Language effects in international testing: The case of PISA 2006 science items

Yasmine H. El Masri,[a]* Jo-Anne Baird,[a] and Art Graesser[b]

[a]*Department of Education, Oxford University, United Kingdom*

[b]*Department of Psychology and Institute for Intelligent Systems, University of Memphis*

* *Corresponding author*

15 Norham Gardens, Oxford, OX2 6PY, United Kingdom, email: yasmine.elmasri@education.ox.ac.uk

Language effects in international testing: The case of PISA 2006 science items

Abstract

We investigate the extent to which language versions (English, French and Arabic) of the same science test are comparable in terms of item difficulty and demands. We argue that language is an inextricable part of the scientific literacy construct, be it intended or not by the examiner. This argument has considerable implications on methodologies used to address the equivalence of multiple language versions of the same assessment, including in the context of international assessment where cross-cultural fairness is a concern. We also argue that none of the available statistical or qualitative techniques are capable of teasing out the language variable and neutralising its potential effects on item difficulty and demands. Exploring the use of automated text analysis tools at the quality control stage may be successful in addressing some of these challenges.

Keywords: science assessment, PISA, international comparison, language bias, test transadaptation

Introduction

International large-scale assessments (ILSAs) such as the Programme for International Student Assessment (PISA) and the Trends in International Maths and Science Study (TIMSS) involve comparisons of student achievement from different language backgrounds. PISA, governed by the Organisation for Economic Cooperation and Development (OECD), assesses literacy of 15-year old students in reading, mathematics, science, and other cognitive constructs. TIMSS, administered by the International Association for the Evaluation of Educational Achievement (IEA), evaluates trends in mathematics and science achievement at fourth (Population 1) and eighth grade (Population 2).

The development and design of ILSAs have instigated much conceptual and technical controversy (e.g. Ercikan & Koh, 2005; Goldstein, 2004; Goldstein, Bonnet, & Rocher, 2007; Kreiner & Christensen, 2014; Prais, 2003). Some of the debates question the possibility of developing a test for which language versions allow fair comparisons of student achievement across countries. An equitable comparison of scores across countries sitting the same test should be established on the premise that the various language forms are equivalent for all participants and offer all parties equal opportunities of success.

Despite the rigorous quality control involved in the translation and adaptation processes of ILSAs, numerous studies have highlighted how items may behave differently across language versions and countries (Asil & Brown, 2016; Ercikan & Koh, 2005; A Grisay, de Jong, Gebhardt, Berezner, & Halleux-Monseur, 2007; Aletta Grisay & Monseur, 2007; Huang, Wilson, & Wang, 2016; Kreiner & Christensen, 2014; Oliveri & Ercikan, 2011; Sandilands, Oliveri, Zumbo, & Ercikan, 2013; Wu & Ercikan, 2006). These studies used statistical techniques generally referred to as differential item functioning (DIF) analyses to investigate potential bias across language versions within countries (e.g. Canada) and across countries. For instance, Ercikan and Koh (2005) identified disparities in the mathematics and science constructs measured by the English and French versions of TIMSS 1995 in Canada, France and USA. Moreover, Grisay and Monseur (2007) and Grisay et al. (2007) analysed PISA data and detected systematic differences in the way items behaved across countries. The magnitude of DIF increased with the distance between languages (e.g. French and Spanish, as romance languages, are considered closely related when contrasted with Arabic, a Semitic language, and hence relatively distant from both French and Spanish). The studies also found that the magnitude of DIF increased when comparing language versions administered in

countries with different GDP per capita. These findings were supported by Asil and Brown's (2016) analyses of PISA 2009 reading tests. In addition, Kreiner and Christensen (2014) detected significant variations in the way some items behaved in language versions of PISA 2006 reading survey and claimed that deleting 'biased' items would have serious implications on the reliability of country rankings. Moreover, Huang et al. (2014) analysed data of PISA 2006 science assessments of USA, Canada, Hong Kong and mainland China and found that curriculum coverage was the main factor that advantaged some countries over others.

The aforementioned studies provide substantial empirical evidence challenging the comparability of language versions of the same assessment. In this paper we argue that language is an inextricable part of the construct measured by science assessments and evidence our position by examining the science construct (namely scientific literacy) of PISA and the role of language in science attainment. We argue that scientific literacy is a composite variable (Maul, 2013a, 2013b), of which language is a dimension. With language an intrinsic part of the science construct, we contend that comparing language versions of the same science test is strictly methodologically indefensible, as translation effects are unavoidable with bias at some level being inevitable (Hambleton, Merenda, & Spielberger, 2005). This claim has potential serious implications for the validity and reliability of international tests. However, we cannot determine the extent from currently available methods for, as we will show later in this paper, these methods do not allow the identification of the sources of bias, nor where there is consistent bias throughout a test. This makes comparability of scores across languages extremely challenging. We thus require more appropriate and sensitive measurement techniques. We support our argument using the literature on cross-cultural comparisons and by using English, French and Arabic versions of some items, including PISA items. The paper concludes by exploring the possibility of using computational linguistics to gauge translation effects. Our evidence draws upon examples from ILSAs and exemplars that we have generated to illustrate specific points of language.

The Construct of Scientific Literacy

Cronbach and Meehl (1955, p. 283) defined a psychological construct as follows:

> A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct.

They further specify that the variable to be measured is formed by a universe of content, from which we need to sample and design items that are domain relevant, and develop from these a test that is representative of the domain and provides an adequate sample of the domain. Psychometricians have focused their attention on whether these and other technical requirements of tests are met. Wiliam (2010) argued that we need to shift the debate from the technical features of assessment (how well we are assessing) to a discussion about what we are assessing.

Constructs reflect the domain of interest and indicate progress in the domain. As such, constructs are the basis upon which scoring criteria can be developed. Grade criteria or level descriptors are operationalisations of the construct. Three distinct approaches to developing constructs can be found in the literature, although they are often blended in practice:

1. Theory-based: constructs are formulated on the basis of a theory, such as Piagetian notions of the development of scientific ideas
2. Empirically-driven: constructs are devised from the results of students on previous tests
3. Views of subject-matter experts: constructs represent disciplinary and/or assessment experts' views on what counts as progression in a particular discipline.

Educational assessments of science are almost always conducted through oral or written language, as in most subjects. We know that there are language effects on question comprehension in surveys in general (Graesser, Bommareddy, Swamer, & Golding, 1996; Graesser, Cai, Louwerse, & Daniel, 2006; Lenzner, Kaczmirek, & Lenzner, 2010) and also effects of language on specific subject matter, including science (Ahmed & Pollitt, 1999; Bramley, Hughes, Fisher-Hoch, & Pollitt, 1998; Eason, Goldberg, Young, Geist, & Cutting, 2012; Graesser, Jeon, Yang, & Cai, 2007). The use of 'academic language' has received considerable attention in research on student achievement (e.g. Frantz, Bailey, Starr, & Perea, 2014; Haneda, 2014; Snow & Ucelli, 2009; Stroud & Kerfoot, 2013) with academic language conceptualised in various ways, such as 'the language of education' (Halliday, 1994), 'the language of school(ling)', 'advanced literacy' (Schleppegrell & Colombi, 2002), 'scientific language' (Halliday & Martin, 1993), 'academic English' (Bailey, 2007; Scarcella, 2003), 'academic communication' (Haneda 2014), and 'formal language' (Graesser et al., 2014; Li et al., 2015). The linguistic and discourse features of academic and formal language have been investigated in some detail. For example, Graesser et al. (2014) have claimed that formal language tends to be in the informational genre (as opposed to oral language and

narrative) and cohesive, with complex syntax and abstract words. Cummins (1979, 1981, 2008) has distinguished between conversational everyday language which he refers to as 'basic interpersonal communicative skills' and the language required for school learning conceptualised as 'cognitive academic language proficiency', the latter being characterised by more complex grammar and sophisticated vocabulary.

The standard view is that language effects in assessments are irrelevant to the construct of the discipline being assessed; that is, they constitute construct-irrelevant variance (e.g. Wiliam, 2010, 261). They are a nuisance, or noise, in the data that cannot be eradicated as long as we have to assess constructs using language. Attempts to minimise language effects through careful writing of items has been advised (Ahmed & Pollitt, 1999; Bramley et al., 1998; Fisher-Hoch, Hughes, & Bramley, 1997; Pollitt, Ahmed, & Crisp, 2007; Pollitt, Entwistle, Hutchinson, & De Luca, 1985). The domain of interest for science assessments is typically defined by a curriculum, syllabus or sometimes more loosely by the discipline generally. Some educational assessments are not curriculum-related, so the domain is ill-defined in those cases, but in many cases there is little explicit definition of the domain and instead the textbooks and assessment materials come to define the domain. From the domain, a testing framework can be constructed which indicates how the domain will be operationalised within a test, what weighting is assigned to certain topics and so on. Counter to the standard interpretation above, it can be argued that language is an intrinsic part of being able to express and justify a scientific argument. Indeed, the high correlations between writing literacy and science literacy in PISA have been justified by the notion that to learn, children must acquire good language skills (Baumert, Lüdtke, Trautwein, & Brunner, 2009).

Language effects have, then, been viewed as construct-relevant, or construct-irrelevant. An alternative position, which we propose is that the method-effect of assessing through language might instead be viewed as integral to a 'composite' construct (Maul, 2013a). This way of viewing the construct is problematic for ILSAs. PISA science tests are designed to assess the construct of scientific literacy using different languages, so comparisons between countries only make sense if language effects are construed as construct-irrelevant method effects. Differential language effects between countries must be construct-irrelevant in PISA testing for the cross-country comparisons to be conceptually coherent. But to what extent are language effects separable from science constructs in international tests? To address this question we carefully examine the construct of scientific literacy as it has been portrayed in the literature and highlight the language dimension in the construct.

Scientific Literacy: The Language Dimension

Several definitions of scientific literacy have been proposed in the literature (Bybee, 1997; Bybee, McCrae, & Laurie, 2009; DeBoer, 2000; Millar & Osborne, 1998; NRC, 1996; Norman, 1998; NSES, 1996).

> Scientific literacy is the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity.
>
> (NSES, 1996, p. 22)

Most of the definitions of scientific literacy emphasise the interaction between science, technology and society (NRC, 1996; Osborne, 2007; Thomas & Durant, 1987) as well as the prime role of understanding the nature of science (Holbrook & Rannikmae, 2007), that is, science being empirical, tentative, value-laden, etc. (Lederman, 1992). More recent definitions have been framed within the emphasis on twenty-first century skills, highlighting the importance of critical thinking, formulating hypotheses and drawing evidence-based conclusions (Osborne, 2013; Pellegrino & Hilton, 2015). Scientific literacy is hence a complex construct that describes the scientific knowledge, critical thinking and processing skills students should acquire to become pro-active citizens, able to evaluate and use scientific information to engage with and reflect on science-related issues in a world largely shaped by technology.

While familiarity with scientific language is central to understanding and engaging with scientific knowledge, only a few definitions of scientific literacy refer explicitly to its importance (e.g. Millar & Osborne, 1998; Thomas & Durant, 1987). From a Systemic Functional Linguistics (SFL) perspective (*cf.* Halliday, 1978; Hasan & Martin, 1989), language in science serves a specific function – that of construing specific knowledge and beliefs (Halliday & Martin, 1993). The specialised grammar achieves specific functions such as precision, conciseness and clarity. Failing to cope with scientific language may have detrimental effects on leaning and performing in science (Norris & Phillips, 2003; Wellington & Osborne, 2001). Norris and Phillips (2003, p.226) argue that reading and writing in science are not simply functional skills that allow scientists to do science; reading and writing are 'constitutive' of science. Acquiring a high proficiency in reading and writing in science fosters attainment in science. Engaging with oral and written arguments, researching

information from written material, writing reports, publishing articles and communicating data in the form of tables, graphs, drawings, diagrams are at the heart of what scientists do and hence how students should learn science (Anderson, 1999; Gaskins et al., 1994; Yore & Treagust, 2006).

Norris and Phillips (2003) assert that literacy in its literal meaning refers to reading and writing and that scientific literacy in its 'fundamental sense' refers to reading, writing and oral communication in science. They claim that reading, writing and orally communicating in science (i.e. the fundamental sense) reinforces the understanding of scientific knowledge and the development of scientific reasoning (what the authors refer to as the derived sense of scientific literacy) and vice versa; a relationship described as symbiotic by Gee (2000). Language, discourse and rhetorical structure are thus considered an integral part of scientific literacy.

PISA's move to assess scientific literacy came amidst a worldwide trend to emphasise the concept of scientific literacy in science curricula (Holbrook & Rannikmae, 2007). This move was driven by an argument that the science curriculum should aim to produce more scientifically literate citizens that are able to solve real-life problems and be engaged in science-related issues. Scientific literacy has quickly become the overarching framework for designing science curricula (cf. Bybee 1997; DeBoer 2000; Bybee et al. 2009). For instance in England, Millar and Osborne's (1998) *Beyond 2000* report led to the implementation of a national science education curriculum in 2003, '*Twenty-First Century Science*', emphasising scientific literacy for fifteen- and sixteen-year old students. In the next section, we examine the way scientific literacy has been defined by OECD and discuss the extent to which the role of reading and writing, i.e. the fundamental sense of scientific literacy (Norris & Phillips, 2003), has been subsumed into the definition.

The Construct of Scientific Literacy in PISA

In both TIMSS and PISA, science constructs are based on frameworks developed by groups of subject experts (e.g. Jones, Wheeler, & Centurino, 2013; OECD, 2006). Unlike TIMSS, OECD claims that the PISA survey is curriculum-independent and is focused on scientific literacy rather than on scientific content (Bybee & McCrae, 2011). That is, instead of assessing students' understanding of scientific knowledge at age 15, PISA measures the

extent to which fifteen-year olds can apply knowledge to everyday situations as well as their readiness to become engaged and reflective citizens (OECD, 1999, 2016). The most recent definition of scientific literacy by OECD is the one adopted in PISA 2015 (OECD, 2016). The first part of the definition describes what OECD considers to be scientific literacy and what a scientifically literate person is able to demonstrate while the second part operationalises the overarching construct of the science assessment of PISA 2015, scientific literacy, into three major competencies, as follows,

> Scientific literacy is the ability to engage with science-related issues, and with the ideas of science as a reflective citizen.
>
> A scientifically literate person, therefore, is willing to engage in reasoned discourse about science and technology which requires the competencies to:
>
> 1. Explain phenomena scientifically – recognise, offer and evaluate explanations for a range of natural and technological phenomena.
> 2. Evaluate and design scientific enquiry – describe and appraise scientific investigations and propose ways of addressing questions scientifically.
> 3. Interpret data and evidence scientifically – analyse and evaluate data, claims and arguments in a variety of representations and draw appropriate scientific conclusions.
>
> (OECD, 2016)

The definition above does not mention language or the relationship between scientific literacy and linguistic proficiency, as there is no explicit mention of language. However, a scientific literacy construct measuring 'engagement in scientific discourse' and the ability to 'explain', 'evaluate' and 'interpret' scientific enquiry and data cannot but presume a high language proficiency. In the next section, we examine PISA science items and scoring guides more closely, highlighting the language demands involved in reading PISA science items and formulating correct responses. We also analyse science proficiency levels as described by PISA and underline the strong relationship between science attainment in PISA and language proficiency.

Language Proficiency and PISA Science Assessments

Examining the language demands on participants when reading PISA science items and producing written responses provides a good indication of the extent to which scoring highly in PISA science tests entails high language proficiency. To provide such evidence, we gauge

the reading load in PISA science items and explore examples of scoring guides and model responses to evaluate the writing skills participants are expected to possess to provide correct responses.

The reading demands of a released PISA item were analysed using the Pearson Reading Maturity Metric (RMM)[1] (Landauer, Kireyev, & Panaccione, 2011; Landauer, 2012). RMM uses computational language models to estimate a measure referred to as 'Word Maturity'. Word Maturity depicts how much experience is needed to reach an adult's understanding of words, sentences and texts. The Word Maturity metric is highly correlated with well-established vocabulary tests ($\rho \sim 0.8$) such as Kaufman Brief Intelligence Test – II[2] (Landauer et al., 2011). Results were compared to the Pearson RMM results of similar analyses carried out on a science item taken from TIMSS administered to a similar age group. The writing demands imposed by PISA and TIMSS science assessments were compared in terms of length and expected complexity of responses. The next section describes the results of these comparisons.

Reading demands of PISA items

PISA items are contextually embedded items, organised within units, which are sets of items referring to a common stimulus and scored independently (OECD 2009b). A stimulus often consists of a text providing context for a science issue as well as a photograph, a figure, a graph, a table, or some other non-textual medium. Appendix 1 presents the stimulus of a released PISA 2006 science unit and one of its items.

By way of contrast Appendix 2 presents a science item from TIMSS 2003 administered to a comparable age group (Grade 8; 14-year-olds) and addressing the same general topic as the PISA item shown above (i.e. health, immune system). We observe how the stimulus is much shorter and much less dense.

Table 1 summarises the comparison of the RMM data between the two items.

[1] Software available online free: http://www.readingmaturity.com/rmm-web/main#/passage/28156.

[2] Kaufman, A. S. and Kaufman, N. L. (1990) Kaufman Brief Intelligence Test—Manual (American Guidance Service, Circle Pines, MN).

Table 1: Comparing language density in PISA and TIMSS items administered to comparable age groups

	Stimulus		Question			
	No. sentences	**No. words**	**No. sentences**	**No. words**	**RMM score**	**Target year (age)**
PISA	17	259	2	23	9.9	Y9-10 (15-16)
TIMSS	2	17	1	30	3.1	K-1 (5-6)

Key: RMM = Pearson Reading Maturity Index

Excluding the titles and the textual material accompanying the graph, the PISA stimulus (Appendix 1) consisted of 17 sentences comprising 259 words and the question comprised 2 sentences with 23 words. The relatively high number of nouns per sentence in PISA is likely to make the text harder to read as higher lexical density renders a text more challenging (Halliday, 1985). Indeed, an analysis of the readability level of the PISA text using the RMM software revealed an overall score of 9.9 suggesting that the text is aimed at Year 9 to 10 students, that is, students of age 15-16. In other words, an average 15-year old student should not find this text challenging. This corresponds to the age group of students participating in PISA. However, to eliminate the possible effect of language on test performance, the text should target a slightly younger age group (e.g. 12-13 year olds). Technical words such as 'puerperal', 'hygiene' and 'extraterrestrial' were highlighted by RMM as being complex. The terms 'puerperal' and 'extraterrestrial' do not reflect any fundamental scientific concept with which 15-year olds would be expected to be familiar. Omitting them or replacing them by more familiar terms would have not altered the construct measured and would have certainly reduced the language demands.

In the TIMSS example (Appendix 2), the context is explained in two sentences (17 words) and the question is only one-sentence long, comprising 30 words. A readability test using RMM resulted in a score of 3.1, suggesting the text is at a level for students in Kindergarten to Year 1 (K-1) aged 5 to 6! The word 'cold' was highlighted as relatively difficult vocabulary for that age group and unlikely to be true for 14 or 15 year olds.

The above analysis was undertaken on only one item to illustrate the different item styles adopted in TIMSS and PISA. The language in the TIMSS items appears to be considerably simplified to minimize construct-irrelevant variance whereas PISA science items clearly impose higher processing demands on participants through, for example, text length and use of graphs. Although the readability level of the PISA item does not exceed the average readability level of the participating age group, it is high enough to potentially cause construct-irrelevant variance. The higher lexical density in PISA and the use of low frequency terms are likely to increase reading demands and therefore affect performance on the science construct.

Language demands are not only experienced when respondents read items, they are also manifested when respondents produce extended responses (see Question Answering Process by Pollitt & Ahmed, 1999). The following section highlights the language demands associated with writing responses in PISA and TIMSS by comparing the types of responses expected in each of the two assessments.

Writing demands in open-ended responses

The scoring of the PISA and TIMSS items discussed in the previous section is now considered in this section. Table 2 below presents the marking scheme for each of the two items.

Table 2: Examples of marking schemes for PISA[3] and TIMSS[4] items

	PISA	TIMSS
Item code	S195Q02	S022154
Item title	Semmelweis Diary	Transmission of Cold in Classroom
Full credit	"Refers to the difference between the numbers of deaths per 100 deliveries in both wards, due to the fact that the first ward had a high rate of women dying compared to women in the second ward, obviously show that it had nothing to do with earthquakes. Not as many people died in ward 2 so an earthquake could not have occurred without causing the same number of deaths in each ward"	"Some students were hanging around Salil [sic] with him sneezing his germs onto them. The ones exposed to the virus caught it."
Partial credit	"It would be unlikely to be caused by earthquakes because earthquakes don't occur frequently"	*No partial credit*
No credit	Earthquakes could not cause the fever"	"Some of his classmates did not like him so probably were not near him a lot."

In the PISA item in order to earn full credit, and in addition to the reading load, participants were required to:

- formulate longer structured responses
- describe and compare observations between the two wards
- articulate clear cause and effect relation
- draw a conclusion based on information in the stimulus material (text and graphs).

This genre of writing in science, i.e. explaining and justifying with evidence, is more demanding than the descriptive genre (i.e. retelling facts) which characterises recall responses (Snow, 2010; Yore & Treagust, 2006). Mahboob's (*in press*) 3D Framework of Language Variation classifies language registers along three continua: (1) user (i.e. audience), (2) mode and (3) use. Users can be placed on a continuum with local/low social distance on one extreme to global/high distance on the other. The mode may be oral or written and the use of

[3] See the following website for released TIMSS Grade 8 items and scoring https://nces.ed.gov/TIMSS/educators.asp

[4] See the following website for released PISA items and scoring.

language may vary from everyday/casual to specialised/technical discourse. According to this analysis, PISA responses may be described as addressing a global audience, as technical and formulated in the written mode. The 'global, written, specialised' domain is one typically found in academic papers and presents the highest comprehension challenges of all other domains (Mahboob, in press).

In TIMSS, expected responses are considerably shorter than those in PISA (see Table 2). Although the TIMSS question (Appendix 2) also requires the respondent to infer a cause, the expected correct response is less structured with use of a more informal genre of writing (e.g. the use of the colloquial expression 'hanging around' instead of 'spending time with' or choice of the term 'caught the virus' instead of the scientific term 'contaminated with the virus'). Using Mahboob's (in press) classification, responses in TIMSS may be said to be in the local, everyday written discourse domain, typical of that used in e-mails between friends. Applying the Cummins' classification (1979, 1981, 2008), it can be argued that Cognitive Academic Language Proficiency is a pre-requisite for writing a full credit response in PISA while conversational language (Basic Interpersonal Communication Skills) is sufficient to gain full marks in TIMSS.

Although most PISA science items are multiple choice questions, a substantial proportion of them (a third) consists of open-ended questions (OECD 2009a; 2009b) and hence require the type of responses described above. In the draft of its most recent scientific framework, OECD has explicitly recognised the reading and writing demands that PISA tests incur and has claimed that language demands will be kept to a minimum for its PISA 2015 cycle (see OECD, 2016, p. 41).

Such statements need to be supported by empirical evidence and such claims verified, i.e. whether OECD's wish to neutralise language effects in science tests are in fact successful. In our view, language effects in international tests are beyond the control of the test designer, as language proficiency is an intricate component woven into the constructs, in this case science constructs. Having presented evidence for high reading and writing demands using PISA and TIMSS items, we next support our argument by highlighting the central role of language proficiency in attaining high proficiency levels in PISA science.

The role of language in PISA science proficiency levels

In PISA science, scores represent six proficiency levels of scientific literacy, which include being able to explain, justify and develop arguments (*cf.* OECD, 2016). One may also note that increasing performance in science requires higher proficiency in the ability to understand scientific language, communicate scientific thoughts, build arguments, interpret data, etc. (Anderson, 1999; Yore & Treagust, 2006; Gaskins et al., 1994; Snow, 2010; Snow & Ucelli, 2009). The increase in language proficiency can be clearly observed when mapping the progression of scientific literacy levels with the language proficiency expected from the lowest (i.e. Proficiency Level 1b) to the highest proficiency level.

For instance, at the lowest proficiency level (Proficiency Level 1b), students are incapable of adequately using scientific content to *formulate explanations* and *interpret* data and are unsuccessful at *describing* simple causal relationships (OECD, 2016). At an intermediate level (Level 3), students are able to use scientific knowledge and data to *explain* facts, *interpret* data and *make inferences* in unambiguous contexts. They are capable of *describing* simple causal relationships. They show evidence of *critical thinking* and are capable of *building partial arguments*. At the highest proficiency level (Level 6), students are capable of *articulating explanations* of complex scientific phenomena and causal relationships (OECD, 2016). They can *interpret* data from a variety of sources and *draw appropriate inferences*. They can *build and formulate* strong arguments and *critique* scientific models, experimental designs and conclusions.

Having language proficiency clearly manifested in PISA science proficiency levels supports the argument that language cannot be assumed to be peripheral to science constructs and hence neutral in terms of its impact on the demands and difficulty of a science item. Language is a construct-relevant variable in PISA science assessments. This conclusion has significant implications for the interpretation of PISA results especially when it comes to comparing results across countries that administer different language versions of the same test. While it may be argued that differential impact of language on the level of difficulty of a question can be detected statistically, we discuss below how the techniques employed in international tests can often miss substantial bias across countries.

Statistical Identification of Language Bias

International tests operate within a psychometrics paradigm that makes powerful claims for the measurement of constructs, but has a number of assumptions that have practical consequences (Baird & Black, 2013). One such assumption is that the test is unidimensional, i.e. measuring a single, indivisible construct. Multi-dimensional item response theory (IRT) is a practicality, but is not the underlying basis for international tests because the aim of international assessment is to compare the performances of countries on a single score associated with a unidimensional construct (e.g. mathematics literacy), not to be able to claim that the French are better at geometry and the Spanish are better at algebra. The possibility that the construct of science literacy is composed of science ability *and* language ability violates the assumption of unidimensionality on which statistical models (i.e. IRT models) used to build and analyse international assessments rest. Moreover, having a multidimensional construct in PISA challenges the conceptual underpinnings and utility value of these tests.

Statistical tests for unidimensionality are conducted by way of fit to the Rasch model and Principal Components Analysis (OECD, 2009a). However, rarely is the alternative of multi-dimensionality properly considered; an exception being the PISA reading literacy tests have been found to contain at least two additional dimensions (age group and gender) in 2000 (Goldstein et al., 2007). A question therefore arises as to whether the data from scientific literacy tests collected across different languages could fit a unidimensional model even if the languages themselves cause differential difficulty. Unidimensionality is a matter of degree for all tests (Andrich, 2014) and it has recently been proposed that constructs can be a composite of related variables, such that they are, to a matter of degree, contemporaneously unidimensional and multidimensional (Maul, 2013b). In this case, scientific literacy can be viewed as unidimensional, but also comprising variables relating to different aspects of science content, reasoning and language.

The standard technique used to detect bias in items is *differential item functioning* (DIF; Dorans, 1989; Holland & Thayer, 1988; Holland & Wainer, 1993) using the total score on the test as a control for ability on the construct and then seeking to identify whether there are group differences in performance on individual items, given the total score. In the case outlined in this article, total score represents the scientific literacy of students and it is possible, using DIF, to explore whether individual items are easier or more difficult for

students who experienced different translations of the test, such as England versus France. Of course, we do not know that it is language that causes any such differences detected, as the cause could equally be a cultural variable, or indeed a curriculum or teaching factor. Nonetheless, DIF can be used as an indicator and is used to investigate group differences, such as gender or ethnicity effects, as a standard procedure in the assessment industry (Cole & Zieky, 2001), and it is also used to detect problematical items in PISA. A number of items, referred to colloquially as 'dodgy items', are discarded on the basis of DIF analyses across countries (OECD, 2009a). Analyses of PISA 2006 scientific literacy indicated that a single dimension best fit the resulting test (Andrich & El Masri, *in preparation*), *in preparation*), although a number of items exhibited DIF by country. Indeed, despite rigorous translation and adaptation techniques followed in PISA (OECD, 2009a; OECD, 2009b) and TIMSS (Yu & Ebbs, 2012), empirical studies detected differential functioning of items, with language being one of the factors contributing to the bias (Grisay & Monseur, 2007; Grisay et al., 2009; Oliveri & Ercikan, 2011; Ercikan & Koh, 2005).

DIF could be caused by a number of explanatory variables that coincide with country. The standard way of dealing with this is to ensure that the test overall does not have too many items favouring a particular group. If a test is biased consistently in favour of one language, DIF will not detect this and it will appear that all is well. Consistent bias in a specific direction could be due to poor translation quality where all items in one version are systematically biased. In the following section, we outline the various sources of bias associated with test translation and highlight how language effects are unavoidable.

Translation and Adaptation of International Tests

Translating source versions of international assessments into other languages and adapting their content to suit cultural sensitivities of target populations can give rise to biases which could be unavoidable and statistical techniques may fail to expose them. Assessments are well translated and adapted, i.e. transadapted, to another group of participants if both linguistic and psychological characteristics of the source text are retained in the target version (Hambleton, 2005; van de Vijver & Poortinga, 2005).

The International Test Commission (ITC) has produced guidelines for sound test translation and adaptation (Hambleton, 2001; ITC, 2005) and these are followed to a great extent in

international tests. For instance, PISA produces two source versions of the surveys in English and French (Grisay et al., 2007; OECD, 2009b). It then performs a double translation of the source versions followed by a reconciliation procedure. This method consists of independently translating the source versions (English and French in this case) into two versions in the target language and then merging them to produce a final version in the target language. This method is claimed to ensure that any idiosyncrasies embedded within a particular language are detected during the transadaptation process and would thus have less impact on the development of target versions (OECD, 2009b).

Despite the rigorous methods of test transadaptation followed by international assessments, Hambleton et al. (2005) contend that bias in test transadaptation can appear at various stages of test development, translation or result interpretation. Bias can even be inherent to the design or methods used to produce the tests. For instance, the selection of competent translators and appropriate transadaptation designs could be very critical. Bias can emerge even with meticulously chosen translators and with carefully designed translation models as we describe below.

Selection of translators

The selection of translators can introduce bias in adapted versions of a test. It is not enough to find qualified translators who are equally proficient in both the source and the target languages. Translators should also have a background in the subject translated with knowledge of test and scale construction (Hambleton, 2005). Otherwise, they may unintentionally introduce elements that make the test easier or more difficult in the adapted version. The example in Appendix 3 illustrates this point: one of the Arabic translations of this item gives inadvertent clues to the correct answer.

Appendix 3 presents the English version of a PISA science 2006 item (S508Q03, Genetically modified crops unit). It consisted of a simple multiple choice question with D as the key (correct answer) and A, B, C as distractors (wrong alternatives). The item was written using a less formal register than typical scientific texts, i.e. limited use of scientific terms. The term genetically modified crops and the acronym GM had been defined in the text preceding the item and, thus, the level of formality of the item would be unlikely to pose comprehension problems to fifteen-year old students.

The adaptation of this item into Arabic was not problematic except for the key (i.e. option D). Figure 1 below presents the adaptation of the key from the English and French source versions into Arabic.

Figure 1: Adaptation of PISA item S508Q03 key into Arabic (age 15)

English	*To include various growth conditions for corn.*
French	*Pour inclure diverses conditions de culture du maïs.*
Arabic	بهدف فحص تأثير الظروف البئية المختلفة على نمو الذرة.

The literal back-translation of the Arabic version of the key would be: '*In the aim of testing the effect of different environmental conditions on the growth of corn*' shown in Figure 1 in Arabic script. Unlike the English and French source versions, the Arabic back-translation consisted of a long nominal phrase highly dense in information.

The differential language complexity between the key and the distractors in the Arabic version was not observed in either of the source versions. The key could have been translated more simply to match the level of formality in the English and French source versions such as: 'لشمل ظروف نمو مختلفة للذرة', which is a direct translation of the source version in English 'To include various growth conditions for corn'.

The poor translation could potentially lead to bias in the item. In this case, it is debatable whether such bias would eventually favour or disadvantage Arabic-speaking participants. At first glance, one would think that by increasing the syntactic complexity and vocabulary density of the key and hence increasing its reading demands, the item would become more difficult in the country administering this Arabic version. This remains a speculation that needs to be confirmed empirically. The item could well have been made easier because of the use of the scientific language in the key only and not the distractors making the key look like the smart option.

Bias resulting from poor choice of translators and low quality in test transadaptation is far from being the only case. While it can be argued that adopting a rigorous model of test transadaptation can limit bias in international tests, we contend that meticulous transadaptation can only minimize potential bias. Indeed, as we describe more thoroughly in

the next section, bias in transadapted tests can arise from factors that are inherent in the process of transadaptation and is hence inevitable at some level.

Transadaptation models and judgement designs

Transadaptation of tests can be carried out by adopting one of two main models: the forward and the backward (or back-) translation models (Hambleton, 2005). The forward model consists of directly translating a test from the source language to the target language while the back-translation model consists of translating the test from the source language to the target language and then back from the target language to the source language. According to Hambleton, each of these models presents weaknesses that could constitute a source of bias towards a particular language group.

While forward translation enables the judgement of the quality of the transadaptation by directly comparing the source and target versions of the test, the process requires bilinguals who may not be equally proficient in both languages. The back-translation option does not involve bilinguals, yet it still presents two main drawbacks. First, the model does not allow direct comparison between source and target versions of the test. Moreover, only versions in the source language are compared and hence any inaccurate translations of idiosyncratic aspects of the source language into the target language would not be detected and would thus be retained.

In addition to issues emerging from the translation model adopted, the design of studies judging item or test equivalence across languages may often miss the bias inherent in the test or even give rise to it (Hambleton, 2005). These designs involve either bilingual or monolingual examinees and hence present their own weaknesses.

Involving bilinguals in judging the equivalence of the source and target version of a test is problematic for two reasons. First, this design assumes that bilingual examinees are equally proficient in both languages despite research evidence pointing to the opposite (Sireci, 1997). Second, data from this design are not representative of any of the monolingual populations because bilingual examinees tend to be, on average, substantially different from their monolingual counterparts (Hambleton, 1993). This limits the extent to which findings can be generalised to establish the equivalence of an assessment across two, very different, language populations.

The shortcomings associated with using bilingual examinees can be overridden by employing a monolingual design instead. This would consist of administering the original version of the test to a source language monolingual sample and administering the adapted version to a target language monolingual sample. In this design, the two monolingual samples would be representative of both the source and target populations and findings would eventually be generalisable to both the source and target populations. However, this design makes use of samples of populations which differ to a greater or lesser extent from each other (e.g. culture, educational background, socioeconomic status, etc.) making direct comparisons of translation quality problematic. For instance, it is likely that students varying on the aforementioned characteristics are not equally proficient in academic language (Snow & Ucelli 2009). This means that one group will inherently be at an advantage when evaluating the linguistic challenges of a particular language form of the test.

In summary, irrespective of the transadaptation model employed and the judgement design adopted, language bias is inevitable in test transadaptation (Hambleton, 2005; van de Vijver & Poortinga, 2005). Quality control scrutiny may fail to identify this type of bias which can be consistent across all items in a particular version and hence escape detection by statistical techniques. Another similar, yet in our view more challenging, source of bias is language idiosyncrasies such as metaphors which prove to be hard to translate into different languages without introducing bias (Spielberger, Moscoso, & Brunner, 2005). The following section introduces a few of these idiosyncrasies and provides examples of items in which they are manifested.

Untranslatable Language Idiosyncrasies

Language idiosyncrasies are countless and often stem from particular syntactic and grammatical properties governing them. Language has been shown to be fundamentally distinct on every level (*cf.* Evans & Levinson, 2009), with some languages being more complex and demanding than others (e.g. Lupyan & Dale, 2010; Trudgill, 2011). If so, then matching the level of complexity of questions in different languages constitutes a difficult challenge that will never be perfectly surmounted. Indeed, the little underlying commonality across languages makes non-biased translation even more difficult to achieve. This bias is not due to lack of rigour in the transadaptation process but emerges rather from dealing with a complex variable like language and attempting to translate the ‘untranslatable’. In this

section, we discuss four different idiosyncrasies – test length, word frequency, abbreviation incongruence and semiotic incongruence – by analysing science items including ones administered in PISA to elucidate our argument.

Test length and speededness effect

Some languages require more words than other languages to express the same meaning (Shala, Rus, & Graesser, 2010). For instance, Shala et al. (2014) found that the word count per sentence of English translations of over 300 speeches of the Former President of Egypt Hosni Mubarak were consistently 10 words shorter than the original Arabic speeches. Sentence length has long been known to be a major component of text grade levels in readability measures such as Flesch-Kincaid grade level (Klare, 1974), presumably because it places a greater burden on working memory (Graesser & McNamara, 2011). The difference in average sentence length may lead to variation in test length across different language versions. Indeed, Eivers (2010) compared three versions of PISA with the English source version and found that the Finish versions were on average 8% longer than the English versions while the Irish versions were 11% longer and the German versions were 17% longer. Longer tests require more time to read; however, all countries are given exactly the same time (two hours) to complete the test. This suggests that the German versions are more prone to being affected by speed constraints than do the Irish, or the Finnish or the English versions. In other words, assuming the same reading speed, participants sitting a German version of PISA will have less time to read and respond to each question than their counterparts completing the Irish, or Finnish or English versions.

Word frequency

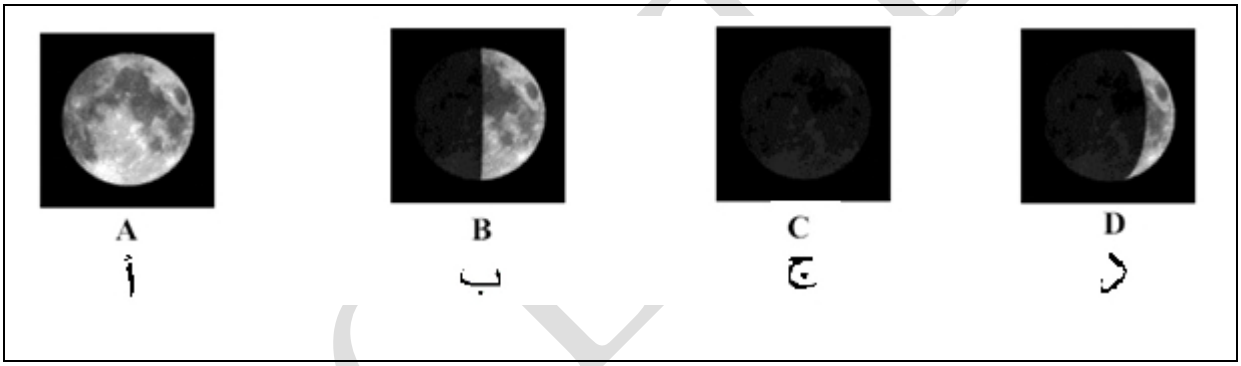
Another type of language idiosyncrasy constituting a possible source of bias is word frequency. Word frequency, age of word acquisition, and word length are robust predictors of text complexity and readability metrics (Graesser et al., 2014; Graesser, McNamara, & Kulikowich, 2011; Klare, 1974; Landauer et al., 2011). Familiar or high frequency words in the source language may be translated into extremely infrequent terms in the target version (or vice versa) notwithstanding high translation quality. Low frequency words might be so unfamiliar to test takers that this impedes their comprehension of the question (Graesser et al., 2006; Lenzner et al., 2010). This could result in much higher language demands across

language groups. To illustrate how word frequency for the same concept could vary across languages and hence result in differential item demands for different language groups, we generated an exemplar test item in three languages related to the lunar cycle (Figure 2).

The lunar cycle is typically taught in Year 6 or 7. Figure 2 depicts the lunar phases in English, French and Arabic on the basis of which an exemplar test item is produced in the three languages. In the following, we analyse the questions and compare the demands associated with each language. The authors deliberately devised simple questions that assess factual knowledge to compare the differential language demands associated with a specific variable, word frequency, across languages. Questions targeting more sophisticated aspects of scientific literacy are likely to introduce additional variables such as concept complexity which may interact with word frequency in unpredictable ways across languages.

In each language version of the exemplar item, examinees are instructed to complete the blanks with the letter corresponding to the correct lunar phase. When analysing the linguistic demands of each question in each language, the stem of the question is constant; that is, in English for instance, the stem '*Which one of the photos above represents the'* is the same in the three questions. This is the case in French and Arabic. Hence, it follows that any difference in the level of demands between questions within each language is determined by the demands imposed by the frequency of words in the lunar phase. For the sake of the argument, we will also assume that the stems are of equivalent difficulties across languages. This means that the variation in question demands across languages is also determined by the difference in word frequency of the lunar cycle across the languages. Let us take each language in turn and consider those demands within and across languages.

Figure 2: Lunar phases and exemplar test items in different languages (age 12-13)



English

1) Which one of the photos above represents the ***new moon***? ___
2) Which one of the photos above represents the ***full moon***? ___
3) Which one of the photos above represents the ***waxing crescent***? ___

French

1) Parmi les photos ci-dessus, laquelle représente la ***nouvelle lune*** ? ___
2) Parmi les photos ci-dessus, laquelle représente la ***pleine lune*** ? ___
3) Parmi les photos ci-dessus, laquelle représente le ***premier croissant***? ___

Arabic

١- أي من الصور أعلاه يمثل ***المحاق*** ؟ ____ [almahaq]

٢- أي من الصور أعلاه يمثل ***البدر*** ؟ ____ [albadr]

٣- أي من الصور أعلاه يمثل ***الهلال المتزايد*** ؟ ____ [alhilal almoutazayed]

English

In questions 1 and 2, the terms 'new', 'full' and 'moon' are high frequency words in English[5]. Hence, the first two questions are easy to process and examinees are very likely to choose letter C for item 1 (moon with no light) and letter A for item 2 (moon 'full' of light). Examinees might hesitate between letters C and D for new moon but the term 'new' will allow them to eliminate A and B as possibilities. However, question 3 seems more complicated as it includes two less familiar terms. The word 'waxing' is associated with 'adding wax' and the word 'crescent' is not a common shape like triangle and square. So this question in English is likely to impose greater linguistic demands than the other two phases of the moon. Examinees might not be able to easily eliminate choices especially if this question was not asked in conjunction with questions 1 and 2.

French

Similar to English, the terms *'nouvelle'*, *'pleine'* and *'lune'* are high frequency words in French[6] and are not expected to impose unnecessary language demands on examinees who would very likely be able to answer questions 1 and 2. Unlike in English, question 3 has also high frequency words. The term *'premier'* (meaning first in English) is very common in French and the term *'croissant'* relates to the famous French pastry which has a crescent shape and therefore examinees would easily associate the shape of the 'premier croissant' (i.e. waxing crescent) with the shape of the French pastry. Hence, in French none of the questions seem particularly demanding or imposing linguistic or cognitive challenges.

Arabic

Arabic is a diglossic language; i.e. the spoken language 'العامية' [*ala'mmiya*] is quite different from the literary one, referred to as 'الفصحى' [*al fuss-ha*] or the modern standard Arabic[7]. Modern standard Arabic is the language taught in schools and is used in science classrooms and assessments when the medium of science instruction is Arabic. Hence, the extent to

[5] http://ucrel.lancs.ac.uk/bncfreq/lists/5_3_all_rank_adjective.txt
[6] http://eduscol.education.fr/cid47915/liste-des-mots-classee-par-ordre-alphabetique.html
[7] Modern standard Arabic is relatively invariant across Arabic countries, spoken Arabic consists of different dialects which vary, sometimes dramatically, across Arabic countries. In the following, the analysis of the questions makes reference to the Levantine dialect; that is, Arabic spoken in Syria, Lebanon, Palestine and Jordan.

which words are familiar in Arabic is also related to how similar the word in standard Arabic is to the one used in the colloquial language.

In question 1, the term 'المحاق' [*almahaq*] (meaning 'new moon' in English) is not commonly used in spoken Arabic. Indeed this word is a scientific term in Arabic[8] and is only likely to be encountered in a context discussing the phases of the lunar cycle. Question 2 includes the term 'البدر' [*albadr*] (meaning 'full moon' in English). This term is quite familiar in Arabic cultures and the word is the same in standard and spoken Arabic. Often, the beauty of a woman is compared to the full moon to imply the brightness and splendour of her face. Hence, question 2 should not entail any particular linguistic demands in Arabic. In question 3, the terms 'الهلال' [*alhilal*] (meaning 'crescent' in English) and 'المتزايد' [*almoutazyed*] (meaning 'increasing' in English) are familiar terms. The term [*alhilal*] is the same in standard and spoken Arabic. It means 'crescent' which is a familiar shape in Arabic-speaking countries, as it is the symbol of Islam, and Arabic-speaking students are likely to be in Muslim majority countries. The term [*almoutazayed*] is similar in both standard and spoken Arabic, meaning 'increasing' and also has high frequency. Hence, in Arabic, it is hypothesised that question 3 does not impose any particular linguistic and cognitive demands.

Although the above discussion refers to a hypothetical example, it is the principle of the effects of word frequency which matter, not the specifics of the example itself. To summarise the findings of the above analysis, it is suggested that item 3 was the most linguistically demanding in English; item 1 the most demanding in Arabic while none of the items seemed to be particularly linguistically demanding in French. Within the same language, the level of demands of a question can vary with the level of frequency of a particular word referring to a given concept (Graesser et al. 2006; Lenzner et al. 2010). This is also true across languages where the translation of the terms from English to French and Arabic leads to different word frequencies in the target languages making some questions inherently less demanding and others intrinsically more challenging linguistically. Wider cultural connotations and use of these words have affected their frequency of usage. Furthermore, the words in the question are specific technical terms with scientific meanings that cannot easily be exchanged for lower or higher frequency terms in the target languages.

[8] http://www.bibalex.org/ica/ar/

The example above represents a case of *identity of concepts* (Poortinga, 1995) where a particular term in a source language, (e.g. waxing crescent in English) matches exactly the meaning of the term used in the translated version (e.g. premier croissant in French). Wiliam (2008) provides a similar example where translations of the English terms 'velocity' and 'speed' into Welsh could advantage one language group over another in a physics assessment. In English, the term 'speed' designates the scalar quantity and is the term used in everyday language. This is not the case in Welsh where the term *cyflymfer* meaning velocity (i.e. the vector quantity) is the vernacular one. Wiliam explains that the issue in this case does not reside in the quality of the translation but rather in what these terms connote in each language.

In this subsection, we discussed how transadaptation issues can be manifested through differential word frequencies across languages. Another case of untranslatable idiosyncrasy that may introduce bias towards or against a particular language group is what we refer to as abbreviation incongruence. Abbreviation incongruence can be observed when translating scientific acronyms and abbreviations, such as the symbol of chemical elements and chemical formulae, which abide by international conventions and make use of Latin characters. We argue that the use of Latin characters in languages that employ a different script (e.g. Arabic, Chinese, Hindi, Japanese, etc.) may place additional cognitive demands for these languages.

Abbreviation incongruence

It might be anticipated that aspects of science, such as the periodic table are an international language. The periodic table is universal and makes use of Latin letters to symbolise the different elements (*H* for hydrogen, *C* for carbon, *O* for oxygen, etc.). Chemical formulae like CO_2 (carbon dioxide) and NaCl (sodium chloride) are also universal. Differential language demands can appear in the translation of the names of chemical elements and compounds especially when the translation occurs into languages that do not have a Latin script.

For instance, the English symbol for lead is 'Pb'. There is incongruence between the word 'lead' and the symbol of the chemical element 'Pb' which is an abbreviation of the Latin origin of the word 'Plumbum' meaning soft metals. Translating an item containing this symbol into French eliminates the incongruence because the French term for lead is *'plomb'*. This makes the item less demanding in French.

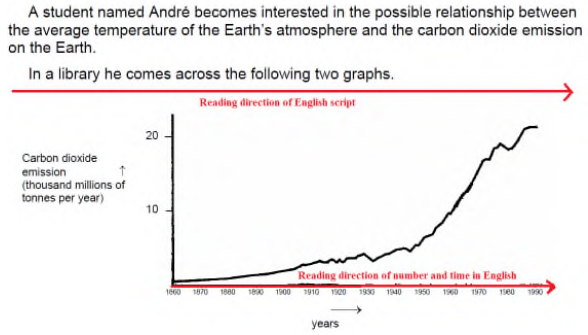
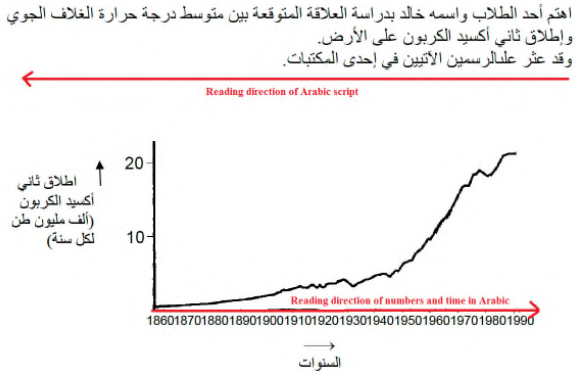
El Masri (2015) provided evidence of how the differential effect of translating the name of chemical compounds was likely to be an issue in some PISA 2006 science items. She discussed how one of the units (Unit S447 Sunscreens) involved the chemical compound zinc oxide which name in English matches with its chemical formula ZnO. The term in French is reversed for this compound, i.e. *'oxyde de zinc'*, but might only create very little additional demands if any. However, the Jordanian Arabic version uses the term 'أكسيد الخارصين' *[oxeed al kharaseen]* which on the one hand, is a different script to the universal chemical formula ZnO and on the other hand, does not include the sound /z/ but the guttural sound /x/. The incongruence between the name and the script of the chemical compound in Arabic and the Latin formula ZnO is likely to increase cognitive load for Arabic-speaking participants. Similar problems are expected to occur in Chinese and Japanese languages.

Another interesting case of incongruence that emerges in transadapted tests is what we termed as semiotic incongruence where reading charts, tables and graphs could create differential demands for different language groups. Theories of graph comprehension in cognitive psychology have been mainly centred on differences between novices and experts within the context of first language learning (e.g. Friel, Curcio, & Bright, 2001; Roth, 2002). Nevertheless, Carpenter and Shah (1998) highlighted the importance of specific characteristics of individuals making sense of the graphs, for instance their language background. In the example below, we analyse the English and Arabic versions of a PISA science item that includes a graph and highlight the additional cognitive demand that may be placed on the Arabic language group.

Semiotic incongruence

Translation of tests can also result in incongruence in the semiotics used. One example is in PISA 2006 (Unit S114) which includes questions requiring students to read graphs. Graphs are semiotic tools with which scientists communicate data and mastering them is a part of acquiring science literacy as discussed earlier in this paper. Reading graphs typically increases the cognitive demands of tasks (e.g. Yu, 2012); however, this activity is not equally demanding in all languages. For instance, in English and French, the time line moves from left to right and hence matches the direction of reading in both languages. In Arabic, this is not the case. Time lines and numbers are read from left to right while the language script is read from right to left (Figure 3).

Figure 3: English and Arabic versions of graphs in PISA unit S114 (age 15)

Excerpt of Unit S114 (English)	**Excerpt of Unit S114 (Arabic)**
	

Source: www.oecd.org/pisa/38709385.pdf.

This aspect of Arabic is not related to the quality of translation but is an integral part of the Arabic language. The additional cognitive demand that is placed upon examinees because of the incongruence between aspects of the items cannot be picked up by generic models that do not focus on the language variables in an international setting. This incongruence, not originally present in the source version, is likely to result in an additional cognitive load for Arabic examinees as similar resources are needed in both tasks, reading a graph and reading a text[9].

Computational Linguistics as a Way Forward

In the previous section, we discussed a number of language idiosyncrasies that are difficult to translate and that introduce bias against particular language groups even when transadaptation methods are rigorous. We contend that the language idiosyncrasies discussed earlier are likely to escape the standard quality control checks often exerted by expert judges (e.g. OECD, 2009b) and suggest here that advances in computational linguistics might provide one way to address these issues.

[9] El Masri (2015) provides several additional examples of how language idiosyncrasies such as differential familiarity with technical acronyms and differential interference between every day and scientific language which can introduce bias in adapted versions of tests.

Landmark advances in computational linguistics (Jurafsky & Martin, 2009), discourse science (McNamara, Graesser, McCarthy, & Cai, 2014), and statistical representations of world knowledge (Landauer, McNamara, Dennis, & Kintsch, 2007) have made it more feasible to develop automated methods of gauging differences between languages in educational assessment and also language translation. There are now several automated measures of text complexity that are used in education applications. In particular, text difficulty was addressed in the Common Core Standards for English Language Arts (www.corestandards.org) when the need for a systematic comparative study of automated text analysis tools was acknowledged, particularly with respect to text difficulty. Also, a systematic comparison study of seven text analysis tools was conducted on hundreds of texts in different discourse genres, including narrative (stories) and informational texts, e.g., social studies and science (Nelson, Perfetti, Liben, & Liben, 2011). Three of the tools have provided multiple dimensions of text complexity in addition to a single dimension: the *Text-Evaluator Tool* of Educational Testing Service (Sheehan, Kostin, Napolitano, & Flor, 2014), the *Pearson Reading Maturity Metric* of Pearson Knowledge Technologies (Landauer et al., 2011), and the *Coh-Metrix* system developed at University of Memphis (Graesser et al., 2014, 2011; McNamara et al., 2014; www.cohmetrix.com). These systems can immediately be used to automatically assess the complexity of questions on tests for any version in English. Meanwhile, these automated natural language processing tools are being developed in other languages, which will broaden the scope of text complexity metrics in other languages and cultures.

The automated text analysis measures can also be systematically applied to translation. For example, in PISA there are versions of a question that are translated from French to the target language (i.e., Arabic in this case) and another version from English to the target language. There are also backward translations from the target language back to English and French. Further research should explore the advantages of implementing automated text analysis tools to assess the extent to which the original English version and the back translations from the target language to English are comparable in terms of overall complexity and complexity on the different levels of language and discourse. In addition to the complexity scales, modern tools of computational linguistics offer the possibility to systematically analyse the similarity of the different English versions of an assessment question at a fine-grained level, such as percentage of specific words that overlap, percentage of syntactic constituents that overlap, semantic similarity of sentences and of the questions as a whole. Future empirical research

should investigate the degree to which modern computational tools are successful in assessing the fidelity of translations in addition to the metrics of question complexity.

Conclusion

In this article, we argue that language is an inextricable part of the construct in science assessments and have examined the implications of this claim on international assessments. Be it in the international or the national setting (e.g. Lebanon, South Africa, Canada and Hong Kong), comparison of student achievement across language versions of the same test is needed. It is important, however, to ensure that these tests stand on solid conceptual and methodological grounds. Cross-cultural comparisons rely on the assumption that transadapted versions of the same test place similar language demands on examinees. However, even when the quality of transadaptation is not a concern, bias at some level is inevitable. So far, available DIF techniques are incapable of revealing the source of bias and may even fail to detect its presence. We have provided examples of items, including PISA items, in three languages to illustrate some of the challenges associated with the transadaptation of idiosyncratic elements. We argue that these particular idiosyncrasies may impose different cognitive demands on examinees in different countries. thereby raising concerns regarding the fairness of international comparisons and some of the conceptual underpinnings of the enterprise.

Exploring the conceptual framework underpinning international tests as well as the methodological shortcomings in developing them, and analysing the data may yield tremendous insight for a way forward. Awareness of the challenge of comparing educational systems of poor developing countries to the ones in wealthier and more developed countries has led to the launching of PISA for Development[10] in 2014 with currently four participating countries in Africa and Latin America. The education scene is nonetheless headed towards more international testing therefore alleviating some of the issues associated with test transadaptation becomes an imperative. Employing computational linguistics may strengthen the quality control stage and eliminate some of the bias inherent in the transadaptation process due to expert judgement. Future research should therefore investigate the prospects of

[10] http://www.oecd.org/pisa/aboutpisa/pisa-for-development-participating-countries.htm

implementing computational linguistics approaches in test transadaptation and to identify the advantages of employing these methods over expert judgement.

Acknowledgements:

This research was carried out under the auspices of a Doctor of Philosophy (DPhil) thesis programme at the University of Oxford in the UK. The authors would like to thank Professor Pauline Rea-Dickens for the invaluable comments on an earlier version of this manuscript.

References

Ahmed, A., & Pollitt, A. (1999). Curriculum demands and question difficulty. *Paper Presented at the International Association for Educational Assessment*. Bled, Slovenia.

Anderson, C. W. (1999). Inscriptions and science learning. *Journal of Research in Science Teaching*, *36*(9), 973–974.

Andrich, D. (2014). A structure of index and causal variables. *Rasch Measurement Transactions*, *28*(3), 1475–1477.

Andrich, D., & El Masri, Y. H. (*in preparation*). PISA and other large-scale assessments: the trade-off between model fit, invariance and validity.

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to other languages: Identifying potential sources of non- invariance. *International Journal of Testing*, *16*(1), 71–93.

Bailey, A. (2007). *The Language Demands of School: Putting Academic English to the Test*. New Haven, CT: Yale University Press.

Baird, J., & Black, P. (2013). Test theories, educational priorities and reliability of public examinations in England. *Research Papers in Education*, *28*(1), 5–21.

Baumert, J., Lüdtke, O., Trautwein, U., & Brunner, M. (2009). Large-scale student assessment studies measure the results of processes of knowledge acquisition: Evidence in support of the distinction between intelligence and student achievement. *Educational Research Review*, *4*(3), 165–176. doi:doi:10.1016/j.edurev.2009.04.002

Bramley, T., Hughes, S., Fisher-Hoch, H., & Pollitt, A. (1998). *Sources of Difficulty in Examination Questions: Science*. Research and Evaluation Division U.C.L.E.S.

Bybee, R. (1997). *Achieving Scientific Literacy: From Purposes to Practices*. Portsmouth, NH: Heinemann.

Bybee, R., & McCrae, B. (2011). Scientific literacy and student attitudes: Perspectives from PISA 2006 science. *International Journal of Science Education, 33*(1), 7–26.

Bybee, R., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching*, *46*(8), 865–883. doi:10.1002/tea.20333

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied*, *4*(2), 75–100.

Cole, N. S., & Zieky, M. J. (2001). The New Faces of Fairness. *Journal of Educational Measurement*, *38*(4), 369–382.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism*, *19*, 121–129.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In California State Department of Education (Ed.), *Schooling and Language Minority Students: A Theoretical Framework*. Los Angeles: Evaluation, Dissemination and Assessment Center California State University.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education* (2nd ed., pp. 71–83). New York: Springer Science + Business Media LLC.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, *37*(6), 582–601. doi:10.1002/1098-2736(200008)37:6<582::AID-TEA5>3.0.CO;2-L

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel-Haenszel method. *Applied Measurement in Education*, *2*, 217–233.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104*(3), 515–528.

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *The Irish Journal of Education / Iris Eireannach an Oideachais, 38*, 94–118. Retrieved from http://www.jstor.org/stable/20789130

El Masri, Y. H. (2015). *Comparability of Science Assessment Across Cultures: The Case of PISA Science 2006*. University of Oxford.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23–35.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 329–492.

Fisher-Hoch, H., Hughes, S., & Bramley, T. (1997). What makes GCSE examination questions difficult? Outcomes of manipulating difficulty of GCSE questions. *Paper Presented at the British Educational Research Association Annual Conference*. University of York: xxx.

Frantz, R. S., Bailey, A. L., Starr, L., & Perea, L. (2014). Measuring academic language proficiency in school-age English language proficiency assessments under new college and career readiness standards in the United States. *Language Assessment Quarterly*, *11*(4), 432–457.

Friel, S. N., Curcio, F. R., & Bright, G. M. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*(2), 124–158.

Gaskins, I. W., Guthrie, J. T., Satlow, E., Ostertag, J., Six, L., Byrne, J., & Connor, B.

(1994). Integrating instruction of science, reading, and writing: Goals, teacher development, and assessment. *Journal of Research in Science Teaching*, *31*, 1039–1056.

Gee, J. P. (2000). Discourse and sociocultural studies in reading . In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research (Vol.3)* (pp. 195–207). Mahweh, NJ: Erlbaum.

Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education: Principles, Policy and Practice*, *11*, 319–330.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel Structural Equation Models for the Analysis of Comparative Data on Educational Performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252–286.

Graesser, A. C., Bommareddy, S., Swamer, S., & Golding, J. M. (1996). Integrating questionnaire design with a cognitive computational model ofhuman question answering . In N. Schwartz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp. 143–174). San Francisco: Jossey-Bass.

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID). A web facility that tests question comprehensibility. *Public Opinion Quarterly*, *70*(1), 3–22. doi:10.1093/poq/nfj012

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal*, *15*, 199–213.

Graesser, A. C., & McNamara, D. S. (2011). Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, *3*(2), 371–398. doi:10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elementary School Journal*, *115*, 210–229.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234.

doi:10.3102/0013189X11413260

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249–266.

Grisay, A., & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation*, *33*(1), 69–86. doi:10.1016/j.stueduc.2007.01.006

Halliday, M. A. K. (1978). *Language as Social Semiotic: The Social Interpretation of Language Meaning*. London: Edward Arnold.

Halliday, M. A. K. (1985). *Spoken and Written Language*. Waurn Ponds, Vic: Deakin University.

Halliday, M. A. K. (1994). A language development approach to education. In N. Bird (Ed.), *Language and Learning* (pp. 5–17). Hong Kong: Institute of Language in Education.

Halliday, M. A. K., & Martin, J. R. (1993). *Writing Science: Literacy and Discursive Power*. Pittsburgh, PA: University of Pittsburgh.

Hambleton, R. K. (1993). Translating achievement tests for use in cross-national studies. *European Journal of Psychological Assessment*, *9*(1), 54–65.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164–172.

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.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Haneda, M. (2014). From academic language to academic communication: Building on English learners' resources. *Linguistics and Education, 26*, 126–135.

Hasan, R., & Martin, J. R. (1989). *Language Development: Learning Language, Learning culture*. Norwood, NJ: Ablex.

Holbrook, J., & Rannikmae, M. (2007). The Nature of Science Education for Enhancing Scientific Literacy. *International Journal of Science Education*, *29*(11), 1347–1362. doi:10.1080/09500690601007549

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), *Test Validity* (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: language, curriculum or culture. *Educational Psychology*, *36*(2), 378–390. doi:10.1080/01443410.2014.946890

International Test Commission. (2005). International Guidelines on Test Adaptation. Retrieved from www.intestcom.org

Jones, L. R., Wheeler, G., & Centurino, V. A. S. (2013). TIMSS 2015 science framework. In I. V. S. . Mullis & M. O. Martin (Eds.), *TIMSS 2015 Assessment Frameworks* (pp. 29–59). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, *10*, 62–102.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210–231.

Landauer, T. K. (2012). Improving text complexity measurement through the Reading Maturity Metric. http://images.pearsonassessments.com/images/tmrs/Word_Maturity_and_Text_Complexity_NCME.pdf: Pearson.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A New Metric for Word Knowledge. *Scientific Studies of Reading*, *15*(1), 92–108. doi:10.1080/10888438.2011.536130

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Mahwah, NJ: Lawrence Erlbaum.

Lederman, N. G. (1992). Students' and teachers' conceptions of the nature of science: A review of the research. *Journal of Research in Science Teaching*, *29*(4), 331–359. doi:10.1002/tea.3660290404

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied Cognitive Psychology*, *24*(7), 1003–1020. doi:10.1002/acp.1602

Li, H., Graesser, A. C., Conley, M., Cai, Z., Pavlik, P., & Pennebaker, J. W. (2015). A new measure of text formality: An analysis of discourse of Mao Zedong. *Discourse Processes*, 1–28. doi:10.1080/0163853X.2015.1010191

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS One*, *5*(1), e8559.

Mahboob, A. (n.d.). Understanding language variation: Implications of the NNEST lens for TESOL teacher education programs. In J. de Dios & A. Mart'nez (Eds.), *Native and Non-Native Speakers in English Language Teaching: Implications and Challenges for Teacher Education*. Boston: De Gruyter Mouton.

Maul, A. (2013a). Method effects and the meaning of measurement. *Frontiers in Psychology*, *4*, 169. doi:doi:10.3389/fpsyg.2013.00169

Maul, A. (2013b). On the ontology of psychological attributes. *Theory and Psychology*, *23*(6), 752–769. doi:doi:10.1177/0959354313506273

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix.* Cambridge, MA: Cambridge University Press.

Millar, R., & Osborne, J. F. (1998). *Beyond 2000: Science Education for the Future*. London: King's College London.

National Research Council. (1996). *National Science Education Standards*. Washington, DC: National Academy of Sciences.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance.* New York, NY: Student Achievement Partners.

Norman, O. (1998). Marginalized discourses and scientific literacy. *Journal of Research in Science Teaching, 35*, 365–374.

Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education, 87*(2), 224–240.

NSES. (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

OECD. (1999). *Measuring Student Knowlede and Skills: A New Framework for Assessment.* Paris, France: Author.

OECD. (2006). *Assessing Scientific, Reading and Mathematics Literacy: A Framework for PISA 2006.* Paris, France: Author.

OECD. (2009). *PISA Data Analysis Manual* (Vol. 2nd). France: Paris: OECD.

OECD. (2016). PISA 2015 Science Framework. In *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy* (pp. 17–46). Paris: OECD Publishing. doi:http://dx.doi.org/10.1787/9789264255425-3-en

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assesments lead to similar conclusions? *Applied Measurement in Education*, *24*(4), 349–366. doi:10.1080/08957347.2011.607063

Osborne, J. (2007). Science education for the twenty first century. *Eurasia Journal of Mathematics, Science & Technology Education*, *3*(3), 173–184.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity*, *10*, 265–279. doi:10.1016/j.tsc.2013.07.006

Pellegrino, J. W., & Hilton, M. L. (Eds.). (2015). *Education for Life and Work: Developing Transferable Knowledge and Skills in the 21st Century*. Washington, DC: The National

Academies Press.

Pollitt, A., & Ahmed, A. (1999). A new model of the question answering process. *Paper Presented at the International Association for Educational Assessment*. Bled, Slovenia.

Pollitt, A., Ahmed, A., & Crisp, V. (2007). The demands on examination syllabuses and question papers. In P. Newton, J.-A. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for Monitoring the Comparability of Examination Standards* (pp. 166–206). London: Qualifications and Curriculum Authority.

Pollitt, A., Entwistle, N. J., Hutchinson, C. J., & De Luca, C. (1985). *What makes exam questions difficult?* Edinburgh: Scottish Academic Press.

Poortinga, Y. H. (1995). Use of tests across cultures. In T. Oakland & R. K. Hambleton (Eds.), *International Perspectives on Academic Assessment* (pp. 187–206). Boston, MA: Kluwer Academic Publishers.

Prais, S. J. (2003). Cautions on OECD's recent educational survey (PISA). *Oxford Review of Education, 29*, 139–163.

Roth, W.-M. (2002). Reading graphs: Contributions to an integrative concept of literacy. *Journal of Curriculum Studies*, *34*(1), 1–24.

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach. *International Journal of Testing*, *13*(2), 152–174. doi:10.1080/15305058.2012.690140

Scarcella, R. (2003). *Academic English: A conceptual framework (Technical Report NO. 2003–1)*. Irvine, CA: The University of California Linguistic Minority Research Institute.

Schleppegrell, M. J., & Colombi, M. C. (Eds.). (2002). *Developing advanced literacy in first and second languages: Meaning with power*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shala, L., Rus, V., & Graesser, A. C. (2010). Automated speech act classification in Arabic. *Subjetividad Y Procesos Cognitivos*, *14*, 284–292.

Shala, L., Vasile, R., & Graesser, A. C. (2014). A bilingual analysis of cohesion in a corpus of leader speeches. In E. William & C. Boonthum-Denecke (Eds.), *Twenty-Seventh International Florida Artificial Intelligence Research Society Conference* (pp. 225–230). Palo Alto, California: The AAAI Press.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool. *Elementary School Journal*, *115*, 184–209.

Sireci, S. G. (1997). Technical issues in linking tests across languages. *Educational Measurement: Issues and Practice*, *16*(1), 12–19.

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*, 450–452.

Snow, C. E., & Ucelli, P. (2009). The challenge of academic language. In D. R. Olson & N. Torrance (Eds.), *The Cambridge Handbook of Literacy* (pp. 112–133). New York, NY: Cambridge University Press.

Spielberger, C. D., Moscoso, M. S., & Brunner, T. M. (2005). Cross-cultural assessment of emotional states and personality traits. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 343–368). Mahwah, NJ: Lawrence Erlbaum Associates.

Stroud, C., & Kerfoot, C. (2013). Towards rethinking multilingualism and language policy for academic literacies. *Linguistics and Education*, *24*(4), 396–405.

Thomas, G., & Durant, J. (1987). Why should we promote the public understanding of science? *Scientific Literacy Papers, 1*, 1–14.

Trudgill, P. (2011). *Sociolinguistics Typology: Social Determinants of Linguistic Structure and Complexity*. Oxford: Oxford University Press.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (pp. 39–63). Mahwah, NJ: Lawrence Erlbaum Associates.

Wellington, J., & Osborne. (2001). *Language and Literacy in Science Education*. Philadelphia, PA: OpenUniversity Press.

Wiliam, D. (2008). International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 253–257. doi:10.1080/09695940802417426

Wiliam, D. (2010). What counts as educational achievement? The role of constructs in the pursuit of equity in assessment. *Review of Research in Education*, *34*, 254–284.

Wu, A. D., & Ercikan, K. (2006). Using Multiple-Variable Matching to Identify Cultural Sources of Differential Item Functioning. *International Journal of Testing*, *6*(3), 287–300. doi:10.1207/s15327574ijt0603_5

Yore, L. D., & Treagust, D. F. (2006). Current Realities and Future Possibilities: Language and science literacy—empowering research and informing instruction. *International Journal of Science Education*, *28*(2-3), 291–314. doi:10.1080/09500690500336973

Yu, A., & Ebbs, D. (2012). Translation and translation verification. In M. O. Martin & I. V. S. Mullis (Eds.), *Methods and procedures in TIMSS 2011 and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. Retrieved from http://timssandpirls.bc.edu/methods/t-translation.html

Yu, G. (2012). The cognitive processes of taking IELTs Academic Writing Task 1. In *IELTS Research Reports*. IDP: IELTS Australia and British Council.

Appendix 1: Released PISA science unit S195 (age 15)

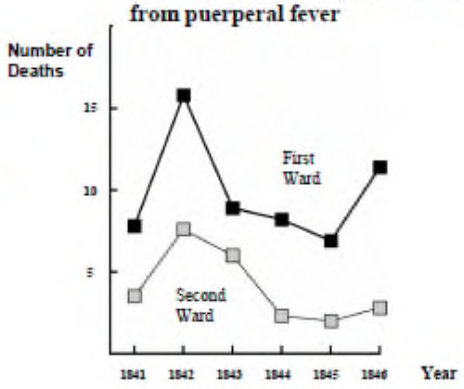
SEMMELWEIS' DIARY

SEMMELWEIS' DIARY TEXT 1

'July 1846. Next week I will take up a position as "Herr Doktor" at the First Ward of the maternity clinic of the Vienna General Hospital. I was frightened when I heard about the percentage of patients who die in this clinic. This month not less than 36 of the 208 mothers died there, all from puerperal fever. Giving birth to a child is as dangerous as first-degree pneumonia.'

These lines from the diary of Ignaz Semmelweis (1818-1865) illustrate the devastating effects of puerperal fever, a contagious disease that killed many women after childbirth. Semmelweis collected data about the number of deaths from puerperal fever in both the First and the Second Wards (see diagram).

Number of Deaths per 100 deliveries from puerperal fever



Diagram

Physicians, among them Semmelweis, were completely in the dark about the cause of puerperal fever. Semmelweis' diary again:

'December 1846. Why do so many women die from this fever after giving birth without any problems? For centuries science has told us that it is an invisible epidemic that kills mothers. Causes may be changes in the air or some extraterrestrial influence or a movement of the earth itself, an earthquake.'

Nowadays not many people would consider extraterrestrial influence or an earthquake as possible causes of fever. But in the time Semmelweis lived, many people, even scientists, did! We now know it has to do with hygienic conditions. Semmelweis knew that it was unlikely that fever could be caused by extraterrestrial influence or an earthquake. He pointed at the data he collected (see diagram) and used this to try to persuade his colleagues.

Question 2: SEMMELWEIS' DIARY S195Q02- 01 02 03 04 11 12 13 21 99

Suppose you were Semmelweis. Give a reason (based on the data Semmelweis collected) why puerperal fever is unlikely to be caused by earthquakes.

.....

.....

.....

.....

Source: www.oecd.org/pisa/38709385.pdf.

Appendix 2: TIMSS item S022154 (Grade 8; age 14)

Transmission of cold in classroom

Scott went to school with a cold. Several days later, half of his classmates also had colds. What is one likely reason some classmates had colds but others did not?

Source: https://nces.ed.gov/TIMSS/educators.asp.

Appendix 3: PISA science 2006 multiple choice item (age 15)

Question 3: GENETICALLY MODIFIED CROPS

Corn was planted in 200 fields across the country. Why did the scientists use more than one site?

A So that many farmers could try the new GM corn.
B To see how much GM corn they could grow.
C To cover as much land as possible with the GM crop.
D To include various growth conditions for corn.