientific practices and discourse						

Appendix C *(Continued)* *

Strands of Scientific Proficiencies	THE NRC 21ST CENTURY SKILLS					Not Aligned (Not Considered Equivalent or Difficult to Align)
	Adaptability	**Complex Communication/ Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	
NAEP 2009						
I. Knowledge, use and interpretation of scientific explanations			**I. Identifying science principles** - Stating or recognizing principles - Demonstrating relationships among closely related science principles specified in the broad science areas **II. Using science practices** - Explain observation of phenomena using science principles **IV. Employing technological design** - Identifying scientific tradeoffs in design decisions and choose among alternative solutions - Apply science principles or data to anticipate effects of technological design decisions			
II. Generate and evaluate evidence and explanations		**I. Identifying science principles** - Demonstrating relationships through different forms of representations (verbal, symbolic diagrammatic) and data patterns (e.g., graphs, tables, formulas, and diagrams)	**I. Identifying science principles** - Describing, measuring, or classifying observations **II. Using science practices** - Predict observations of phenomena - Propose, analyze, and evaluate alternative explanations or predictions - Suggest example of observations that illustrate a science principle **III. Conducting scientific inquiry** - Conducting investigation using appropriate tool and techniques - Identifying patterns in data and/or relate patterns in data to theoretical models			
III. Understand the nature and development of scientific knowledge			**III. Conducting scientific inquiry** - Designing and critiquing aspects of scientific investigations - Using empirical evidence to validate or criticize conclusions about explanations and predictions			
IV. Participate productively in scientific practices and discourse			**IV. Employing technological design** - Propose and critique solutions to problems, given criteria and scientific constraints			

* Roman numbers correspond to each strands of scientific proficiency.

Appendix C *(Continued)* *

Strands of Scientific Proficiencies	THE NRC 21ST CENTURY SKILLS					Not Aligned (Not Considered Equivalent or Difficult to Align)
	Adaptability	**Complex Communication Social Skills**	**Non-Routine Problem Solving**	**Self-Management Self-Development**	**Systems Thinking**	
TIMSS 2007						
I. Knowledge, use and interpretation of scientific explanations			**I. Knowing** - Describe – Describe organism, physical materials and science process that demonstrate knowledge of properties, structures, function, & relationships - Use tools & procedures – Demonstrate knowledge of the use of science apparatus, equipment, tools, procedures, measurement devices, and scales. **II. Applying** - Relate – Relate knowledge of a concept to an observed or inferred property, behavior, or use of objects - Use models – Use a diagram or model to demonstrate understanding of a science concept, structure **III. Reasoning** - Integrate/synthesize – Make associations or connections between concepts in different areas of science - Integrate/synthesize – Demonstrate understanding of unified concepts and themes across science domains			**I. Knowing** - Recall/Recognize - Identifying accurate statements about science facts, relationships, processes, and concepts; identify the characteristics or properties of specific organisms - Define – Providing or identifying definitions of scientific terms; recognize and use scientific vocabulary, symbols, abbreviations, units, and scales **II. Applying** - Compare/contrast/ classify – Identify and describe similarities
II. Generate and evaluate evidence and explanations			**II. Applying** - Interpret information – Interpret relevant textual, tabular, or graphical information in light of science concepts or principles - Find solutions – Identify or use a science relationship, equation, or formula to find a solution/demonstration of a concept - Provide or identify an explanation for an observation or natural phenomenon demonstrating understanding of the underlying science concept, principle, law, or theory **III. Reasoning** - Analyze/Solve problems – Determine the relevant relations, concepts, and problem solving steps; develop and explain problem solving strategies - Integrate/synthesize – Provide solution that requires consideration of different factors or related concepts. Integrate mathematical concepts or procedures in the solutions to science problems - Hypothesize/Predict – Combine knowledge of science concepts, experience, and observation to formulate questions that can be answered through investigations. Formulate hypotheses as testable assumptions; make predictions about effects of changes. - Design/Plan – Design or plan investigations for answering scientific problems. Recognize characteristics of good designs. Make decisions about measurements. - Draw conclusions – Detect patterns of data, describe or summarize data trends, interpret data (interpolate/extrapolate from data), make valid inferences, draw appropriate conclusions - Generalize – Make conclusions that go beyond the given conditions, apply conclusions to new situations - Evaluate – Weigh advantages and disadvantages to make decisions about diverse alternatives - Justify – Use evidence to justify explanations and problem solutions **IV. Scientific Inquiry** - Formulating question and hypothesis, design experiments, represent data, analyze and interpret data, and drawing conclusions and developing explanations			
III. Understand the nature and development of scientific knowledge						
IV. Participate productively in scientific practices and discourse						

Appendix C *(Continued)* *

Strands of Scientific Proficiencies	THE NRC 21ST CENTURY SKILLS Adaptability	Complex Communication/ Social Skills	Non-Routine Problem Solving	Self-Management Self-Development	Systems Thinking	Not Aligned (Not Considered Equivalent or Difficult to Align)
PISA 2006						
I. Knowledge, use and interpretation of scientific explanations			**I. Identifying scientific issues** - Recognizing issues that it is possible to investigate scientifically **II. Explaining phenomena scientifically** - Applying knowledge of science in a given situation - Identifying appropriate descriptions, explanations, and predictions **IV. Attitudes - Support for scientific inquiry** - Supporting the use of factual knowledge and rational explanations	**I. Identifying scientific issues** - Identifying keywords to search for scientific information		
II. Generate and evaluate evidence and explanations			**II. Explaining phenomena scientifically** - Describing or interpreting phenomena scientifically and predicting changes **III. Using scientific evidence** - Interpreting scientific evidence and making and communicating conclusions - Identifying the assumptions, evidence and reasoning behind conclusions			
III. Understand the nature and development of scientific knowledge			**I. Identifying scientific issues** - Recognizing the key features of scientific investigation **IV. Attitudes - Support for scientific inquiry** - Expressing the need for logical and careful processes in drawing conclusions			
IV. Participate productively in scientific practices and discourse			**III. Using scientific evidence** - Reflecting on the social implications of science and technological development **IV. Attitudes - Support for scientific inquiry** - Acknowledging the importance of considering different perspectives and arguments			

Appendix C *(Continued)* *

Strands of Scientific Proficiencies	THE NRC 21ST CENTURY SKILLS: Adaptability	Complex Communication/ Social Skills	Non-Routine Problem Solving	Self-Management/ Self-Development	Systems Thinking	Not Aligned (Not Considered Equivalent or Difficult to Align)
CLA-CCLA-CWRA						
I. Knowledge, use and interpretation of scientific explanations			**I. Evaluation of evidence** - Determining what information is or is not pertinent to the task at hand **I. Analysis and synthesis of evidence** - Drawing connections between discrete sources of data and information			**II. Mechanics** - Using vocabulary and punctuation correctly, effectively - Demonstrating a strong understanding of grammar - Using proper transitions - Structuring paragraphs logically and effectively **II. Interest** - Structuring syntax and organization to add interest to the writing - Using colorful but relevant metaphors, similes - Craft writing that engages the reader - Using writing that leaves the reader thinking
II. Generate and evaluate evidence and explanations		**II. Presentation** - Articulating clearly the argument and the context for the argument - Using evidence correctly and precisely to defend the argument - Present evidence comprehensively and coherently - Citing sources correctly and consistently **II. Development** - Organizing the argument logically and coherently - Avoiding extraneous elements in the argument - Presenting evidence in an order that contributes to a persuasive and coherent argument **III. Persuasiveness** - Presenting supportive evidence effectively - Drawing thoroughly and extensively from the available range of evidence - Analyzing evidence rather than only presenting it - Considering counter-arguments and addressing weakness in his/her argument **II. Interest** - Using creative and engaging examples	**I. Analysis and synthesis of evidence** - Presenting his/her own analysis of the data or information - Committing or failing to recognize logical flaws - Breaking down the evidence into its component parts - Attending to contradictory, inadequate or ambiguous information **I. Drawing conclusions** - Constructing convincing arguments based on data - Selecting the strongest set of supporting data - Prioritizing components of the argument - Avoiding overstated or understated conclusions - Identifying holes in the evidence and subsequently suggesting additional information that might resolve the issue			
III. Understand the nature and development of scientific knowledge			**I. Evaluation of evidence** - Distinguishing between rationale claims and emotional ones, fact from opinion - Recognizing the ways in which the evidence might be limited or compromised - Spotting deception and holes in the arguments of others - Considering all sources of evidence **I. Acknowledging alternative explanation/viewpoints** - Recognizing that complex problem do not have a clear answer - Proposing other options and weighting them in the decision - Qualifying responses and acknowledging the need for additional information in making absolute decisions		**I. Acknowledging alternative explanation/viewpoints** - Considering all stakeholders or affected parties in suggesting a course of action	
IV. Participate productively in scientific practices and discourse						

Appendix D

Assessment Task Dimensions for Retrospective Logical Analysis

Task Dimensions	Examples for Designing Tasks or for Coding Decisions			
Task Demands: What the task asks the test taker to do, what it elicits from the student	- Defining concepts - Identifying facts	- Executing procedures in familiar tasks - Executing procedures in unfamiliar tasks	- Selecting an appropriate procedure to solve a problem - Determining the theoretical position of a manuscript - Drawing diagrams illustrating a process - Writing an equation from a statement - Constructing an interpretation - Drawing conclusions - Justifying or predicting	- Evaluating the validity of a conclusion - Evaluating products, or proposals - Producing alternative solutions to a given problem - Designing an experiment to solve a non-routine or novel problem
***Cognitive Demands*:** Inferred cognitive processes that students likely act upon to provide responses	Less Cognitive Demanding ⟵			⟶ More Cognitive Demanding
	Remembering Recognizing Recalling	*Applying* Executing or Implementing more or less routine procedures	*Reasoning Using Mental Models* Explaining Interpreting Inferring Organizing/Classifying Comparing/Contrasting Exemplifying	*Assembling Knowledge in New/Creative Ways* Planning Generating Producing Monitoring
Item Openness Gradation in the constraint exerted in the nature and extent of the response	- The continuum from multiple-choice tests to constructed response test format such as long essays, projects, and collection of products over time - Require one correct solution versus multiple correct solutions/approaches - The continuum of the structureness/directedness of the task (following instructions or steps)			

Appendix E

Definitions of the Nature of Problem Dimensions

Problem Solving Dimensions

Structureness.[12] Problems can be classified as well-structured versus ill-structured. *Well-structured problems* are those in which the student has a clear sense of what an acceptable answer would be and a clear idea of what it takes to get to that answer. *Ill-structured problems* do not have one definitive correct answer, therefore the criteria to accept an answer as correct is not straightforward and there is no correct algorithm to solve it. It is important to note that structureness, or level of problem definition, does not depend on the student's (problem-solver) knowledge, but on the problem characteristics. However, students (problem-solvers) can have different knowledge about the characteristics of the problem.

Routiness. A *routine problem* is one for which problem-solvers already have a solution procedure, a *non-routine problem* is the one for which there is no previously learned procedure. The definition of routine or nonroutine problem depends on the knowledge of the student. Real-world problem are in general nonroutine (Mayer & Wittrock, 2006).

Richness. There is no clear definition of rich problems, but here I consider rich problems those that require problem solver to: (a) conduct diverse activities to solve it (e.g., locate and synthesize information, conduct an experiment, read oriented scientific materials), (b) have more than one subtask (e.g., students need to conduct, say several small experiments to solve the problem), and (c) provide real-world context. Lean problems focus on one task that does not require multiple activities but may or may not provide a real-world context.

Schoolness. *Schooled problems* deal with familiar or taught procedures that required academic exposure; a textbook type of problem. Problems that do not require academic exposure to solve them have been named unschooled (Alexander, 2006). However, it should not be expected that students have been taught explicitly how to solve *unschooled problems*; but it should be expected that the information gained through formal and information education helps them to solve it.

Collaboration. Problems which solution is approached without the input or assistance of another person are named *independent problem solving*. Problems which solution is approached in groups (students work together) are named *collaborative problem-solving*.

Time. When there is a time constraint to solve the problem it is said that the problem is *timed*. Problems and tasks that are unrestricted in terms of time are considered *untimed*.

Communication Dimensions

Extent of Writing. This dimension focuses on the relative intensity or amount of writing that is required (e.g., an explanation that may require three lines versus and explanation that requires a rationale around diverse conclusions with persuasiveness intend). It is assumed that the larger the extent of the response required, the higher the complexity of the communication due to the different aspects that need to be considered in scoring the quality of the response (e.g., presentation, development, persuasiveness; see for example the CLA Scoring Rubric).

---

[12] This type of problem has been named also well-defined or ill-defined (e.g., Mayer & Wittrock, 2006).