Ruiz-Primo, Maria-Araceli & Li, Min (2016). PISA science contextualized items: the link between the cognitive demands and context characteristics of the items. *RELIEVE*, *22*(1), art. M11. DOI: http://dx.doi.org/10.7203/relieve.22.1.8280

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

PISA science contextualized items: the link between the cognitive demands and context characteristics of the items

Items contextualizados de ciencias en PISA: la conexión entre las demandas cognitivas y las características de contexto de los items

Ruiz-Primo, Maria-Araceli [(1)] **& Li, Min** [(2)]

(1) Stanford University. (2) University of Colorado Boulder.

Abstract

The ubiquitous use of contexts in test items is based on the premise that contextualizing items is an effective strategy to test whether students can apply or transfer their knowledge. In this paper, we continue a research agenda focusing on testing this premise. We present a study of the context characteristics in a sample of 2006 and 2009 PISA science items and how these characteristics as well as student performance may be related to the cognitive demands of the items. The study was guided by two research questions: (1) What are the cognitive demands of the sampled PISA contextualized items and what is the students' performance linked to these items? (2) Are the items' cognitive demands associated with certain characteristics of the contexts of the items that proved to be linked to students' performance? Using 52 released and secured PISA science items, we captured information about three context dimensions of items (i.e., level of abstraction, resources, and nature of the context) and the cognitive demands of the items. A multinomial logistic regression with cognitive demand as the outcome variable, context characteristics as the predictors, and percent of correct responses as the covariant indicated that certain context characteristics are linked to the cognitive demands of items. For example, we found that items in which contexts involve only concrete ideas were associated with items with low cognitive demands; these items are unlikely to require content knowledge to be responded. We also found that the type of resource (e.g., tables, graphs) was associated with the cognitive demands of the items: schematic representations seem to be linked to items tapping procedural knowledge rather than to items tapping declarative or schematic knowledge. We concluded that further research is needed to better understand the influence that context characteristics have on the cognitive processes in which students are asked to engage and in their performance.

Keywords: PISA; science items; context characteristics of the items; cognitive demands; validity

Reception Date
2016 May 15

Approval Date
2016 June 25

Publication Date:
2016 June 25

Resumen

El uso frecuente de contextos en items de una prueba se basa en la premisa de que agregar contextos a los items es una estrategia eficaz para comprobar si los estudiantes pueden aplicar o transferir sus conocimientos. En este trabajo, seguimos una línea de investigación que se centra en probar esta premisa. Se estudian las características de los contextos en una muestra de items de ciencias de PISA 2006 y 2009 y cómo estas características, así como el desempeño de los estudiantes, pueden estar relacionados con las demandas cognitivas de los items . El estudio se guío por dos preguntas de investigación: (1) ¿Cuáles son las demandas cognitivas de una muestra de items con contexto de la prueba PISA y cuál es el desempeño de los estudiantes en estos items? (2) ¿Están asociadas las demandas cognitivas de los items con ciertas características de los contextos, que previamente han demostrado estar relacionadas con el desempeño de los estudiantes? Se codificaron 52

Fecha de recepción
15 Mayo 2016

Fecha de aprobación
25 Junio 2016

Fecha de publicación
25 Junio 2016

Autor de contacto / Corresponding author

Ruiz-Primo, Maria-Araceli. Graduate School of Education. 485 Lasuen Mall. Stanford University, Stanford, CA (USA). aruiz@stanford.edu

items liberados y no liberados de PISA en tres dimensiones de los contextos de los items (nivel de abstracción, recursos y naturaleza del contexto) además de las demandas cognitivas de los items. Una regresión logística multinomial, con la demanda cognitiva como variable de resultado, las características contextuales como los predictores, y el porcentaje de respuestas correctas como la covariante, indicó que ciertas características del contexto están vinculados a las demandas cognitivas de los items . Por ejemplo, se encontró que los contextos en los que sólo se presentan ideas concretas están asociadas a items con bajas demandas cognitivas; en estos items es poco probable que se requiera de un conocimiento del contenido para ser respondidos. También se encontró que el tipo de recurso (por ejemplo, tablas, gráficos) se asocia con las demandas cognitivas de los items las representaciones esquemáticas parecen estar vinculadas a items de tipo procedural y no a items de tipo declarativo o esquemático. Se concluye que se necesita más investigación para comprender mejor la influencia que tienen las características de contexto en los procesos cognitivos de los estudiantes y en su desempeño

Palabras clave: PISA; items de ciencias; características contextuales de los items; demandas cognitivas; validez

The argument that all teaching, learning, and testing should be done in a meaningful context (see for example Boaler, 1993, 1994; Greeno, 1989; Hembree, 1992; Taber, 2003; Wiggins, 1993) leads to the creation of items that include a context that is recognizable by students as something realistic, interesting, relevant, and useful (Haladyna, 1997). It is widely believed that using meaningful contexts during instruction (context-based learning) enhances motivation and confidence (Boaler, 1993, 1994; Kelly, 2007). If such contexts do enhance learning, it follows that, at least hypothetically, contexts can do the same in test items. For years, contextualizing items has been considered an effective strategy to measure complex thinking (Haladyna, 1994) and to test whether students can apply their knowledge (Ahmed & Pollitt, 2000, 2007; Boaler, 1994; Haladyna, 1997).

Because of these beliefs, items with contexts have become widely used in science testing. In the United States, 70% of the science items released by the National Assessment of Educational Progress (NAEP) for Grade 8 and 71% for Grade 4 are items with contexts (Wang & Li, 2014). Similarly, 78% of the 2011 *Trends in International Mathematics and Science Study* (TIMSS) science items released for Grade 8 are also contextualized (Wang & Li, 2014). Indeed, contextualized items are the primary item type in large-scale assessment programs such as the *Programme for International Students Assessment (PISA)* (Ruiz-Primo & Li, 2012, 2015; Wang & Li, 2014).

Despite the widespread use of contexts in items, their utility, and practice, their underlying assumptions have been called into question by some (Ahmed & Pollitt, 2001; Gerofsky, 1996; Wiliam, 1997) due to the lack of knowledge about this type of item (Haladyna, 1994; but see Ahmed & Pollitt, 1999, 2000, 2007; Mevarech & Stern, 1997). For example, do we know the types of contexts used in science items? Do we know how the degree of contextualization may facilitate or impede the performance of students with different abilities as the intended construct and interact with some other important abilities, such as reading proficiency? In other words, do we know what characteristics or types of contexts have the potential to cause an *invalid source of difficulty* (Hembree, 1992; Fisher-Hoch & Hughes, 1996; Mevarech & Stern, 1997)? A review of current research indicates that we know little about what types of context are helpful or detrimental in serving the purpose of accurately assessing the intended construct. As Ahmed and Pollitt (2000) put it, “A good context allows us to measure the student’s ability to apply their knowledge, but a bad context can prevent us from measuring anything at all” (p. 1). If contextualized items contain something that makes them work in a different direction than expected, no amount of good administration, good teaching, or wise judgment can compensate for the wrongness of the item (Ahmed & Pollitt, 2007). If context can affect the way items are interpreted and, therefore, responded to, then we are dealing with a validity issue that should be investigated deliberately. Learning more about the effect of contexts and their characteristics on students' performance seems necessary.

We (Ruiz-Primo & Li, 2012, 2015; Ruiz-Primo, Li, & Minstrell, 2014) have started a research agenda focusing on learning more about contexts in items. This type of research can potentially offer a robust framework for characterizing item contexts and studying their effects on student performance. As part of this research agenda, in this paper we focus on how contexts' characteristics are related to the cognitive demands of PISA items. Unlike most achievement tests (e.g., NAEP), some PISA items have up to three levels of contexts and therefore provide an important source of information with which to explore these ideas. Further, contexts in PISA items contain much more information than their equivalents in other achievement tests.

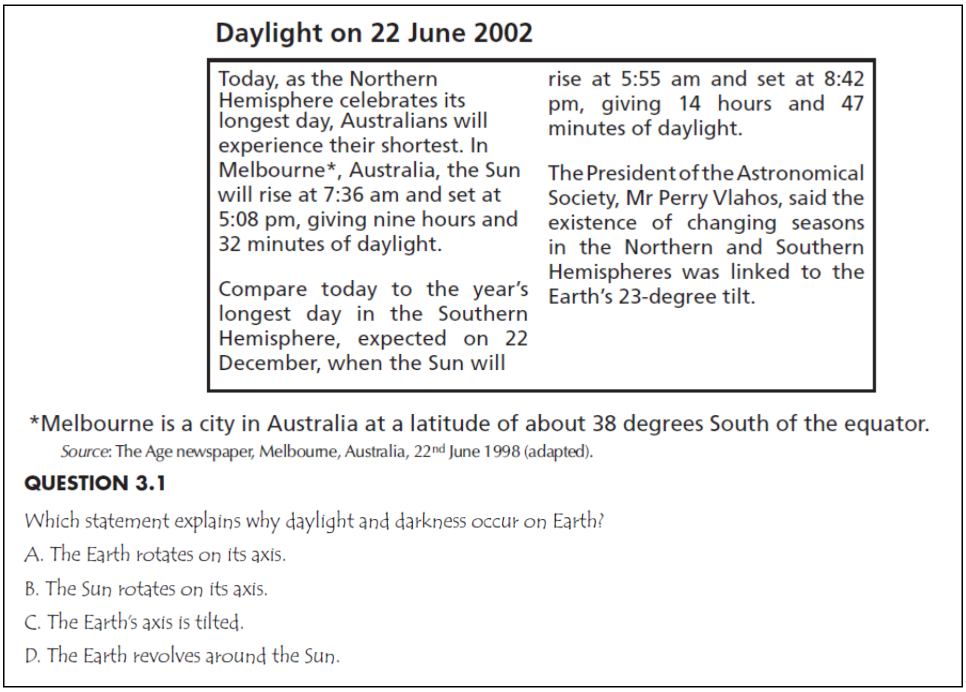
The study presented in this paper focuses on the following research questions: (1) *What are the cognitive demands of the sampled PISA contextualized items and what is the students' performance linked to these items*? (2) *Are the items' cognitive demands associated with certain characteristics of the contexts of the items that proved to be linked to students' performance?* To learn about context characteristics, we use a strategy, the *context profiling approach*, to gather information about items, based on a logical analysis. We provide information about a sample of the 2006 and 2009 PISA science items, including both released and secured items, to which we applied the coding approach. We focus on items with context characteristics that we found were related to students' performance and explore how such characteristics relate to the cognitive demands of the items. We provide exploratory evidence of the possible relationship between context characteristics, cognitive demands, and test takers' performance. The paper contributes to the field in at least two ways: (1) it advances our understanding of contextualized items and their impact on students' performance; and (2) it contributes to our understanding of item development in science specifically, but potentially in other STEM disciplines. In what follows, we first provide information about contextualized items. We then describe the coding approach used to analyze context items, followed by the description of the study.

Contextualized Items

In previous papers, we (Ruiz-Primo & Li, 2012, 2015) defined *item context* as an item component of supplemental information that precedes or follows the item question, such as a description of a lab setup, a natural phenomenon, or a practical problem[1.] Common related terms include scenario, background, vignette, or cover story. Items with contexts have been called *contextualized* (e.g., Ahmed & Pollitt, 2007), *scenario-based* (e.g., Fulkerson, Nichols, Haynie, & Mislevy, 2009; McMartin, McKenna, & Youssefi, 2000), *narrative* (e.g., Terry, 1980), *realistic* (e.g., Cooper & Dunne, 2000; Hembree, 1992), or *content-dependent* (e.g., Haladyna, Downing, & Rodriguez, 2002). It has been argued that contextualizing items makes the questions more concrete and less demanding (Ahmed & Pollitt, 2000, 2007). Figure 1 provides an example of a contextualized item[2].

[1] Contexts can be included in the options as well. In this paper, we only focus on contextualized items that present contextual information in the prompts

[2] Example items in this paper were taken from "PISA take the test: sample questions from OECD's PISA assessment" (http://www.oecd.org/pisa/pisaproducts/pisatakethetestsamplequestionsfromoecdspisaassessments.htm)



Daylight on 22 June 2002

Today, as the Northern Hemisphere celebrates its longest day, Australians will experience their shortest. In Melbourne*, Australia, the Sun will rise at 7:36 am and set at 5:08 pm, giving nine hours and 32 minutes of daylight.

Compare today to the year's longest day in the Southern Hemisphere, expected on 22 December, when the Sun will rise at 5:55 am and set at 8:42 pm, giving 14 hours and 47 minutes of daylight.

The President of the Astronomical Society, Mr Perry Vlahos, said the existence of changing seasons in the Northern and Southern Hemispheres was linked to the Earth's 23-degree tilt.

*Melbourne is a city in Australia at a latitude of about 38 degrees South of the equator.

Source: The Age newspaper, Melbourne, Australia, 22nd June 1998 (adapted).

QUESTION 3.1

Which statement explains why daylight and darkness occur on Earth?

A. The Earth rotates on its axis.

B. The Sun rotates on its axis.

C. The Earth's axis is tilted.

D. The Earth revolves around the Sun.

Figure 1. An example of a contextualized item

Unfortunately, contextualized items are constructed based mainly on either conventional wisdom or on writing rules established based on non-contextualized items. Therefore, contexts used may interfere with the target construct. Contexts can potentially evoke students' relevant knowledge, thereby enhancing their understanding of the content and/or process skills that are required to respond to what the items are asking. However, contexts can also elicit irrelevant information and mislead students in ways that can result in their providing incorrect responses due to misinterpreting the required task, which can lead to inaccurate inferences about student learning (Ahmed & Pollitt, 2001; Leighton & Gokiert, 2005).

The use of context in the literature of item development has been associated with item sets, also called item bundles, items clusters, or testlets. We use the term *testlet* (Wainer & Kiely, 1987) to refer to the group of items (bundles or clusters) related to a single context. Testlets have one or more introductory paragraphs that state the problem or describe a scenario, followed by a set of items that can be presented in any format. A testlet can contain five to ten items (Haladyna, 1994). However, single items, each with its own context, are also common. The PISA items in a testlet can have up to three levels of contexts: (a) *general context* that introduces the entire testlet; (b) *subtestlet context* that may be introduced for a subset of, say, two items in the testlet; and (c) *item context*, a context presented at the individual item level.

Approach to Analyzing Contexts in Items

The *context profiling coding approach* we proposed (Ruiz-Primo & Li, 2012) uses logical analysis based on judgmental evidence to profile the contexts. The approach proposes dimensions and aspects to guide the judgment of the characteristics of the *item*, the *context*, and the *context in relation to the item question*[3].

The *item characteristics* coding focuses on capturing information on five aspects of the item: format, layout, context level, cognitive

[3] In other papers, we (Li, 2001; Ruiz-Primo, 2003, 2007; Shavelson, Ruiz-Primo, Li, & Ayala, 2002) have used the term *logical analysis* to describe our approach to analyzing the features of items. We have contrasted logical analysis with *empirical analysis*, the latter of which involves collecting and summarizing students' responses to the assessment item. The latter analysis empirically examines both the assessment-evoked cognitive activities and the observed student performance

demands, and item location in the testlet. The *context characteristic* focuses on four main dimensions of context characteristics: complexity, abstractness, resources, and nature. Each of these dimensions involves different aspects that provide more detailed descriptive features for the item contexts. For example, the dimension of complexity includes characteristics of extension of the context, type of vocabulary, number of main ideas, and reading load. These dimensions enable the coding of characteristics of the context without considering the item question. Unlike the first two sets of codes, the third set of codes requires the coding of *context in relation to the item question*. These codes focus on judging the context's function and properties in relation to each of the item questions in a testlet. Table 1 provides a brief description of the dimensions used in this study. A detailed explanation of the codes is provided in Ruiz-Primo and Li (2012, 2015).

When testlets are used, it is important to judge the role of the general context (or the subcontext and/or item context) relative to each of the items forming a testlet. The profiling approach uses items as the unit of analysis based on the idea that contexts, at any level, should be analyzed in relation to the item that they are linked with since information provided in the contexts needs to be processed or filtered eventually by examinees to generate responses to each individual item. Therefore, this proposed approach captures the characteristics of every level of context identified in any given testlet.

In our previous papers using the coding approach, we have acknowledged the difficulty and subjectivity involved in profiling contexts. The coding approach should be thought of as an exploratory strategy for identifying those context characteristics that deserve more attention as potential factors that can be linked to students' performance. We think this is a necessary step that has not been done previously. The coding approach should be viewed "as the first step in operationalizing these dimensions" (Ruiz-Primo & Li, 2015, p. 14). Further studies using student talk-aloud protocols, interviews, and experimental design of test items are needed to improve and refine the coding approach.

Table 1. *Examples of Dimensions and Aspects Used in This Study to Capture Characteristics of Context*

Dimension and Aspects	**Description**	**Examples of Coding Criteria**
Characteristics of the Item…Judged based on		
Cognitive demand	The type of knowledge judged to be tapped by the item.	• General reasoning • Declarative • Procedural • Schematic
Characteristics of the Context…Judged based on		
Abstractness		
Abstractness of main ideas	Presence of abstract ideas.	• All concrete • Some abstracts, some concrete • All abstract
Use of concrete reference	Whether the abstract ideas identified have any concrete reference, such as objects or events.	• Link between all abstract ideas and concrete reference • Link between some of the abstract ideas and concrete reference • No link between abstract ideas and concrete references
Resources		
Type of resource	Any other nonverbal material(s) provided.	• Diagram • Graph • Map
Nature of Context		
Setting	The setting described in the context.	• Everyday house activity • Professional/workplace • Scientific informational

Context of the Study

The study presented in this paper uses 52 released and secured *PISA science items* that were the focus of our previous exploratory study (Ruiz-Primo & Li, 2015)[4].In this study, we analyze the cognitive demands of those items and how these demands are linked to the characteristics of context and the students' performance. We approached the analysis of the cognitive dimensions of items using a framework that we have used in the past and that has been empirically tested in the context of science assessments, with confirming results around the categories proposed (Li, 2002; Li & Shavelson, 2001; Li, Ruiz-Primo, & Shavelson, 2006; Yin, 2005). We chose to focus only on context characteristics at the general context level for two reasons: (1) any analysis of items with multiple levels of contexts would reduce the sample size of items from 52 to 40, leaving even more constrained statistical power to adequately evaluate the association between cognitive demand of items and context, and (2) because we think the results have a higher potential to impact item developers' decisions when constructing the items and their contexts.

The PISA Framework. The current PISA science framework differs (OECD, 2015 from the one for the 2006 and 2009 administrations. For the purposes of this study, we describe the framework used for the 2006 and 2009 items included in this study[5]. Overall, the PISA science framework focuses on scientific literacy: "What is it important for citizens to know, value and be able to do in situations involving science and technology?" (OECD, 2009, p. 126). The framework is divided into three components: knowledge, competencies, and attitudes. The items used in this study tap the following dimensions of the framework: (1) disciplinary areas (i.e., physical systems, living systems, earth and space systems, and technological systems), (2) science practices (e.g., scientific inquiry and scientific explanations), and (3) competencies (i.e., identify scientific issues, explain phenomena scientifically, and use scientific evidence). In this study, we focus on one dimension of the PISA items, the performance competencies, because of the importance that these competencies have on scientific literacy to articulate the underlying construct of interests.

The importance of scientific literacy in the PISA's framework makes the use of context in items a critical aspect in the test. These contexts are organized using two dimensions: (1) types of context – personal (self, family, and peer groups), social (the community), and global (life across the world), crossed with (2) areas of application – health, natural resources, environment, hazard, and frontiers of science and technology.

Context Characteristics and Students' Performance. In a previous exploratory study (Ruiz-Primo & Li, 2012, 2015), we probed the importance of PISA context characteristics. We learned that, in the case of the PISA items, contexts should be analyzed by level. This result was fairly consistent across many of the coded aspects. Whether a context is located at the general or specific level, its characteristics are connected to noticeable differences in student performance. For example, a general context may *not be necessary* to respond to an item in a testlet, but the context at the item level can be directly related to the item question.

We also learned that the roles that contexts play vary by level of context, and items with differing context roles tend to be differently correlated with how well students perform on the items. For general contexts, student performance was statistically lower on items with contexts playing a description role (43.61 as the percentage of correct responses [PCR]) than on those with contexts playing a scenario

[4] Having access to secured items proved to be extremely difficult for the first study. Assuming that processes have not changed, getting new secured items would require months before the permission is granted by the Organisation for Economic Cooperation and Development (OECD).

[5] The 2015 PISA science framework labels the three competencies as explaining phenomena scientifically, evaluating and designing scientific enquiry, and interpreting data and evidence scientifically. In addition, it includes three types of scientific knowledge, which are content, procedural, and epistemic

role (59.28 as the PCR) with Cohen's *D* = 1.01; this difference was insignificant for specific contexts, although the same trend appeared as well.

Resources and *level of abstraction* are dimensions worth considering as critical sources of difficulty or sources of easiness. In general contexts, number of resources (whether drawings, graphs, maps, or tables) and how relevant the resource was to the item question were significantly correlated with the difficulty level of the items in opposite directions; that is, an item becomes more difficult when more resources are presented in its context, but easier if the resources are relevant to the question asked. This finding was confirmed by a separate analysis by gender. Level of abstraction also affects item difficulty, at least for specific contexts; item contexts with mainly abstract ideas that lacked references to concrete objects or events were associated with a lower average difficulty. The other two dimensions, complexity and connectivity, may require development, elaboration, reexamination, and refinement of aspects or codes to help disentangle their relationship with student performance.

The next section describes our exploratory study using the PISA items and explains how we approach the context characteristics and the study of the relationships between different context characteristics and student performance.

Methodology

Items. The original sample included 52 items embedded in 17 science testlets from PISA 2006 and PISA 2009, including 39 secured items and 13 released items. The sample included 4 testlets administered in PISA 2006, 11 administered in both PISA 2006 and PISA 2009, and 2 administered in PISA 2009 and previous PISA administrations. Sixty-five percent of the items required students to select an option, and 23% had students construct a response. Forty-eight percent of the items were classified by PISA as "explaining phenomena scientifically," 36% as "using scientific evidence," and the rest as "identifying scientific issues."

For the student performance, we performed the analysis with seven English-speaking countries: Australia, Canada, England, Hong Kong-China, Singapore, New Zealand, and the United States. Examining data patterns across student groups allowed us to examine how student characteristics interact with context features. For simplicity, we only report and discuss the findings from the United States sample; we note differences with other countries where they arise at the end of the results section.

Coding Cognitive Demands. Analysis of the items' cognitive demands was done during item coding. The approach used to code cognitive demands is based on three interdependent types of knowledge. In this approach we assume that type of knowledge lies on a continuum from concrete to abstract, from bits of information to high levels of organized knowledge. It is assumed that performance at higher levels of scientific literacy should be linked to certain types of knowledge. The cognitive demands were coded using three types of codes about knowledge:

1. **Declarative Knowledge – Knowing What**. This type of knowledge ranges from discrete and isolated content elements such as terminology or facts, to more organized knowledge forms such as statements, definitions, and knowledge of classifications, categories, and principles (e.g., mass is a property of an object).
2. **Procedural Knowledge – Knowing How**. This type of knowledge involves knowledge of skills, algorithms, techniques, and methods. Usually, this knowledge takes the form of "if-then" production rules or a sequence of steps that are the result of consensus, agreement, or disciplinary norms (Anderson et al., 2001). It is viewed as skill knowledge (Royer, Ciscero, & Carlo, 1993), and ranges from motor procedures (e.g., massing an object with a triple-beam scale) to more complex procedures (e.g.,

implementing a multistep sequence of actions to find out the density of an object when neither the mass nor the volume is known). Procedural knowledge can be automatized by practice, which allows it to be retrieved and executed without deliberate attention.

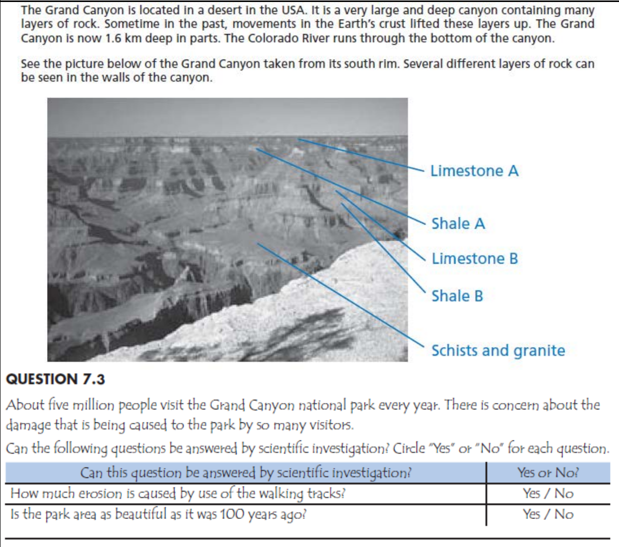
3. **Schematic Knowledge – Knowing Why**. This type of knowledge involves information that is interconnected in more organized bodies of knowledge, such as schemas, mental models, or "theories" (implicit or explicit). This type of knowledge allows students to approach novel and ill-defined problems, explain phenomena, or predict an outcome. Schematic knowledge and procedural knowledge are combined in the process of reasoning to design experiments or to solve mathematical problems.

Two more codes were used to capture item characteristics:

4. **General Reasoning** – This code was used to account for those items that could be responded to, even correctly, with no or minimal knowledge. These items require only careful reading or logical reasoning to figure out the correct response.
5. **Hard to Know**. This code was used to account for those items in which it was hard to know exactly what type of knowledge was required to respond, but we were sure that the item could not be responded to by test-taking skills.

The codes were mutually exclusive; only one code per item could be selected. Figure 2 presents examples of four items and the cognitive demand codes. The cognitive demand codes were judged based on a logical analysis of what knowledge students need to bring after examining both the context provided and the task asked in the prompt. For instance, the example in Figure 2a was coded as tapping general reasoning because the wording of the two questions in provides multiple cues to students to arrive at the right answers without substantive scientific knowledge (e.g., the first question indicates a causal relationship and uses a scientific term whereas the second question mostly implies a subjective judgment). The example in Figure 2b requires students to recall pieces of factual knowledge related to the diseases possibly associated with exposure to ultra-violet rays whereas and was therefore coded as declarative knowledge. Example in Figure 2 c was coded as a procedural knowledge item asks students to identify the laboratory equipment needed to determine the electric conductivity of materials. Example in Figure 2d was coded as a schematic knowledge since students most likely need to explain a described phenomenon by applying their mental model for phase change, temperature, and volume of water.

(a) Example of a PISA item coded as Tapping General Reasoning



The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.

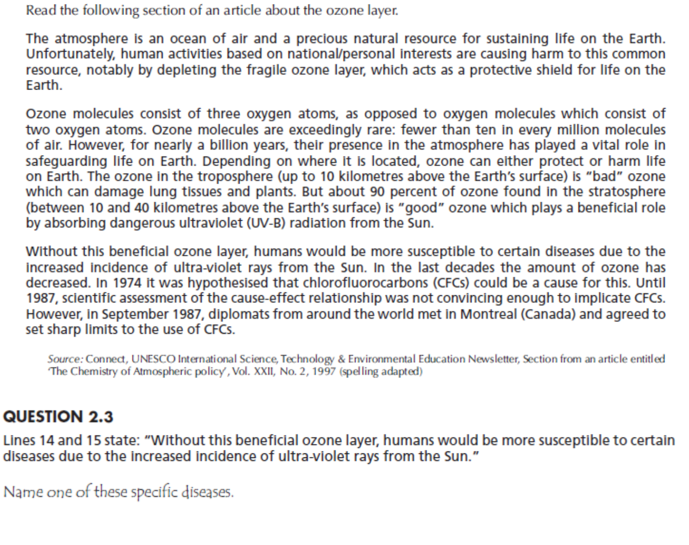
QUESTION 7.3

About five million people visit the Grand Canyon national park every year. There is concern about the damage that is being caused to the park by so many visitors.

Can the following questions be answered by scientific investigation? Circle "Yes" or "No" for each question.

Can this question be answered by scientific investigation?	Yes or No?
How much erosion is caused by use of the walking tracks?	Yes / No
Is the park area as beautiful as it was 100 years ago?	Yes / No

(b)Example of a PISA item coded as Tapping Declarative Knowledge



Read the following section of an article about the ozone layer.

The atmosphere is an ocean of air and a precious natural resource for sustaining life on the Earth. Unfortunately, human activities based on national/personal interests are causing harm to this common resource, notably by depleting the fragile ozone layer, which acts as a protective shield for life on the Earth.

Ozone molecules consist of three oxygen atoms, as opposed to oxygen molecules which consist of two oxygen atoms. Ozone molecules are exceedingly rare: fewer than ten in every million molecules of air. However, for nearly a billion years, their presence in the atmosphere has played a vital role in safeguarding life on Earth. Depending on where it is located, ozone can either protect or harm life on Earth. The ozone in the troposphere (up to 10 kilometres above the Earth's surface) is "bad" ozone which can damage lung tissues and plants. But about 90 percent of ozone found in the stratosphere (between 10 and 40 kilometres above the Earth's surface) is "good" ozone which plays a beneficial role by absorbing dangerous ultraviolet (UV-B) radiation from the Sun.

Without this beneficial ozone layer, humans would be more susceptible to certain diseases due to the increased incidence of ultra-violet rays from the Sun. In the last decades the amount of ozone has decreased. In 1974 it was hypothesised that chlorofluorocarbons (CFCs) could be a cause for this. Until 1987, scientific assessment of the cause-effect relationship was not convincing enough to implicate CFCs. However, in September 1987, diplomats from around the world met in Montreal (Canada) and agreed to set sharp limits to the use of CFCs.

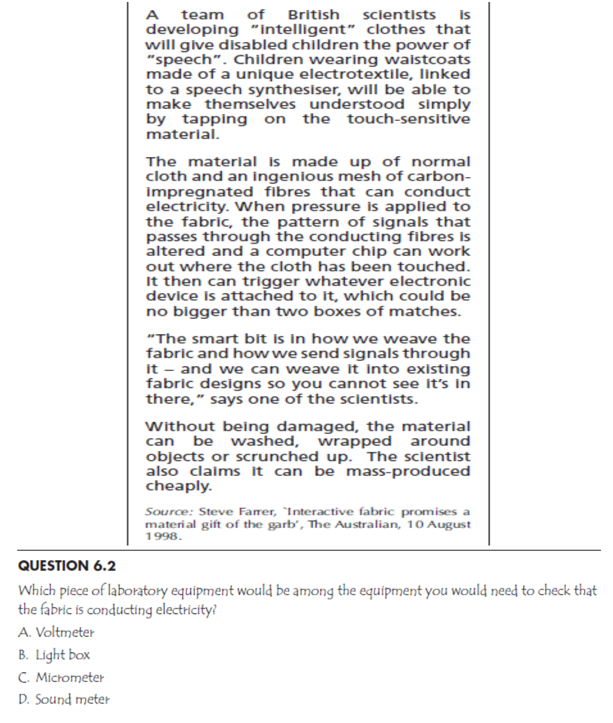
Source: Connect, UNESCO International Science, Technology & Environmental Education Newsletter, Section from an article entitled 'The Chemistry of Atmospheric policy', Vol. XXII, No. 2, 1997 (spelling adapted)

QUESTION 2.3

Lines 14 and 15 state: "Without this beneficial ozone layer, humans would be more susceptible to certain diseases due to the increased incidence of ultra-violet rays from the Sun."

Name one of these specific diseases.

(c) Example of a PISA item coded as Tapping Procedural Knowledge



A team of British scientists is developing "intelligent" clothes that will give disabled children the power of "speech". Children wearing waistcoats made of a unique electrotextile, linked to a speech synthesiser, will be able to make themselves understood simply by tapping on the touch-sensitive material.

The material is made up of normal cloth and an ingenious mesh of carbon-impregnated fibres that can conduct electricity. When pressure is applied to the fabric, the pattern of signals that passes through the conducting fibres is altered and a computer chip can work out where the cloth has been touched. It then can trigger whatever electronic device is attached to it, which could be no bigger than two boxes of matches.

"The smart bit is in how we weave the fabric and how we send signals through it – and we can weave it into existing fabric designs so you cannot see it's in there," says one of the scientists.

Without being damaged, the material can be washed, wrapped around objects or scrunched up. The scientist also claims it can be mass-produced cheaply.

Source: Steve Farrer, 'Interactive fabric promises a material gift of the garb', The Australian, 10 August 1998.

QUESTION 6.2

Which piece of laboratory equipment would be among the equipment you would need to check that the fabric is conducting electricity?

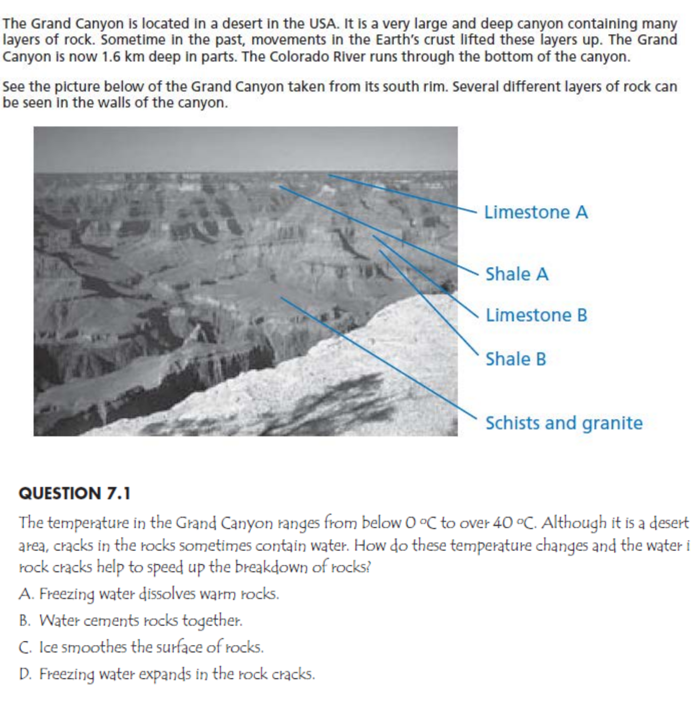
A. Voltmeter

B. Light box

C. Micrometer

D. Sound meter

(d) Example of a PISA item coded as Tapping Schematic Knowledge



The Grand Canyon is located in a desert in the USA. It is a very large and deep canyon containing many layers of rock. Sometime in the past, movements in the Earth's crust lifted these layers up. The Grand Canyon is now 1.6 km deep in parts. The Colorado River runs through the bottom of the canyon.

See the picture below of the Grand Canyon taken from its south rim. Several different layers of rock can be seen in the walls of the canyon.

QUESTION 7.1

The temperature in the Grand Canyon ranges from below 0 ºC to over 40 ºC. Although it is a desert area, cracks in the rocks sometimes contain water. How do these temperature changes and the water i rock cracks help to speed up the breakdown of rocks?

A. Freezing water dissolves warm rocks.

B. Water cements rocks together.

C. Ice smoothes the surface of rocks.

D. Freezing water expands in the rock cracks.

Figure 2. Examples of PISA items with different cognitive demands

Context Profiling Coding. Four released item testlets of 12 items were used to train coders and refine the coding approach. Due to the tuning of the coding system, it was not possible to formally test the consistency between coders before the secured items were coded at the National Center for Education Statistics (NCES) in the USA. Coding took place over 12 hours in 1.5 days, the maximum time allowed by NCES to code items. Items were coded as they were presented in carpets by year. To assure coding agreement, two coders conducted the consensus coding with 50% of the items, eight testlets of 26 items. A single coder coded the remaining testlets.

Context Codes Selected. We chose to study four aspects of the contexts, three of which were found to be significantly related to student performance, two from the *level of abstractness* dimension, and one from the *resources* dimension. We also included a fourth aspect, the *setting* of the context because we considered this context characteristic to be an important factor that should be taken into account during the item writing phase. In addition, we studied the *PISA performance competencies*, which were directly taken from the item description as defined by the item developers as part of the PISA item specification.

It is important to note that for those codes with two or less PISA items, we recoded the values by collapsing categories. For example, nature of setting originally had five codes (e.g., classroom or school, daily house-related), but we recoded into only two codes: *personal context* that combined both daily life, role playing, and school settings (12 and 4 items, respectively) and *non-personal* that combined professional/workplace and scientific/informational texts (34 and 2 items, respectively). At the end, we recoded all aspects of the contexts in three dimensions: *abstractness*, *resources*, and *nature of context*. This recoding allowed avoiding cells with fewer than two observations and met the necessary assumptions for conducting the selected statistical analysis. We did not recode cognitive demands or the performance competencies.

All of the codes used in the analysis are nominal variables. We also include the PCR, a continuous variable. A higher PCR indicates that items are relatively easier or students perform better compared to lower values of PCR. Table 2 provides a summary of the variables used in this study organized as outcomes, predictors, and covariant.

Table 2. *List of Variables Included in Data Analysis and Their Values*

	Variables Used	Description	Values
Outcome	Cognitive demand	The type of knowledge judged to be tapped by the item.	• General reasoning • Declarative • Procedural • Schematic
Predictors	**Abstractness of the Context**		
	Abstractness of main ideas[a]	Presence of abstract ideas	• All concrete • At least some abstract ideas are present, along with some or no concrete ideas
	Easiness of connectedness[a]	Whether the different main ideas presented in the context are easily linked	• Easy to connect • Some or all abstract ideas are hard for students to make connections
	Resources in the Context		
	Type of resource involved[a]	Any other nonverbal material provided	• All information is textual • Visual representation • Schematic representation (e.g., data table or graph)
	Nature of the Context		
	Type of setting[a]	The setting described in the context	• Personal context • Non-personal context (i.e., social context or global context classified by PISA)
	PISA Framework		
	PISA performance competencies	Categories of student performance based on the PISA framework	• Explaining phenomena scientifically • Identifying scientific issues • Using scientific evidence
Covariant	Percentage of correct responses (PCR)	The percent of examinees who responded to an item correctly	Ranging from 0 to 100

[a] Values were recoded to collapse some levels for these variables so that each code level had a reasonable number of assessment items

In order to evaluate how the cognitive demand of items is linked to context characteristics and student performance, we first performed the crosstable to produce the descriptive statistics items differing in cognitive demand. We then ran the multinomial logistic regression with the cognitive demand as the outcome variable, context characteristics as the predictors, and PCR as the covariant. As an extension to the logistic regression, the multinomial logistic regression is suitable for a nominal variable. This method enabled us to determine the relationship among categorical variables by predicting the probabilities of the different possible cognitive demand of items, given their context characteristics and the covariant of PCR. In other words, the regression analysis estimates the likelihood of an item's cognitive demand with a given profile of context characteristics and item difficulty.

Results

Analyses were carried out with both delta statistic (the difficult parameter estimated based on the Item Response Theory approach; see PISA 2006 technical report) and the percent of correct responses, which is equivalent to the passing rate of students who correctly answered the questions (see PISA 2006 technical report). The correlation between the two indicators, the delta statistics and the percentage of correct responses, was very high (r = −0.994), suggesting that the items were ranked in terms of difficulty very similarly by the two indicators. To respond to our research questions, we report the percentage of correct responses, interpreting this measure is comparitively straightforward.

The first research question, *What are the cognitive demands of the sampled PISA released and secured items and what is the students' performance linked to these items*?, is answered by presenting the percent of items by type of knowledge. This information is presented in Table 3. From the 52 items coded, only one was found to be too ambiguous to determine its cognitive demand in terms of the underlying knowledge required. The highest percentage of items was found for procedural knowledge and the lowest for declarative knowledge. It is interesting to note that similar percentages are found for items requiring schematic knowledge to respond and those requiring only testing skills. There is no difference observed, as revealed by the one-way analysis of variance across the four levels of cognitive demand, in the percent of correct responses by type of knowledge tapped by the items (F(3, 47) = 0.418, p = 0.741).

Table 3. *Released and Secured PISA Items and PCA by Type of Knowledge (N = 52)*

Type of Knowledge	Frequency	Percentage	Averaged PCR
Declarative	8	15.4	50.97
Procedural	16	30.8	55.96
Schematic	14	26.9	50.67
General Reasoning	13	25.0	48.98
Hard to Know[a]	1	1.9	-

a. The item with the "hard to know" code was excluded for the further analysis

To address the second question, *Are the items classified by cognitive demands associated with certain characteristics of the contexts of the items that proved to be linked to students performance?*, we focus on the four context characteristics and the PISA performance competency. We first report the descriptive statistics of the predictors broken down by items' cognitive demand. We then describe the findings from the multinomial logistic regression in which we evaluated how PCR as a covariate and item context characteristics are associated with items' cognitive demand in the form of the type of knowledge required.

Profiling PISA Items Tapping Different Cognitive Demands

Level of Abstraction. For this dimension, we focused on abstractness of main ideas and the use of concrete reference(s) in an item's context(s). Both characteristics were found to be significantly associated with student performance patterns in our previous study. Item contexts with mainly abstract ideas that lacked references to concrete objects or events were associated with a lower average difficulty. We hypothesized that if the ideas presented in the context were too abstract and not appreciably linked to concrete objects or events, then this could potentially interact with students' comprehension of the information presented in the context and therefore, with their understanding of the question asked and their response to that question. Table 4 provides information about the joint distribution broken down by type of knowledge and abstractness of the context. The majority of the contexts that present mostly concrete ideas and involve references to concrete objects or event were found to assess either general reasoning or procedural knowledge. It is interesting to note that there is not much difference in the profiles between items tapping procedural and schematic knowledge.

Table 4. *Items Classified According by Type of Knowledge and Abstractness of Ideas in the Context of the Items*

Type of Knowledge	Abstractness of Main Ideas: All Ideas Are Concrete	Abstractness of Main Ideas: At Least Some Are Abstract	Abstractness of Main Ideas: Total	Use of Concrete Reference: All Linking to Concrete Objects or Events	Use of Concrete Reference: Some or No Links to Concrete Obj./Events	Use of Concrete Reference: Total
General reasoning	9	4	13	3	10	13
Declarative	2	6	8	4	4	8
Procedural	6	10	16	4	12	16
Schematic	5	9	14	6	8	14
Total	22	29	51	17	34	51

Context Resources. We labeled as *context resources* all diagrams, tables, pictures, graphs, or photographs that were part of the general context of the testlets. In our previous study, we hypothesized that resources could potentially affect students' performance in either direction, focusing the attention of the student on appropriate item aspects or misleading students' attention by having ambiguous resources that were unnecessary to respond to the items. Our hypothesis was confirmed, and we found a link between *student performance and amount of resource as well as relevance of the resource* (i.e., unnecessary to understand the text of the contexts, merely supplementary to understanding the text of the context; and essential to understanding the text of the context and may provide information not otherwise conveyed in the text).

In this paper we focus on the type of resource by recoding items whose contexts without any resource as "textual information only." This allows us to include all the items in the analysis instead of only those items with at least one resource. Table 5 provides information about the joint distribution between the two variables. Items in which schematic representations are included in contexts appear to primarily assess students' procedural knowledge in contrast to declarative and schematic knowledge. On PISA science tests, it is more common for the procedural items to ask students to interpret data presented in either a table or a graph.

Table 5. *Items Classified by Type of Knowledge and Type of Resource in the Context*

Type of Knowledge	Textual Information Only	Graphic Representation(s) Included	Schematic Representation(s) Included	Total
General reasoning	4	7	2	13
Declarative	4	1	3	8
Procedural	5	4	7	16
Schematic	3	8	3	13
Total	16	20	15	51

Setting of Context. We grouped the contexts into two types of settings: (1) *personal* as everyday household or students' school work that examinees are familiar with; and (2) *non-personal* as those contexts that describe events, phenomena, or activities at a professional workplace or using informational scientific texts. The non-personal setting can be easily aligned to the PISA framework of social and global contexts. Our non-personal category includes both. In our previous study, we found that items with scientific information tend to be more difficult than those with daily household settings, a finding that might be related to the level of reading demanded by texts or students' familiarity with household settings in daily life. In the present study, we focus on the association of item contexts by type of knowledge tapped in the item and the context settings. Table 6 provides information about the distribution between type of knowledge and type of context setting. Personal settings are seemingly less suitable for assessing schematic knowledge compared to contexts that are concerned with social or global matters, whereas the contrary is observed in the non-personal settings. The highest frequency of non-personal settings is observed for items tapping schematic knowledge.

Table 6. *Items Classified by Type of Knowledge and Setting of the Context*

Type of Knowledge	Personal	Non-personal	Total
General reasoning	4	9	13
Declarative	5	3	8
Procedural	6	10	16
Schematic	0	14	14
Total	15	36	51

Linking Cognitive Demand of Items to Context Characteristics and Student Performance

In this section, we undertake a closer examination of the possible association between the cognitive demand of items and context characteristics and student performance. A test of the full model with the four context characteristics and the PISA performance competencies as predictors and the PCR as the covariant was statistically significant, indicating that these characteristics overall were associated with the cognitive demands tapped by the items (x^2 [24, N = 51] = 78.55, p = .000). Nagelkerke's R^2 of .841 indicated a strong relationship between prediction and grouping of four cognitive demands.

Table 7 reports the main effects for the predictors and covariant in predicting the level of cognitive demand. All five context or item characteristics except *use of concrete reference* significantly contribute to the prediction of the cognitive demand of items, among which the setting of context and PISA performance competencies show the strongest association with the levels of cognitive demand.

Table 7. *Likelihood Ratio Tests for the Main Effects in Predicting the Type of Knowledge*

Main Effect Tested	Likelihood Ratio Chi-Square Test	df	Statistical Significance
Predictors			
Abstractness of main ideas	10.366	3	.016
Use of concrete reference	1.695	3	.638
Type of resource included	15.222	6	.019
Setting of context	21.974	3	.000
PISA performance expectation	40.523	6	.000
Covariant			
PCR (percentage of correct responses)	9.416	3	.024

A detailed summary of the regression results is reported in Table 8, where each level (code) of each context aspect (e.g., for the aspect *abstractness of main ideas*, the code included is “all ideas are concrete”) was evaluated against each item’s cognitive level (e.g., declarative knowledge) to determine how each code was associated to the items’ cognitive demands. That is, for each code used for each aspect of the contexts evaluated, we carried out pairwise comparisons across the four levels of cognitive demand. To avoid presenting all the possible comparisons, the table includes only the results for which statistical significance (p < .05) and statistical trending (p < .10) were found. For example, for the code “all ideas are concrete,” instead of presenting the six comparisons (all the possible comparisons across all the types of cognitive demand), we provide only those that show a p value of < .10. Therefore, when the context has the characteristic of “all ideas are concrete,” the odds of being classified as a general reasoning item are more than 30 times greater as a declarative knowledge item.

Table 8. *Selected Pairwise Comparisons of the Main Effects: Abstractness, Resources, Setting of Contexts, Performance Competencies, and Percent of Correct Responses**

Focal Type of Cognitive Demand	Comparison Type of Cognitive Demand	B[a]	Standard Error of B	Wald Test	Statistical Significance (df = 1)	Exp(B) as the Change In Odds[b]
Abstractness of main ideas: All are concrete						
Reasoning	Declarative	8.939	4.392	4.143	.042	>30
Reasoning	Procedural	3.389	1.649	4.223	.040	29.623
Reasoning	Schematic	5.968	3.513	2.886	.089	>30
Resources: Only visual representations are included						
Reasoning	Declarative	10.765	5.165	4.344	.037	>30
Reasoning	Procedural	2.678	1.448	3.421	.064	14.561
Declarative	Schematic	-9.047	4.615	3.844	.050	.000
Setting of context: Personal						
Reasoning	Schematic	27.051	1.302	431.663	.000	>30
Declarative	Schematic	30.106	2.321	168.220	.000	>30
Procedural	Declarative	-4.118	2.321	3.147	.076	.016
PISA performance competencies: Explaining phenomena scientifically						
Declarative	Procedural	29.290	3.551	68.037	.000	>30
Percent of correct responses						
Reasoning	Schematic	.109	.064	2.910	.088	1.115
Procedural	Schematic	.160	.066	5.874	.015	1.174
Procedural	Declarative	.137	.066	4.334	.037	1.146

*The table only reports findings with a p < .10.

a. B, as the logit slope coefficient, indicates the effect of a unit of change in the predicting variable on the predicted logits with the other variables in the model held constant.

b. Odds ratios are the effect of one unit of change in the predicting variable in the predicted odds ratio when other variables held constant.

We found that contexts in which the main ideas are all concrete are more likely to be classified as eliciting general reasoning skills instead of any type of knowledge. This suggests that item developers need to incorporate some abstract ideas in order to maintain the cognitive demand of assessment items. When contexts are situated in personal settings, our results indicated a higher likelihood of the items being associated with general reasoning or declarative knowledge compared to schematic knowledge. With respect to the type of resource involved, results revealed that contexts with only graphic representations are more likely to be associated with general reasoning skills rather than declarative or procedural knowledge. Finally, items with visual representations have a statistically higher probability of eliciting schematic knowledge than declarative knowledge.

The performance competencies proposed by PISA were found to be a strong predictor for item cognitive demand as well. Items designed to assess students' competency in *explaining phenomena scientifically* were significantly associated with declarative knowledge instead of procedural knowledge. Lastly, it is interesting to note that students performed relatively better on procedural items; in other words, items with the demand of procedural knowledge were easier for students than those requiring declarative and schematic knowledge. A similar trend was identified for items that require general reasoning when compared against items of schematic knowledge[6].

We replicated the regression analysis with other six countries for the entire sample as well as for each of the two gender groups. Interestingly, the results of four OECD counties (i.e., Australia, Canada, England, and New Zealand) were found comparable to these of the USA sample. The 15-year olders in Singapore and Hong Kong-China performed differently on items varying in cognitive demand, therefore the performance patterns on items of different cognitive demands were more apparent than their peers from US and other four countries. Due to this very reason, item difficulty became a stronger predictor of the level of items' cognitive demand in the regression model. For instance, the setting of item contexts was found no longer significantly associated with items' cognitive demands for Singapore students. Likewise, the presence of visual representations in contexts was not a predictor for cognitive demands for Hong Kong students.

Discussion

Despite the years of experience developing assessments and the rules and principles developed to construct assessment items, what it is known about how to develop good items and good context remains somewhat of a mystery. Indeed, as many as 40% of new items fail to perform as intended when first tried (Haladyna, 1994). Developing effective test items is still more art than science. We still do not know how to write items that unequivocally pass all the quality tests. Reviewing test items accounts for testing companies' largest expenditure of time and money. Different guidelines, models, or frameworks proposed to develop items and assessments over time (Anderson, 1972; Downing, 2006; Haladyna, 1994; Haladyna & Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Mislevy & Riconscente, 2006; Shoemaker, 1975) do not provide clear guidance regarding how to effectively develop assessment items, and none of the proposed models or frameworks address how to develop

[6] This finding about the link between the item's cognitive demand and item difficulty can bring insights when comparing student performance patterns across countries. For instance, among the seven countries that we analyzed, students from Hong Kong performed overwhelmingly well on items that require general reasoning, procedural knowledge, and schematic knowledge, but appeared to be mediocre on declarative items. In contrast, students from Singapore excelled on items of general reasoning, declarative, and procedural knowledge, but less so for schematic knowledge items, which is more comparable to students from USA and lower than students from the other five countries. These observations were just based on the descriptive statistics without considering the sampling weights for inferential statistical analysis

appropriate item context. Even using experienced item writers does not guarantee the production of high quality items (Haladyna, 1997). As Welch stated, "Considering that items are the backbone of the assessment industry, there is relatively little research on item writing" (2006, p. 306). In fact, not even in the various recent editions of the handbooks of *Educational Measurement* has the item writing process been carefully addressed. There is no doubt that item writing has not received the same attention that assessment statistical models have (Haladyna, 1994; Osterlind, 1998), and even less attention has been given to writing effective contextualized items. Of the writing rules usually cited in the literature, only a few are empirically based (Haladyna, 1994), while most are a collection of rules based on experiences and wisdom (Bormuth, 1970).

With this background in mind, it should not be a surprise, then, how little we know about the development and use of *context* in item prompts. It has been argued that contextualizing items makes the questions more concrete and less demanding (Ahmed & Pollitt, 2000, 2007). Furthermore, contextualizing items has been considered an effective strategy to test whether students can apply (transfer) their knowledge (Ahmed & Pollit, 2000, 2007; Boaler, 1994; Haladyna, 1997) and to measure complex thinking (Haladyna, 1994). However, we know little about whether any of these claims are true. There is virtually no empirical evidence about how much contextualization in items can help to obtain richer and more accurate assessment information about students' knowledge application or knowledge transfer. In the case of PISA, this type of evidence is critical in order to justify the use of contexts at so many different levels and with such abundance.

With this study, we contribute to the accumulation of evidence about the characteristics of contexts and their impact on student performance, evidence that we hope helps the assessment development field. We acknowledge that the sample size of the items coded is very small and that the results presented should be treated as exploratory. For example, constrained by the number of items, we were unable to examine the interaction effects of context characteristics, although in practice many characteristics can collectively affect student performance in a complex manner. Still, our study demonstrates that assessment items engage test takers in different types of cognitive processes (ranging from retrieving memorized information to conjecturing, interpreting, justifying, etc.). What context characteristics are associated with certain cognitive process? Are these characteristics and cognitive demands associated with student performance? These are the main questions we tried to respond to in this study.

First, we linked different characteristics of context with the cognitive demands of the items. We learned that items with contexts involving only concrete ideas are unlikely to require the use of content knowledge in the response. Items that tap procedural and schematic knowledge had contexts in which some abstract ideas were presented. If this result can be replicated, then we can start developing some guidelines to assist item developers. Why is this type of information important? Because the context used in an item may reduce the intended cognitive demands.

We also learned the importance of the type of resource used in defining the cognitive demands of the items: schematic representations seem to be linked to items tapping procedural knowledge rather than to items tapping declarative or schematic knowledge. Learning about the specific characteristics of the resources used in the contexts of items helps better tailor the targeted type of knowledge test takers need to engage in appropriate cognitive processes. It is also interesting to learn how the type of setting, personal or non-personal, is linked to certain cognitive demands. This study provides evidence that personal settings are less suitable than non-personal settings for assessing schematic knowledge.. Finally, it was interesting to learn that use of evidence was associated, as expected, with items tapping procedural knowledge but it was also associated with general reasoning. What do the contexts of items tapping procedural knowledge need to have to avoid correct responses based only on general reasoning?

Independent of the framework used to code cognitive demands in items, this exploratory study has demonstrated the importance of learning more about the characteristics of contexts and how they are related to the cognitive processes in which students need to engage. Learning more about the context characteristic in items is essentially a validity issue that should be considered by the OECD assessment development team. Furthermore, this team should share and report the pilot testing findings related to contexts to stimulate the measurement community to ask and explore why some contexts were found unhelpful or misleading. Allowing researchers to have access to secured items will help to move the field forward as well, ideally, with unconditional access. Limiting the time of access to study items does not support learning more deeply about the characteristics of the PISA items and how they can be improved.

References

Ad-Hoc Technical Committee on the Development of Technical Criteria for Examining Cultural Validity in Educational Assessment. (2015). *Promoting and evaluating cultural validity in the activities performed by the National Institute for Educational Evaluation (INEE)*. Submitted to the National Institute for Educational Evaluation. Mexico City, Mexico, January 16.

Basterra, M. R. (2011). Cognition, culture, language, and assessment. In M. R. Basterra, E. Trumbull, & G. Solano-Flores (Eds.), *Cultural validity in assessment* (pp. 72-95). New York: Routledge.

Bialystok, E. (2002). Cognitive processes of L2 users. In V. J. Cook (Ed.), *Portraits of the L2 user* (pp. 145-165). Buffalo, NY: Multilingual Matters.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice*, *9*(3), 387-399.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Paper Number 71*. Retrieved from OECD website: http://www.oecd-ilibrary.org/education/the-policy-impact-of-pisa_5k9fdfqffr28-en

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer Verlag.

Camilli, G. (2006). Test fairness. En R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 221-256). Westport, CT: American Council on Education and Praeger.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Thousands Oaks, CA: Sage.

Capacity Development Group (2007, May). *Capacity assessment methodology: User's guide*. Bureau for Development Policy, United Nations Development Program. New York, September 2005. Retrieved from the United Nations Development Programme website: https://www.unpei.org/sites/default/files/PDF/institutioncapacity /UNDP-Capacity-Assessment-User-Guide.pdf

Carnoy, M. (2015). *International Test Score Comparisons and Educational Policy: A Review of the Critiques*. Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/publication/international-test-scores

Clarke, M. (2012). *What matters most for student assessment systems: A framework paper*. Retrieved from https://openknowledge.worldbank.org/bitstream/handle/10986/17471/682350WP00PUBL0WP10READ0web04019012.pdf?sequence=1

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Darling-Hammond, Linda (2014). What can PISA tell us about U.S. education policy? *New England Journal of Public Policy*: *26*(1), Art. 4. Retrieved from http://scholarworks.umb.edu/nejpp/vol26/iss1/4

Dogan, E., & Circi, R. (2010). A blind item-review process as a method to investigate invalid moderators of item difficulty in translated assessment. In M. von Davier & D. Hastedt (Eds.), *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments* (Vol. 3) (pp. 157-172). Hamburg: IERI.

Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about uses of international assessments. *Teachers College Record*, *117*(1), 1-28.

Ercikan, K., Roth, W.-M., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education*, *27*, 275-285.

Ercikan, K., & Solano-Flores, G. (2016). Assessment and sociocultural context: A bidirectional relationship. In G. T. L. Brown & L. Harris (Eds.), *Human Factors and Social Conditions of Assessment*. New York: Routledge.

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: PREAL. Retrieved from http://www.uis.unesco.org/Education/Documents/Ferrer.pdf

Figazzolo, L. (2009). *Impact of PISA 2006 on the education policy debate*. Retrieved from http://download.ei-ie.org/docs/IRISDocuments/Research%20Website%20Documents/2009-00036-01-E.pdf

Gebril, A. (2016). Educational assessment in Muslim countries: Values, policies, and practices. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Gilmore, A. (2005). The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Back support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS). Retrieved from http://www.iea.nl/fileadmin/user_upload/Publications/Electronic_versions/Gilmore_Impact_PIRLS_TIMSS.pdf

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*. Mahwah, NJ: Lawrence Erlbaum.

Husén, T. (1983). *An incurable academic: Memoirs of a professor*. Oxford, UK: Pergamon Press.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25. doi: http://dx.doi.org/10.1086/648471

Kane, M. T. (1982). A sampling model of validity. *Applied Psychological Measurement*, *6*, 125-160. doi: http://dx.doi.org/10.1177/014662168200600201

Kane, M. T. (2006). *Validation*. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian-heritage societies. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Lingard, B., & Lewis, S. (2016). Globalization of the American approach to accountability: The high price of testing. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*. New York: Routledge.

Martínez-Rizo, F. (2015). Las pruebas ENLACE y EXCALE: Un estudio de validación. Retrieved from http://publicaciones.inee.edu.mx/buscadorPub/P1/C/148/P1C148.pdf

Messick, S. (1989). Validity. En R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education, Macmillan.

Messick, S. (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741-749. doi: http://dx.doi.org/10.1037/0003-066X.50.9.741

Mullis, I. V. S., & Martin, M. O. (2011). *TIMSS 2011 item writing guidelines.* Retrieved from http://timssandpirls.bc.edu/methods/pdf/T11_Item_writing_guidelines.pdf

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., & Preuschoff, C. (2009). *TIMSS 2011: Assessment frameworks.* Retrieved from http://timssandpirls.bc.edu/timss2011/downloads/TIMSS2011_Frameworks.pdf

National Project Managers' Meeting (2010, October). *Translation and adaptation guidelines for PISA 2012*. Doc: NPM10104e. PISA Consortium. Budapest, Hungary. Retrieved from https://www.oecd.org/pisa/pisaproducts/49273486.pdf

Organisation for Economic Co-operation and Development (OECD). (n.d.). *Programme for international student assessment (PISA): Results from PISA 2012, Country note: United States.* Retrieved from http://www.oecd.org/pisa/keyfindings/PISA-2012-results-US.pdf

Organisation for Economic Co-operation and Development (2006). *PISA released items: Mathematics*. Retrieved from http://www.oecd.org/pisa/38709418.pdf

Organisation for Economic Co-operation and Development (2010). *Translation and adaptation guidelines for PISA 2012.* Retrieved from http://www.oecd.org/pisa/pisaproducts/49273486.pdf

Lockheed, M., Prokic-Bruer, T., & Shadrova, A. (2015). *The experience of middle-income countries participating in PISA 2000-2015* (PISA series). Washington, D.C. & Paris: The World Bank & OECD Publishing. doi: http://dx.doi.org/10.1787 /9789264246195-en

Ravela, P. (Ed.). (2001). Los próximos pasos: ¿Hacia dónde y cómo avanzar en la evaluación de aprendizajes en América Latina? Document No. 20. Working Group on Assessment and Standards. Santiago: PREAL. Retrieved from http://campus-oei.org/calidad/grade.PDF

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, *16*(2), 5-8, 13. doi: http://dx.doi.org/10.1111/j.1745-3992.1997.tb00585.x

Sjøberg, S. (2007). PISA and "real life challenges": Mission impossible. In S. T. Hopmann, G. Brinek, & M. Retzl (Eds.), *According to PISA—Does PISA keep what it promises?* Berlin: LIT Verlag.

Solano-Flores, G. (2008, July). A conceptual framework for examining the assessment capacity of countries in an era of globalization, accountability, and international test comparisons. Comunicación presentada en la *6th Conference of the International Test Commission*, Liverpool, UK.

Solano-Flores, G. (2011). Assessing the cultural validity of assessment practices: An introduction. In M. R. Basterra, E. Trumbull, and G. Solano-Flores, *Cultural validity in assessment* (pp. 3-21). New York: Routledge.

Solano-Flores, G. (2016). Generalizability. En L. E. Suter, D. Wyse, E. Smith, & N. Selwyn (Eds.), *The BERA/SSAGE Handbook of Educational Research* (chap. 47). London: Sage.

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2006). Traducción y adaptación de pruebas: Lecciones aprendidas y recomendaciones para países participantes en TIMSS, PISA y otras comparaciones internacionales. *Revista Electrónica de Investigación Educativa (REDIE)*, *8*(2). Retrieved from http://redie.uabc.mx/redie/article/download/143/246

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L.A. (2009). Theory of test translation error. *International Journal of Testing*, *9*, 78-91.

Solano-Flores, G., Contreras-Niño, L.A., & Backhoff, E. (2013). The measurement of translation error in PISA-2006 items: An application of the theory of test translation error. In M. Prenzel, M. Kobarg, K. Schöps, & S. Rönnebeck (Eds.), *Research in the context of the programme for international student assessment* (pp. 71-85). Springer Verlag.

Solano-Flores, G., & Gustafson, M. (2013). Assessment of English language learners: A critical, probabilistic, systemic view. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large scale assessment in education: Theory, issues, and practice* (pp. 87-109). New York: Routledge.

Solano-Flores, G., & Li, M. (2006). The use of generalizability (G) theory in the testing of linguistic minorities. *Educational Measurement: Issues and Practice*, *25*(1), 13-22.

Solano-Flores, G., & Li, M. (2009). Generalizability of cognitive interview-based measures across cultural groups. *Educational Measurement: Issues and Practice*, *28* (2), 9-18.

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, *19*(2-3), 245-263. doi: http://dx.doi.org/10.1080/13803611.2013.767632

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, *38*(5), 533-573. doi: http://dx.doi.org/10.1002/tea.1018

Stachelek, A. J. (2010). Exploring motivational factors for educational reform: Do international comparisons dictate educational policy? *Journal of Mathematics Education at Teachers College*, *1*, 52-55.

Suter, Larry E. (2000). Is student achievement immutable? Evidence from international studies on schooling and student achievement. *Review of Educational Research*, *70*(4), 529-545. doi: http://dx.doi.org/10.3102/00346543070004529

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*, 295-312. doi: http://dx.doi.org/10.1016/0959-4752(94)90003-5

Tatto, M. T. (2006). Education reform and the global regulation of teachers' education, development and work: A cross-cultural analysis. *International Journal of Educational Research*, *45*, 231-241. doi: http://dx.doi.org/10.1016/j.ijer.2007.02.003

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment* (Chap. 21). New York: Routledge.

van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. En G. Brown & L. Harris (Eds.), *Handbook of human factors and social conditions of assessment*, (Chap. 25).New York: Routledge.

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

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind*. Cambridge, MA: Harvard University Press.

Wuttke, J. (2007). Uncertainties and bias in PISA. En S. T. Hopmann, G. Brinek, and M. Retzl (Eds.), *According to PISA – Does PISA keep what it promises?* Berlin: LIT Verlag.

Authors / Autores

To know more / Saber más

Ruiz-Primo, Maria-Araceli (aruiz@stanford.edu).

She is an associate professor at the Graduate School of Education at Stanford University. Her work, funded mainly by the National Science Foundation and the Institute of Education Sciences, examines the assessment of student learning both in the classroom and in large-scale assessment programs and classroom assessment practices. Her publications address the development and evaluation of diverse learning assessment strategies (e.g., concept maps and students' science notebooks), and the study of teachers' informal and formal formative assessment practices (e.g., the use of assessment conversations and embedded assessments). Her recent work focuses on the development and validation of assessments that are instructionally sensitive and instruments intended to measure teachers' formative assessment practices. Her address is: Graduate School of Education. 485 Lasuen Mall. Stanford University, Stanford, CA

Li, Min (minli@uw.edu)

She is associate professor at College of Education, University of Washington, Seattle. She is interested in understanding how student learning can be accurately and adequately assessed both in large-scale testing and classroom settings. Her work reflects a combination of cognitive science and psychometric approaches in various projects, including examining the cognitive demands of large-scale science items, using science notebooks as assessment tools, parameterizing the design of contextualized tasks, analyzing teachers' classroom assessment practices, and validating complex performance-based tasks for teachers and educational leaders. Her address is: University of Washington. College of Education. Seattle, WA 98195-3600. https://education.uw.edu/people/faculty/minli

RELIEVE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa

E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]

© Copyright, RELIEVE. Reproduction and distribution of this articles it is authorized if the content is no modified and their origin is indicated (RELIEVE Journal, volume, number and electronic address of the document).

© Copyright, RELIEVE. Se autoriza la reproducción y distribución de este artículo siempre que no se modifique el contenido y se indique su origen (RELIEVE, volumen, número y dirección electrónica del documento).