The Challenges of Measuring Outside-School-Time Educational Activities: Experiences and Lessons from the Programme for International Student Assessment (PISA)

MARK BRAY, MAGDA NUTSA KOBAKHIDZE, AND LARRY E. SUTER

Recent years have brought a major shift in the field of international comparative education with the rise of quantitative assessments of student achievement. Within these studies, outside-school-time (OST) as a supplement and complement to schooling has been included. However, OST is difficult to define and measure; and challenges multiply when comparisons are made across countries and cultures. The Programme for International Student Assessment (PISA) managed by the Organisation for Economic Cooperation and Development (OECD) has sought to confront these challenges. It has had some success, and the 2015 PISA iteration showed advances over its predecessors; but it also showed significant shortcomings. With a methodological thrust, this article examines the challenges of preparing appropriate questions in English, translating those questions into other languages, and preparing meaningful cross-national comparisons. The article has pertinence not only within the specific domain of OST studies but also the wider thrusts of quantification in international comparisons.

Many scholars have stressed the value of quantification within the social sciences and the broad field of educational studies (e.g., Franklin 2008; Gomm 2017), although others, including in the specific subfield of comparative education, have reservations about "datafication" and have critiqued trends and patterns (e.g., Shields 2015; Piattoeva et al. 2018). Carnoy (2019) has focused on international testing, which, he remarked, "has become the dominant force in the field of comparative and international education" (567). Yet although the architects of international testing sought to strengthen the empirical study of educational systems, Carnoy observed that comparative educators "got more than they bargained for" (569). Carnoy highlighted some of the political issues arising from oversimplification of findings and methodological issues, including selection bias and validity of data.

The present article elaborates on these concerns with specific focus on outside-school-time (OST) educational activities as measured by the Programme for International Student Assessment (PISA), which is operated under the auspices of the Organisation for Economic Co-operation and Development (OECD). Existing literature addresses not only ways in which OST

Received April 16, 2018; revised October 13, 2018, and May 8, 2019; accepted July 20, 2019; electronically published January 21, 2020

Comparative Education Review, vol. 64, no. 1.

0010-4086/2020/6401-0005$10.00

activities support in-school learning but also ways in which such activities contribute to rounded personal development (e.g., Park et al. 2016; Suter 2016). In the process, the literature stresses the value of supplementation and complementarity. However, this literature faces significant definitional challenges in assessing the nature and contributions of OST education, especially in cross-national settings. Several publications (e.g., Zhou and Zou 2016; Chen and Zhi 2018; Liao and Huang 2018) have presented misleading findings from analysis of PISA data about OST activities, and the current article sends a note of caution to authors undertaking similar studies. Some of the challenges faced by OST measurement reflect wider shortcomings in efforts to quantify dimensions of education across national boundaries. As such, examination of these challenges sheds light on broader issues in international comparisons of learning experiences.

PISA is renowned for its scope and its importance to both policy makers and practitioners.[1] Since the first PISA round in 2000, assessments have been administered every 3 years. The 2015 iteration sampled students in 72 countries and self-governing economies (OECD 2016, 11).[2] PISA assessments focus on the learning of 15-year-olds, most of whom are concentrated in a single grade but some of whom are in higher or lower grades. Each round has covered science, mathematics, and literacy, with particular stress on one of these domains in each iteration. In addition, students and school principals provide contextual information. Particularly relevant to the present article was the Educational Career Questionnaire (ECQ), which in 2015 was administered in redesigned form to students in 22 countries.

The OECD has declared (2016, 10) that "stringent quality-assurance mechanisms are applied in translation, sampling and data collection," and that as a consequence "the results of PISA have a high degree of validity and reliability." These remarks have justification since PISA is well resourced in much professional expertise, but problems still arise in the framing of questions, translations, sampling, and interpretation of findings. This article builds on an earlier publication in the *Comparative Education Review* (Bray and Kobakhidze 2014). That publication focused not only on PISA but also on the Trends in Mathematics and Science Study (TIMSS) and noted challenges in cross-national measurement of supplementary private tutoring (parts of which are widely known as shadow education; see, e.g., Bray 2009, 2017; Lee et al. 2009). The present article has a wider focus on OST educational activities, and it identifies methodological lessons from the 2015 PISA iteration.

[1] Meyer and Benavot (2013); Sellar et al. (2017); Maddox (2019); Waldow and Steiner-Khamsi (2019).

[2] For brevity, the rest of this article refers only to countries but recognizes that some jurisdictions (such as Hong Kong) are self-governing jurisdictions rather than countries.

OST Scope and Definitions

Around the world, the term OST is most commonly used in the United States. One origin of research interest in that country has been concern about undesirable activities of youth with excessive free time and about the care of children whose parents are working or otherwise unavailable. Government investment in OST activities and corresponding evaluations expanded during the mid-1990s (Afterschool Alliance 2015, 1), and much research attention shifted from prevention of undesirable consequences from unattended OST to assessment of outcomes of specific extracurricular programs (e.g., Mahoney et al. 2005, 2009). A major benefit of these studies was recognition that much learning, especially of nonacademic dimensions but even in academic domains, occurred beyond the school classroom. Broad definitions also helped commentators to view developmental patterns for children and youth in integrated ways.

Other studies have focused more exclusively on academic learning. Specifically in the domain of PISA achievement rankings, commentators have pointed out that science, mathematics and/or literacy scores may reflect out-of-school supplementary tutoring as well as in-school learning (Ma et al. 2013; Park 2013). This factor has been especially evident in parts of East Asia and was a major reason for including sharper questions in later iterations of PISA. However, problems arose in definitions even of out-of-school supplementary tutoring, let alone broader forms of OST; one challenge for the OECD is to cater to the emphases in different societies. Thus, for example, US-based advocates of focus on OST may have very different modes in mind from South Korean advocates.

These matters were to some extent identified in the article by Bray and Kobakhidze (2014) on which the present article draws. That article was specifically concerned with private supplementary tutoring and began with definitional issues. For example, while many people define private in a financial light, that is, whether the consumers have to pay for the service, others define private as being outside the public space even if free of charge. Second, supplementary may imply additional content within the confines of the existing school curriculum, or it may mean additional content beyond the existing curriculum. And third, while tutoring implies to many people one-to-one instruction, it can also mean small group or even large-class provision. When the wider lens of OST is brought into use, the possible ranges of content and format multiply further.

The OECD (2011) recognized the challenges of diverse content and format in a publication titled *Quality Time for Students: Learning in and out of School*, which drew on data from the 2006 PISA iteration. The chapter on patterns of students' learning time compared regular school lessons with OST lessons and individual study; the chapter on different population subgroups

examined socioeconomic status, immigrant background, and other dimensions; and the chapter on relationships between students' learning time and performance sought insights on whether students who studied longer performed better. The report observed that "differences in the time spent in regular school lessons, according to students' and schools' characteristics, tend to look similar to those spent in individual study" (40), and that this finding was largely consistent across countries. However, the report added "the same is not true for the time spent in out-of-school-time lessons. Because the nature, meaning and function of out-of-school time lessons are not necessarily the same across countries, differences in the time spent in out-of-school-time lessons . . . are more complex and the results vary across countries" (40).

Subsequent PISA iterations sought more clarity on the complexities and variations. As this article shows, however, even the 2015 iteration encountered major problems in design for this particular task.

Structure and Content of the OST Questions in PISA 2015

Framing the Questions

The modular structure of the PISA 2015 context assessment design had 19 areas that were considered important and may be grouped as student background, education processes, activities, actors, educational outcomes, resources, and career aspirations (Jude 2016, 47; Klieme 2017, 9; OECD 2017, 60). Different items about the student, school, and other questionnaires addressed these 19 areas. The 2015 PISA "context" questionnaire included items on student background, educational processes, and strategies. The process items included grade repetition, program attended, learning time at school (mandatory lessons and additional instruction), and "out-of-school learning." The framework also included specific outcomes for school subject domains. The 2015 PISA round gave particular emphasis to science and less attention to literacy and mathematics.

Most OST questions in the 2015 PISA survey were in the ECQ addressed only by students in 22 countries. However, one question about additional study time was contained in the Student Questionnaire (StQ) administered in all 72 participating countries (fig. 1). It asked in general terms about OST hours per week spent on specific school topics (science, mathematics, and languages; "other" was included but not defined). As indicated in stem of the question, the OST hours were to include homework, additional instruction, and private study beyond the required school schedule. The present article is more concerned with additional instruction than with homework and private study, and from this perspective it is regrettable that the three categories were combined. Furthermore, the inclusion of homework in the category obstructed analysis of forms of additional instruction. And while the question

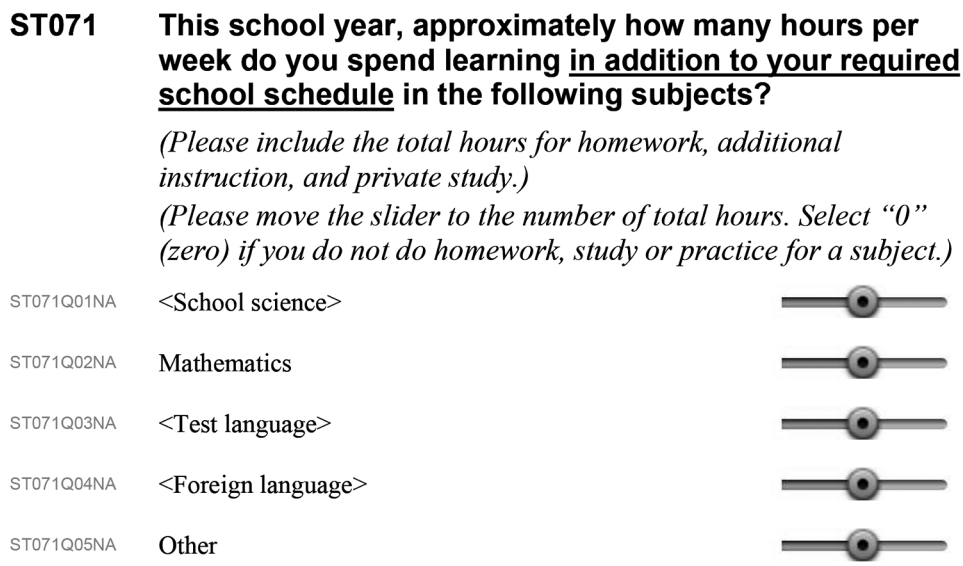

FIG. 1.—StQ 071: Hours of additional learning, by subject. Source: OECD (2014). A color version of this figure is available online.

specified "this school year," it did not allow for variations in examination season, midterm break, vacations, and so forth. This arrangement allowed for inconsistent and confusing reporting of additional instruction during breaks in the school year.

ECQ question 001 (fig. 2) focused explicitly on the number of hours of additional instruction rather than on learning that included homework and self-study in the current school year. Alongside academic studies were music, sports, performing arts, visual arts, and others. Subsequent questions were asked to students who had indicated attendance in each subject area. For science, these subsequent questions addressed:

- the subject-components of the instruction (physics, chemistry, biology and others);
- whether content was additional to that in regular school courses;
- the type of instruction, with options including one-on-one tutoring, video-recorded instruction, and large-group study;
- whether the instruction was provided in the regular school building or elsewhere;
- the identity of the instructor, with options including one of the respondent's regular teachers in this year's courses, a person working mainly for a business, or a nonspecialist instructor such as a student;

EC001	In this school year, approximately how many hours per week do you attend additional instruction in the following domains in addition to mandatory school lessons? *(An hour here refers to 60 minutes, not to a class period.)* *(Please move the slider to the number of hours you attend, move it to "0" [zero] if you don't attend any additional instruction.)*	
EC001Q01NA	<School science> or <broad science>	
EC001Q02NA	Mathematics	
EC001Q03NA	<Test language>	
EC001Q04NA	<Foreign languages>	
EC001Q05NA	Social sciences (e.g. history, sociology, politics)	
EC001Q06NA	Music (e.g. musical instrument, choir, composition)	
EC001Q07NA	Sports (e.g. in clubs, lessons, team)	
EC001Q08NA	Performing arts (e.g. dancing, acting)	
EC001Q09NA	Visual arts (e.g. photography, drawing, sculpting)	
EC001Q10NA	Other	

FIG. 2.—ECQ question 001: Hours of additional instruction, by subject. Source: OECD (2014). A color version of this figure is available online.

- the ways that teachers behaved and lessons proceeded in additional instruction compared with regular school classes;
- the nature of teacher-student interactions; and
- the reasons for attending additional instruction or, for those who did not attend, the reasons for not doing so.

The questionnaire thus solicited a great deal of information, much of which had not previously been sought in cross-national surveys. Analysts interested in marketization of education would have been disappointed that no question directly asked whether the additional instruction required payment (or how much), but, as noted, one question did ask if the instructor mainly worked for a business. A question about payment had been included in the semifinal version of the instrument but was among items dropped from a questionnaire that was already long.

All the questions about science were then repeated for those who took additional study in mathematics and test language with the exception of the initial question about subject-components of the instruction; and questions about the type of instruction, the reasons for attending, and the identity of the instructor were then asked about test language. These sets of questions thus permitted comparison of patterns in science, mathematics and test language.

Having asked about current receipt of additional instruction, the questionnaire asked whether respondents had received instruction at earlier stages and, if so, for how many years. It also asked who in the family helped regularly with homework or private study.

Potential Gaps between Intended and Received Meanings

Specialists in test design are always aware of the danger that intended meanings may not completely match the interpretations by the respondents (Karabenick et al. 2007; Koretz 2008). This challenge arises even when everything is conducted in single language and a single culture; challenges escalate when questions are translated and applied in diverse cultures (van de Vijver et al. 2011). The PISA team undertook piloting but could not remove all ambiguities. This section first addresses the questions in English and then notes some problems in translation.

Like other international assessments, PISA employs quality control mechanisms to ensure comparability of items across different languages. The PISA design includes double translations of the source versions in English and French; after translations have been done, multiple reviews of test items are conducted by experts and translators in each participating country. The PISA technical reports claim that the OECD takes contextual and linguistic factors into consideration seriously and makes extra efforts to establish cross-cultural equivalences (e.g., OECD 2017). However, problems evidently remained.

Cross-cultural measurement is considered valid if test items in different languages measure the same construct with comparable level of difficulty. Items should be equally easy or difficult and should require similar mental effort (Sweller 1988). If an item is difficult to understand, a student needs more time and effort and a considerable working memory. This objective may be challenging, and scholars have pointed out language-related biases and translation problems in various international assessments.[3]

Issues of cross-cultural equivalences are closely related to validity and reliability, as well as to the extent to which inferences and interpretations are meaningful and appropriate. Arffman (2012) compared PISA 2000 items in Finnish with the English and French versions and found that linguistic variations in the Finnish version might have affected students' performance. Other scholars have compared PISA 2006 items in Swedish, Danish, and Norwegian and have found inaccurate and awkward translations (Sjøberg 2015, 118–20). As Arffman (2012, 1) remarked, "a failure to establish equivalence jeopardizes the validity of inferences made on the basis of the results of the test." Avvisati and colleagues (2019) and Rutkowski and Rutkowski (2019) have also examined this matter, noting advances in PISA design and processes but also ongoing challenges.

Concerning the specific focus of the present article, figure 3 reproduces the ECQ question 005 about types of additional instruction. The subquestion about "live instruction by a person" potentially overlaps with "one-on-one tutoring with a person" and perhaps small-group and large-group study. It might seem surprising that the question survived the piloting as a priority item when the question about whether the tutoring required payment was cut.

[3] Arffman (2012); Sjøberg (2015); El Masri et al. (2016); Solano-Flores (2019).

EC005	→ Questions EC003-EC012 only apply, if a student attends any "additional science instruction". Otherwise skip EC003-EC012 and proceed to EC0013. **Which type of additional science instruction do you participate in during this school year?** *(Please select all that apply.)*	
EC005Q01NA	One-on-one tutoring with a person	☐1
EC005Q02NA	Internet tutoring with a person (including e.g. <Skype™>)	☐1
EC005Q03NA	Internet or computer tutoring with a programme or application	☐1
EC005Q04NA	Live instruction by a person	☐1
EC005Q05NA	Video-recorded instruction by a person	☐1
EC005Q06NA	Small group study or practice (2 to 7 students)	☐1
EC005Q07NA	Large group study or practice (8 or more students)	☐1
EC005Q08NA	Other additional science instruction	☐1

FIG. 3.—ECQ question 005: Types of additional instruction in science. Source: OECD (2014). A color version of this figure is available online.

Turning to translations, table 1 shows how the ECQ question 005 categories were presented in a number of languages. These versions were selected, first, because the countries to which they apply had an apparently adequate number of responses in contrast to high nonresponse rates elsewhere and, second, because they reported high OST participation rates. The translations were undertaken by professionals from the field who were native speakers of the target languages and fluent in English. Translations of the first item, "one-on-one tutoring with a person," were largely consistent across languages. However, the Greek translators asked about "private lessons," which may be explicitly or implicitly defined in different ways. "Private" can be interpreted financially to signal fee-based tutoring, but it may also be understood as tutoring received privately, that is, away from schools and public space. Furthermore, as noted, for some people the word "private" is especially related to one-on-one tutoring (cf. Bray and Kobakhidze 2014, 592). The Bulgarian phrase "personal private lessons" seems to signal individual tutoring, but the meaning may be ambiguous. The Polish translation, "individual meetings with a teacher, tutor, or another person" could include general meetings with teachers that were not necessarily related to tutoring.

Another option, "live instruction by a person," not only overlaps with "one-on-one tutoring by a person" but also was confusing in some languages. In most languages, "live" was translated as "in the presence of a teacher"; but whether this means individual, small-, or large-group tutoring was open to respondents' interpretation. In Korea, it was the most favored option, and the Korean scholars consulted by the authors related it to popular forms of lecture-type tutoring by commercial centers called hagwons (see Park et al. 2011; Choi et al. 2012). In China, this option was translated as "live *mass* tutoring by one

person" (emphasis added), indicating large-group tutoring and contrasting with the Bulgarian version that was clearly related to individual consultation.

On a third dimension, the architects of the PISA 2015 questionnaire intended to distinguish internet (e.g., via Skype™) from video-recorded tutoring. However, the Greek and Italian versions did not specify that the video instruction was recorded, and this omission would have caused overlap with the category "Internet tutoring with a person." Thus, the questions varied across countries, with corresponding implications.

As table 1 shows, the authors of this article conducted systematic linguistic comparisons between English and other selected languages. The findings show that some OST items differed linguistically from the source items in English. There is reason to believe that these translations affected students' understanding of those items, which is a serious threat to test validity. Differences may arise not only because of translation errors but also because of the idiosyncrasies of meanings and connotations of OST items in different languages. Poor translations and errors commonly result in item bias (El Masri et al. 2016, 438), and nonequivalent translations raise questions about comparability of items across countries.

Samples and Administrative Issues

The next challenge for analysts lies in sampling. As noted, the StQ was administered as part of the package to all respondents in the 72 countries, but the ECQ was an option and thus administered in only 22 countries and in some cases to limited samples within them. The sample in Australia was reduced by selecting only every third student from the original sample; in the United Kingdom the ECQ was administered in England, Wales, and Northern Ireland but not in Scotland; German students were not asked the follow-up questions about their additional study; and the Belgian authorities administered the ECQ in the French-speaking school system but not the German-speaking or Flemish-speaking systems.

A further challenge concerns nonresponses among the students who did receive the questionnaire. For example, students who responded "no" to the screening question about hours of additional study were omitted from follow-up items; students did not reach some items because of time; and some students did not answer some items because they thought that a nonresponse meant "no."[4] Furthermore, no code existed for "question not administered

[4] Since no follow-up was conducted, analysts have no way to check the reliability of the initial screening item. It is likely that some students selected No to avoid answering questions, in which case the responses to the main items would be based on a biased sample. The opposite also happened. In Thailand, for example, it appears that some students were biased to answer "yes" to everything possible. Also, the ECQ was answered only after the StQ and the information-and-communication-technology questionnaire (ICQ). The StQ was supposed to be answered in 30 minutes, but no standard was in place to handle students who had not completed. Test administrators who granted more time for the StQ added pressure to the ECQ component; yet the strategies of administrators were not recorded, and therefore could not be taken into account during analysis.

TABLE 1
ECQ Question 005: Which Type of Additional Science Instruction Do You Participate in during the School Year?

English	Korean	Thai	Chinese (Hong Kong)	Chinese (China)	Italian	Greek	Bulgarian	Slovenian	Polish
One-on-one tutoring with a person	One-on-one tutoring with a person	One-on-one with an instructor	One-on-one tutor's guidance	One-on-one live tutoring	One-on-one tutoring with a person	Private lessons	Personal private lessons	Individual lessons with an instructor	Individual meeting with a teacher, tutor or another person
Internet tutoring with a person (including, e.g., Skype™)	Internet lessons with your teacher (e.g., Skype™)	Internet teaching with an instructor such as Skype™	Internet tutor's guidance (e.g., using Skype™)	One-on-one Internet tutoring (including, e.g., QQ)	Internet tutoring with a person (including, e.g., Skype™)	Internet private lessons (e.g., via Skype™)	Education over Internet with a teacher (including, e.g., Skype™)	Internet tutoring (e.g., via Skype™)	Lessons taught by a teacher, tutor or another person via internet (e.g., Skype™)
Internet or computer tutoring with a program or application	Internet or computer lessons using program or application without a teacher	Internet or computer instruction using a program or application	Learning through the Internet/ computer program or application	Internet or computer tutoring with a program or application	Internet or computer lessons with a program or application	Internet or computer lessons with computer program or application	Education over the Internet or computer-based education with the help of a special program	Internet or computer lessons using program or application	Teaching with internet or computer, using program or application
Live instruction by a person	Live instruction (lecture) by a person	Live teaching with an instructor	Live teaching	Live mass tutoring by one person	Live lessons by a person	Lessons with (or in) the presence of a teacher	Consultations with a teacher	Lessons in the presence of an instructor	Lessons conducted in person ("live") by a teacher, tutor or another person

Video-recorded instruction by a person	Video-recorded lessons	Video recorded instruction by an instructor	Video-recorded teaching	Video-recorded mass tutoring by one person	Video lessons by a person	Video lessons by a teacher	Education with a teacher on video-recording	Learning via recorded lessons	Lessons recorded on video, run by a teacher, tutor or another person
Small group study or practice (2–7 students)	Small group study or practice (2–7 people)	Small group study or practice (2–7 people)	Small group learning or practice (2–7 students)	Small group learning or practice (2–7 students)	Small group study or practice (2–7 students)	Study or practice in a small group (2–7 people)	Education in a small group (2–7 students)	Small group study (2–7 students)	Learning/lessons in a small group (2–7 students)
Large group study or practice (8 or more students)	Large group study or practice (8 or more people)	Large group study or practice (8 or more people)	Large group learning or practice (8 or more students)	Large group learning or practice (8 or more students)	Large group study or practice (8 or more students)	Study or practice in a large group (8 or more students)	Education in a big group (8 or more students)	Large group study (8 or more students)	Learning/lessons in a large group (8 or more students)
Other additional science instruction	Other additional science instruction	Other additional science subjects	Other additional science teaching	Other science extracurricular/ out-of-class tutoring	Other additional science lessons	Other additional types of natural sciences	Other forms of supplementary education in natural sciences	Other forms/ types of additional lessons or activities in science	Other types of additional lessons in science

Source: Questionnaires in the original languages are found in the database of the German Institute for International Educational Research (DIPF), http://daqs.fachportal-paedagogik.de/search/show/survey/177?language=en.

because country deleted it." As a result, the meaning of nonresponses was not always clear. Accounting correctly for nonresponse meanings is especially important for defining the denominator for rates of participation. For example, when calculating the percentages of students attending additional study with a one-on-one tutor in science, the researcher must decide whether to count only those answers that included zero hours or to include additional nonresponse categories in the denominator.[5]

A challenge for the 2015 iteration of PISA in contrast to its predecessors arose from the decision to make all ECQ responses (and most StQ responses) computer-based. Care was taken in advance to explore the implications, but some dimensions were overlooked. Responses on the number of hours were requested on a slider. The questionnaire did request respondents to move the slider to zero if they did not receive any additional OST instruction for that component (see, e.g., figs. 1 and 2), but many respondents did not do so. This experience again shows the potential gaps between a questionnaire instruction and comprehension and/or compliance. The data-processors had to code many responses as missing when, in fact, many of them probably should have been zero.

A further question is whether lack of familiarity with computers caused some students, particular in rural areas of lower-income countries, to be impeded when making their responses. Zhang (2017) examined data in China and concluded that, indeed, students in such areas were impeded, with the result that their data were less reliable.[6] Other scholars have also compared student academic achievements based on paper-and-pencil and computer tests. For example, Jerrim (2016, 513–14) analyzed PISA 2012 mathematics items and found significant differences resulting from mode differences in terms of mean scores and covariation with key demographic characteristics. These examples demonstrate that computer-based testing potentially affects how students respond and interpret test items and background questionnaires.

Methodological Insights from Analysis of the Data

Full evaluation of the reliability and validity of the PISA items for additional study would ideally include with follow-up interviews. That cannot easily

[5] One possibility for the denominator is to include the entire sample. Another possibility is to include only those who answered that they attended or did not attend additional study. A third possibility is to include only those who checked 'zero' or more hours in the denominator, omitting the nonresponses. The nonresponse rates varied by country, and therefore comparisons on rates between countries are affected by the decision on the denominator. Imputation of missing values would also require decisions on the above options. The categories of nonresponse given on the data base are not sufficiently clear to know which respondents were supposed to answer. Some countries, such as Australia, reduced the sample by two thirds. Differences in rate of attending OST between countries are high, but the true sizes of the differences and also the true errors in response to the items are impossible to calculate. That is one reason why the examination of relationships between items for those who responded was chosen for analysis. The respondents should be consistent across categories and countries if they understood the items the same way.

[6] OECD (2010); Bennett (2015); Herold (2016); Zhao (2016).

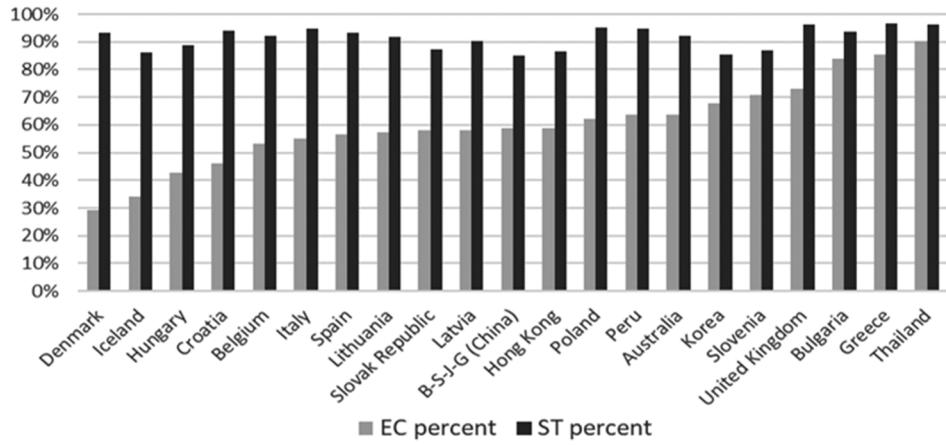

FIG. 4.—Percent of responses to ECQ and StQ forms that reported attending "additional study" in science for 1 or more hours in past week.

be accomplished by the present authors, especially on a post hoc basis. Nevertheless, indications can be achieved through examination of response rates to each set of categories of OST.

Comparison between the StQ and ECQ Forms

Looking across the full StQ and ECQ instruments, analysts may ask what StQ question 071 indicates and how far it matches the data from ECQ question 001 (fig. 4). The answer depends on the country. The differences in student responses to number of hours of additional study in the StQ and ECQ were inconsistent from country to country. For example, Belgium, China, and Slovak Republic reported extremely high average hours of additional study from the StQ questions compared with other countries.[7] The StQ item was expected to be more frequently reported because the instructions to respondents asked for homework to be included while homework was specifically excluded from the ECQ item. The differences between the two forms were not consistent across countries. The largest differences were in Croatia, Denmark, Iceland, and Hungary, while response rates were about the same in Bulgaria, Greece, and Thailand. In other words, students in some countries made little distinction between hours of additional study and specific hours in the eight types

[7] The StQ data average about 2 hours greater than the ECQ data, most likely because the StQ question included homework. For example, the difference between the two measures among countries ranges from one hour (Belgium) to seven hours (Croatia). The two measures are correlated with each other, but the range of relationships varies from .3 in Belgium to .6 in Thailand, suggesting that one measure cannot be a reliable indicator of the other. The variation in reporting within countries is also very high, with standard deviations 2–6 times the average. This suggests that country averages are not good representations of individual students' reported time in either indicator.

specified in the ECQ. By contrast, students in other countries viewed the items as completely different from each other.

Comparison of Similar Groupings

Another method for examining differences in interpretation of the OST response categories is comparison of the magnitudes of differences in response rates for similar items. Eight items identified sources of additional instruction in science, mathematics and literacy using yes/no responses (fig. 3). Seven of the response categories may be combined into three OST categories: tutor or personal assistance; use of technology; and attending group sessions. The level of consistency in student responses to the common OST instructional types is a rough indicator of the degree of common understanding among students in different countries. However, this comparison provides little evidence of either consistency or lack of consistency across countries. Some countries, especially Korea, have very high proportions in one category (tutor with a live person) and few other responses. Thus, the 2015 categories do appear to capture meaningful country differences in type of additional study, but the reliability of the response categories is impossible to gauge with available data.

The next question is what the magnitudes of response rates on the ECQ questions indicate about which data can be used and which should be discarded on the grounds that insufficient numbers of respondents appear to have understood the questions with adequate clarity. This question cannot be answered in general because of diverse patterns. The ECQ responses can provide rough country estimates of participation rates in OST study and of the numbers of hours in each domain, but estimates are rough because the researcher must make judgments about whether to include nonresponses as a negative or not. The meaning of a nonresponse seems to be different in each country, and specific information from country managers is not in the public domain.

Conclusion

This article commenced by noting trends in "datafication" of education and Carnoy's (2019, 569) observation that international testing "has become the dominant force in the field of comparative and international education" in which comparative educators "got more than they bargained for." Concerning the specific focus of this article, the PISA architects are to be applauded for their recognition (OECD 2011, 40) that "the nature, meaning and function of out-of-school time lessons are not necessarily the same across countries" and for their desire to understand the complexities more fully. However, the 2015 PISA iteration did not adequately achieve the goal. Indeed, there is a danger of further damage to the nuances sought by specialists in the field of comparative education because the apparently authoritative numbers

are in practice potentially misleading. This observation echoes and adds to remarks by Rutkowksi and Rutkowski (2019, 137) that "there is a growing body of evidence that some of the constructs (or latent traits) developed by PISA . . . are not comparable across, or at time[s] even within, systems."

Elaborating, the conceptual, practical, and political power of PISA has greatly expanded over the period since its launch in 2000. Much effort has been devoted to improvements in questionnaire design, sampling, administration, and analysis, and corresponding advances have been achieved. However, this article, building on Bray and Kobakhidze (2014), has pointed out that problems with measurement of "additional study time" (after-school, shadow education, etc.) have remained significant. Concerning OST activities, the article has highlighted problems of focus, question construction, translation, and response rates. In some respects, these observations echo remarks by earlier scholars (e.g., Nardi 2008; Sellar et al. 2017). This article has given particular attention to private supplementary tutoring, which in contemporary times is recognized to have much greater significance than it was accorded in the initial years of PISA's work (Kuger and Klieme 2016, 30).

The current article underlines the importance of data literacy for analysts and widens the debate about measurement aspects of large-scale assessments. The context questionnaires and particularly the ECQ addressed in this article provide significant background information. In the domain of OST activities, the questions are especially important because researchers often seek to identify the effects (and roles) of tutoring on students' test scores by correlating tutoring with academic achievement (e.g., Lee et al. 2009; Choi et al. 2012). Appropriate orientation, measures across time, and adequate precision in background questionnaires are vital for well-grounded claims. Jude (2016, 49) recommended improvement of the context questionnaires "by applying advanced scaling methods, by using imputation methods for missing data, and by in-depth analysis of cross-national equivalence." Rutkowski and Rutkowski (2016, 256) took the notion further by recommending the PISA authorities to publish dedicated limitations chapters with the PISA reports. Both of these recommendations deserve echo and emphasis.

Another important methodological dimension concerns cross-national equivalences. Table 1 demonstrated varied translations of ECQ question 005 items, further highlighting threats to accurate measurement of concepts across countries and overall validity of international comparisons. Criticism of PISA translations is not new. Nardi (2008) identified issues related to culture and language that disadvantaged non-Anglophone countries. Fischman (2017) echoed these concerns; and Sellar (2017) further questioned the accuracy of comparing culturally sensitive test items across different languages, noting that translations can make some items more difficult or even confusing for students. This article has presented examples of items in different languages to illustrate some of the challenges associated with the PISA background

questionnaires. It aims to increase awareness of the threat that translation poses to the validity of PISA and related assessments. It is argued that language-related difficulties seem to be the most problematic differences to identify and deal with (Arffman 2012, 2). The article emphasizes that more clarity is needed on translation comparability, better revisions and verification of translations, and better guidelines and training for translators. Complete elimination of language-related bias might be impossible, but recognition of problematic items is a way to reduce wrong data interpretation by consumers of secondary data.

Valid translations also depend on clear definition of concepts in the original language. To some extent the problem of definition lies in the nature of OST activities themselves, since the category takes various forms and is very sensitive to cultural context. These problems may be related to what Van de Vijver and He (2016, 233–34) called construct bias (an item with different meanings in different contexts) and item bias (with ambiguous connotations due to linguistic and cultural factors).

These remarks lead to a larger question about the challenges of culture-free assessment when procedures aim to measure the same concept across diverse cultures. Tröhler (2013, 156) argued that the OECD's annual flagship publication *Education at a Glance* was inadequately sensitive to cultural differences, and that it reduced social meanings to statistics and figures. Similarly, Meyer and Benavot (2013, 10) noted that critics "question the possibility of a culturally neutral educational platform in which the same test and test questions are used in countries whose social, economic, cultural, and colonial backgrounds are so vastly different." This article has shown that apparently simple questions do not always generate correspondingly simple data.

These observations demonstrate broader challenges in translating and adapting the instruments, and raise questions about the reliability and comparability of data. Although the OECD boasts a rigorous process of translation and verification (e.g., OECD 2017, 91–99), continued limitations must be acknowledged. The process has long included back translations and revision to increase the accuracy of the national instruments, and in 2015 for the first time in PISA history included a translatability assessment. The process included work by linguists and item developers to identify and tackle translation and adaptation difficulties (OECD 2017, 63). Van de Vijver and He (2016, 235–36) explained in detail not only the statistical but also the nonstatistical strategies that had been employed for PISA 2015 to enhance cross-cultural equivalence. However, this article has pointed out that ambiguities remained in the English versions of the questionnaires and were compounded by the translations.

Another dimension concerns the possibility of measuring trends over time. Many scholars have taken data from different iterations of large-scale assessments and have arrayed them on bar charts and tables (e.g., Baker et al.

2001; Runte-Geidel 2013; Vest et al. 2013). It is indeed valuable to have had questions on OST activities in the 2006, 2009, 2012 and 2015 PISA iterations, but much caution is needed when comparing data across time. One of the coeditors of a volume focused on the PISA 2015 assessment framework explicitly counseled against comparisons across different cycles (Jude 2016, 48): "Measuring trends can only be guaranteed when the measures are kept stable. However, as the underlying frameworks have changed over time, only a limited set of questionnaire indicators can be compared between the cycles. Moreover, changing policy interests, and also changes in the learning context itself need to be accounted for." Jude and colleagues are to be applauded for such methodological clarity and indication of limitations.

This article contributes to informed scholarly debate and need for critical examination of one of the most influential international assessments. PISA and related enterprises have had significant impact on educational systems and have generated public discussions across the globe. They have also contributed significantly to the field of comparative education. However methodological deficiencies, measurement errors and biases should be acknowledged with the hope that shedding light on such areas will improve the quality of data across countries and help stakeholders to make appropriately informed decisions. At the same time, the measurement challenges highlighted in this article stress the importance of OST analysts themselves finding ways to improve definitions and to identify conceptual and organizational differences across countries and cultures with greater clarity.

References

Afterschool Alliance. 2015. "Evaluations Backgrounder: A Summary of Formal Evaluations of Afterschool Programs' Impact on Academics, Behavior, Safety and Family Life." http://afterschoolalliance.org//documents/Evaluation_Backgrounder.pdf.

Arffman, Inga. 2012. "International Education Studies: Increasing Their Linguistic Comparability by Developing Judgmental Reviews." Finnish Institute for Educational Research, University of Jyväskylä.

Avvisati, Francesco, Noémi Le Donné, and Marco Paccagnella. 2019. "Conclusion: An OECD Conference on the Cross-Cultural Comparability of Questionnaire Measures in Large-Scale Assessments." In *Invariance Analyses in Large-Scale Studies*, ed. F. J. R. van der Vijver. OECD Education Paper no. 201, Paris: OECD.

Baker, David P., Motoko Akiba, Gerald K. LeTendre, and Alexander W. Wiseman. 2001. "Worldwide Shadow Education: Outside-School Learning, Institutional Quality of Schooling, and Cross-National Mathematics Achievement." *Educational Evaluation and Policy Analysis* 23 (1): 1–17.

Bennett, Randy E. 2015. "The Changing Nature of Educational Assessment." *Review of Research in Education* 39 (March): 370–407.

Bray, Mark. 2009. *Confronting the Shadow Education System: What Government Policies for What Private Tutoring?*. Paris: UNESCO.

Bray, Mark. 2017. "Schooling and Its Supplements: Changing Global Patterns and Implications for Comparative Education." *Comparative Education Review* 62 (3): 469–91.

Bray, Mark, and Magda Nutsa Kobakhidze. 2014. "Measurement Issues in Research on Shadow Education: Challenges and Pitfalls Encountered in TIMSS and PISA." *Comparative Education Review* 58 (4): 590–620.

Carnoy, Martin. 2019. "The Uneasy Relation between International Testing and Comparative Education Research." In *SAGE Handbook of Comparative Studies in Education*, ed. L. E. Suter, E. Smith, and B. Denman. Los Angeles: SAGE.

Chen, Chunjiu, and Ting jin Zhi. 2018. "Supplementary Tutoring in International Comparison." *Guangming Daily*, 14 June [in Chinese].

Choi, Álvaro, Jorge Calero, and Josep-Oriol Escardíbul. 2012. "Private Tutoring and Academic Achievement in Korea: An Approach through PISA-2006." *KEDI Journal of Educational Policy* 9 (2): 299–322.

El Masri, Yasmine H., Jo-Anne Baird, and Art Graesser. 2016. "Language Effects in International Testing: The Case of PISA 2006 Science Items." *Assessment in Education: Principles, Policy and Practice* 23 (4): 427–55.

Fischman, Gustavo. 2017. FreshEd Podcast #70: "The Power and Perils of International Large Scale Assessments." https://soundcloud.com/freshed-podcast/freshed-70-the-power-and.

Franklin, Mark. 2008. "Quantitative Analysis." In *Approaches and Methodologies in the Social Sciences: A Pluralist Perspective*, ed. D. Porta and M. Keating. Cambridge: Cambridge University Press.

Gomm, Roger. 2017. "A Positivist Orientation: Hypothesis Testing and the 'Scientific Method.'" In *The BERA/SAGE Handbook of Educational Research*, ed. D. Wyse, N. Selwyn, E. Smith, and L. E. Suter. London: SAGE.

Herold, Benjamin. 2016. "Comparing Paper and Computer Testing: 7 Key Research Studies." *Education Week* 35 (22): 8.

Jerrim, John. 2016. "PISA 2012: How Do Results for the Paper and Computer Tests Compare?" *Assessment in Education: Principles, Policy and Practice* 23 (4): 495–518.

Jude, Nina. 2016. "The Assessment of Learning Contexts in PISA." In *Assessing Contexts of Learning: An International Perspective*, ed. S. Kuger, E. Klieme, N. Jude, and D. Kaplan. Dordrecht: Springer.

Karabenick, Stuart A., Michael E. Woolley, Jeanne M. Friedel, Bridget V. Ammon, Julianne Blazevski, Christina R. Bonney, Elizabeth De Groot, Melissa C. Gilbert, Lauren Musu, Toni M. Kempler, and Kristin L. Kelly. 2007. "Cognitive Processing of Self-Report Items in Educational Research: Do They Think What We Mean?" *Educational Psychologist* 42 (3): 139–51.

Klieme, Eckhard. 2017. "Adolescents' Extra-Curricular Activities, Well-Being and Educational Outcomes: Comparative Findings from PISA 2015." Keynote address at WERA-IRN Conference, "Extended Education from an International Comparative Point of View," World Education Research Association (WERA) International Research Network (IRN), University of Bamberg, December 2.

Koretz, Daniel. 2008. *Measuring Up: What Educational Tests Really Tell Us*. Cambridge, MA: Harvard University Press.

Kuger, Susanne, and Eckhard Klieme. 2016. "Dimensions of Context Assessment. " In *Assessing Contexts of Learning: An International Perspective*, ed. S. Kuger, E. Klieme, N. Jude, and D. Kaplan. Dordrecht: Springer.

Lee, Chong-Jae, Hyun-Jeong Park, and Heesook Lee. 2009. "Shadow Education Systems." In *Handbook of Education Policy Research*, ed. Gary Sykes, Barbara Schneider, and David N. Plank. New York: Routledge.

Liao, Xiangyi, and Xiaoting Huang. 2018. "Who Is More Likely to Participate in Private Tutoring and Does It Work? Evidence from PISA (2015)." *ECNU Review of Education* 1 (3): 69–95.

Ma, Xin, Cindy Jong, and Jing Yuan. 2013. "Exploring Reasons for the East Asian Success in PISA." In *PISA, Power, and Policy: The Emergence of Global Educational Governance*, ed. Heinz-Dieter Meyer and Aaron Benavot. Oxford: Symposium.

Maddox, Bryan, ed. 2019. *International Large-Scale Assessments in Education: Insider Research Perspectives*. London: Bloomsbury.

Mahoney, Joseph L., Reed W. Larson, Jacquelynne S. Eccles, and Heather Lord. 2005. "Organized Activities as Development Contexts for Children and Adolescents." In *Organized Activities as Contexts of Development: Extracurricular Activities, After-School and Community Programs*, ed. J. L. Mahoney, R. W. Larson, and J. S. Eccles. Mahwah, NJ: Erlbaum.

Mahoney, Joseph L., Deborah L. Vandell, Sandra Simpkins, and Nicole Zarrett. 2009. "Adolescent Out-of-School Activities." In *Handbook of Adolescent Psychology*, ed. R. Lerner and L. Steinberg. New York: Wiley.

Meyer, Heinz-Dieter, and Aaron Benavot, eds. 2013. *PISA, Power, and Policy: The Emergence of Global Educational Governance*. Oxford: Symposium.

Nardi, Emma. 2008. "Cultural Biases: A Non-Anglophone Perspective." *Assessment in Education: Principles, Policy and Practice* 15 (3): 259–66.

OECD. 2010. *PISA Computer-Based Assessment of Student Skills in Science*. Paris: OECD.

OECD. 2011. *Quality Time for Students: Learning in and out of School*. Paris: OECD.

OECD. 2014. *Student Questionnaire for PISA 2015: Computer-Based Version*, Main Survey Version. Paris: OECD. http://www.oecd.org/pisa/data/CY6_QST_MS_STQ_CBA_Final.pdf.

OECD. 2016. *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, and Financial Literacy*. Paris: OECD.

OECD. 2017. *PISA 2015 Technical Report*. Paris: OECD.

Park, Hyunjoon. 2013. *Re-Evaluating Education in Japan and Korea: Demystifying Stereotypes*. London: Routledge.

Park, Hyunjoon, Claudia Buchmann, Jaesung Choi, and Joseph J. Merry. 2016. "Learning beyond the School Walls: Trends and Implications." *Annual Review of Sociology* 42 (July): 231–52.

Park, Hyunjoon, Soo-yong Byun, and Kyung-keun Kim. 2011. "Parental Involvement and Students' Cognitive Outcomes in Korea Focusing on Private Tutoring." *Sociology of Education* 84 (1): 3–22.

Piattoeva, Nelli, Ezekiel Doxon-Román, and Noah Sobe. 2018. "The Datafication of Comparative Education." Webinar of Comparative and International Education Society (CIES) Post-Foundational Special Interest Group (SIG). https://postfoundational.weebly.com/webinars.html.

Runte-Geidel, Ariadne. 2013. "La Incidencia De Las Clases Particulares en España a Través de los Datos de PISA." [The incidence of private tuition in Spain through the PISA data]. *Revista Española de Educación Comparada* 21:249–82.

Rutkowski, Leslie, and David Rutkowski. 2016. "A Call for a More Measured Approach to Reporting and Interpreting PISA Results." *Educational Researcher* 45 (4): 252–57.

Rutkowski, Leslie, and David Rutkowski. 2019. "Methodological Challenges to Measuring Heterogeneous Populations Internationally." In *SAGE Handbook of Comparative Studies in Education*, ed. L. E. Suter, E. Smith, and B. Denman. Los Angeles: SAGE.

Sellar, Sam. 2017. FreshEd Podcast #75: "The Global Education Race." http://www.freshedpodcast.com/samsellar.

Sellar, Sam, Greg Thompson, and David Rutkowski. 2017. *The Global Education Race: Taking the Measure of PISA and International Testing*. Toronto: Brush Education.

Shields, Robin. 2015. "Measurement and Isomorphism in International Education." In *The SAGE Handbook of Research in International Education*, ed. M. Hayden, J. Levy, and J. Thomson. Los Angeles: SAGE.

Sjøberg, Svein. 2015. "PISA and Global Educational Governance: A Critique of the Project, Its Uses and Implications." *Eurasia Journal of Mathematics, Science and Technology Education* 11 (1): 111–27.

Solano-Flores, Guillermo. 2019. "The Participation of Latin American Countries in International Assessments: Assessment Capacity, Validity, and Fairness." In *SAGE Handbook on Comparative Studies in Education*, ed. L. E. Suter, E. Smith, and B. Denman. Los Angeles: SAGE.

Suter, Larry E. 2016. "Outside School Time: An Examination of Science Achievement and Noncognitive Characteristics of 15-Year-Olds in Several Countries." *International Journal of Science Education* 38 (4): 663–87.

Sweller, John. 1988. "Cognitive Load during Problem Solving: Effects on Learning." *Cognitive Science* 12 (2): 257–85.

Tröhler, Daniel. 2013. "The OECD and Cold War Culture: Thinking Historically about PISA." In *PISA, Power and Policy: The Emergence of Global Educational Governance*, ed. H. D. Meyer and A. Benavot. Oxford: Symposium.

van de Vijver, Fons J. R., Athanasios Chasiotis, and Seger M. Breugelmans, eds. 2011. *Fundamental Questions of Cross-Cultural Psychology*. Cambridge: Cambridge University Press.

van de Vijver, Fons J. R., and Jia He. 2016. "Bias Assessment and Prevention in Noncognitive Outcome Measures in Context Assessments." In *Assessing Contexts of Learning: An International Perspective*, ed. S. Kuger, E. Klieme, N. Jude, and D. Kaplan. Dordrecht: Springer.

Vest, Andrea E., Joseph L. Mahoney, and Sandra D. Simpkins. 2013. "Patterns of Out-of-School Time Use around the World: Do They Help to Explain International Differences in Mathematics and Science Achievement?" *International Journal for Research on Extended Education* 1 (1): 71–85.

Waldow, Florian, and Gita Steiner-Khamsi, eds. 2019. *Understanding PISA's Attractiveness: Critical Analyses in Education Policy Studies*. London: Bloomsbury.

Zhang, Minxuan. 2017. Personal communication to the authors, Shanghai Normal University.

Zhao, Yong. 2016. "Did the Shift from Paper to Computer Bring Down East Asia's (China's) PISA Performance?" http://nepc.colorado.edu/blog/did-shift.

Zhou, Jinyan, and Xue Zou. 2016. "Comparing the Private Tutoring Options between Students in China and the United States: Evidence from 2012 PISA Survey and Investigation." *Education and Economy* 2:44–52 [in Chinese].