Scandinavian Journal of Educational Research

ISSN: 0031-3831 (Print) 1470-1170 (Online) Journal homepage: https://www.tandfonline.com/loi/csje20

Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment

Therese N. Hopfenbeck, Jenny Lenkeit, Yasmine El Masri, Kate Cantrell, Jeanne Ryan & Jo-Anne Baird

To cite this article: Therese N. Hopfenbeck, Jenny Lenkeit, Yasmine El Masri, Kate Cantrell, Jeanne Ryan & Jo-Anne Baird (2018) Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment, Scandinavian Journal of Educational Research, 62:3, 333-353, DOI: 10.1080/00313831.2016.1258726

To link to this article: https://doi.org/10.1080/00313831.2016.1258726

SCANDINAVIAN JOURNAL OF EDUCATIONAL RESEARCH, 2018
VOL. 62, NO. 3, 333–353
https://doi.org/10.1080/00313831.2016.1258726

Routledge
Taylor & Francis Group

Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment

Therese N. Hopfenbeck, Jenny Lenkeit, Yasmine El Masri, Kate Cantrell, Jeanne Ryan and Jo-Anne Baird

Department of Education, Oxford University Centre for Educational Assessment, University of Oxford, Oxford, UK

ABSTRACT

International large-scale assessments are on the rise, with the Programme for International Student Assessment (PISA) seen by many as having strategic prominence in education policy debates. The present article reviews PISA-related English-language peer-reviewed articles from the programme's first cycle in 2000 to its most current in 2015. Five literature bases were searched, and results were analysed with SPSS. Results map the frequency of publications according to journal, country, and scientific discipline. They also summarise major themes within three identified categories: secondary analysis, policy impact, and critiques. Findings indicated that studies based on the PISA dataset has led to progress in educational research while simultaneously pointing to the need for caution when using this research to inform educational policy.

ARTICLE HISTORY
Received 3 January 2016
Accepted 30 October 2016

KEYWORDS
PISA; policy impact; secondary analysis; critiques

1. Introduction

In the past two decades, international large-scale assessments (ILSA) have been on the rise, with the Programme for International Student Assessment (PISA) seen by many as having strategic prominence in international education policy debates (Meyer & Benavot, 2013). The PISA was launched by the Organisation for Economic Co-operation and Development (OECD) in 1999 with the aim to assess "aspects of preparedness for adult life" (OECD, 2000, p. 3). The purpose in the initial documents stated that PISA would provide evidence on students' capacities to continue learning throughout their lives, because the "parents, students and the public and those who run the education systems need to know the answer to these questions" (p. 3) regarding how the education system prepares students for lifelong learning. The OECD launched PISA "in response to this demand" (p. 3).

Unlike the International Association for the Evaluation of Educational Achievement (IEA)'s Progress in International Reading Literacy Study (PIRLS), which has a curriculum approach, the PISA study has a literacy approach, in which the test content is independent of the participating countries' school curricula, with a focus upon assessing whether 15-year-olds are able to *apply* what they have learned in school in real life situations by the time they have finished their compulsory schooling.[1]

CONTACT Therese N. Hopfenbeck therese.hopfenbeck@education.ox.ac.uk Department of Education, Oxford University Centre for Educational Assessment, University of Oxford, Oxford, UK
Supplemental data for this article can be accessed http://dx.doi.org/10.1080/00313831.2016.1258726
[1]Retrieved from http://www.oecd.org/pisa/aboutpisa/pisafaq.htm

The PISA is a triennial study that measures students' knowledge and skills in three main domains: reading, scientific, and mathematical literacy. In each cycle, one of the domains is in focus, with reading in 2000 and 2009, mathematics in 2003 and 2012, and science in 2006 and 2015. In 2018, reading will be the main domain for the third time, giving researchers the ability to analyse trend results from three cycles of reading in which it has been the main focus, making the analyses more robust. In addition to the three domains, problem solving has been included in some of the cycles, as well as financial literacy and the new domain of global competency, which will be introduced in 2018.

Data from PISA are used in education policy formation in many countries; especially in grey literature, as opposed to academic journals (Baird et al., 2016; Lindblad, Pettersson, & Popkewitz, 2015; Ozga, 2012). However, the extent to which the research field has made progress using PISA data is less well documented. Two previous studies reviewed the number of academic journal articles published using PISA data. Both provided descriptive statistics on the kinds of articles published (Domínguez, Vieira, & Vidal, 2012; Lindblad et al., 2015) and one limited the discussion of articles to those comparing countries (Lindblad et al., 2015). This paper adds to the findings of these studies and investigates the frequency, nature, and content of PISA-related research articles published in peer-reviewed journals and summarises the most prominent topics in PISA-related research. Our focus is upon articles dealing with the cognitive domains tested in PISA (reading, science, and mathematics literacy). With this paper we aim to map how PISA-related research has evolved over the past 15 years in the research community and how it is distributed across different countries and disciplinary fields. We further aim to summarise key research themes addressed with PISA data, critical discourses of the study itself and the discussions concerned with the role of PISA in policy formation.

There is a spectrum of approaches to systematic literature reviews (Gough, Oliver, & Thomas, 2012). With ours, we intent to provide a comprehensive bibliographical synthesis of PISA-related research in academic journals, rather than a systematic review of specific themes around the study. Coinciding with the release of the new PISA 2015 results towards the end of the year 2016, this synthesis serves as an assessment of how PISA has (1) affected the research literature so far, (2) advanced research around educational topics in national and cross-national contexts, (3) initiated intense discourse around technical aspects of measurements in international comparisons, and (4) forged scholarly discussion around the impact of global actors on educational policies. It may serve as an overview for researchers new to the field of international large-scale assessments and may encourage others to reflect upon relationship between the scope of PISA-related research and knowledge gained from it. We use a similar methodology to that presented by Lenkeit, Chan, Hopfenbeck, and Baird (2015), who reviewed research published on the PIRLS, and proceed as described in the following section.

2. Methods

2.1. Literature Database

An electronic search was conducted in five scientific databases: ERIC, PsycINFO, Scopus, Web of Science, and Zetoc. The keywords employed in the search were "PISA" and "Programme for International Student Assessment." The search terms were set to all fields (title, abstract, and keywords, etc.) and the date range spanned a period of 15 years from January 1999 to September 2015. The five databases included were selected due to their extensive coverage of research literature in the social sciences, as well as their coverage of some areas in science. The publication records were managed in reference management software, EndNote X7. In the second phase of development, the initial search was repeated, but with two digital search engines: Google Scholar and SOLO (Oxford University's library catalogue). The search was conducted to exhaust publications that were not indexed or abstracted in the five databases. Due to the broader and more inclusive nature of these search engines, additional items were located in the search and added to the EndNote database. These items comprised a range of publications, including journal articles, conference proceedings,

government reports, feature articles, book chapters, book reviews, and editorials. In the final phase, a targeted search was conducted in three hand-picked journals that were known by the authors to publish PISA-related research. These journals were the *Journal for Educational Research Online* (*JERO*), *Large-Scale Assessments in Education, and* the *IERI Monograph Series* (Volumes 1 to 5). Again, these journals were omitted in preliminary searches because they were not indexed in the databases or the search engines.

Once the search phase was completed, the duplicate entries were removed from EndNote and full-text copies of the items were sourced. At this stage, an initial count of database items revealed a total of 1,001 publications (910 journal articles, 46 reports, 24 conference papers, 14 book chapters, 3 book reviews, 3 newspaper articles and 1 editorial). Thus, the search rounds revealed a large volume of publications across a range of forms and from a variety of research projects. However, it was clear from the outset that these publications were related to PISA research in different ways and to varying extents. In other words, although "PISA" appeared somewhere in each of the items, the nature and purpose of this inclusion varied considerably. For example, in several journal articles, PISA was mentioned briefly as an in-text reference or example. In such a paper, a standard phrase might include a popular maxim, such as: "In recent years, international studies of educational achievements, such as PISA, have become important benchmarks for policy makers." In other instances, the PISA acronym simply appeared in a cited work in the reference list only. It was clear that the foci of these articles were not PISA research or the common themes and topics that PISA extends to. Further, many of the articles, while drawing on PISA research, focused primarily on other large-scale assessments, such as the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science study (TIMSS). To address this issue, misplaced articles were removed from the database and specific criteria for data inclusion were established.

2.2. Inclusion Criteria

As in both Domínguez et al. (2012) and Lenkeit et al. (2015) , a publication was retained in the review if it met three criteria: (1) the item was an original peer-reviewed article that was published in an academic journal, (2) the language of the publication was English, and (3) the article was "genuinely" concerned with PISA, meaning that PISA-related research formed part of the content and was not used merely as a source of citation. In addition, two exclusion criteria were determined. Self-identified PISA articles refer to research that uses part of the study's material for a small independent sample (e.g., Vidal-Abarca, et al., 2014). Such articles, as well as commentaries on previous articles published in an academic journal, were excluded from the database. Filtering articles in our database through these inclusion criteria was an important part of the screening process because it worked as a mechanism for quality control and helped to identify grey literature.

Although PISA is an international survey, articles published in languages other than English were excluded from the study for two reasons. First, English is the lingua franca of the academic community and is now well established as the global medium of academic publishing (Galloway & Rose, 2015; Mauranen, 2003). This reality is unlikely to change in the near future, regardless of arguments for or against the use of English as both a global language and an academic lingua franca. Second, non-English language PISA publications were found in at least 14 different languages, which raised logistical problems concerning the possibility of translation. Fortunately, some of the papers included English versions, and when a translation was located, it was included in the review. Non-peer-reviewed journal articles, conference proceedings, working papers, book chapters, and project reports were also excluded.

2.3. Categorisation Scheme, Data, and Measures

A database was created in SPSS and the final analytical sample of articles consists of 654 original peer-reviewed articles published in peer-reviewed journals reporting PISA-related research. A full

list of articles in the Endnote-database can be found in supplemental material Appendix A. The SPSS database with articles was categorized along the following variables:

- *Year of publication*: variable indicating in which year the article was published.
- *Country affiliation of the first author*: variable indicating the country in which the first author's institution is situated.
- *Journal name*: name of the journal the article was published in.
- *Journal discipline*: variable indicating the main disciplinary area of the journal the article was published in. These were established as educational research, educational psychology, sociology, comparative research, literacy/reading research, mathematics research, science research, economics, teaching and instruction, statistics, psychology, policy studies, assessment and testing, educational administration, law, health, information systems, and technology and interdisciplinary.
- *Thematic focus*: variable indicating which of the following three categories the article belonged to:

Secondary data analysis: research making use of the PISA database to address a specific research question

(1) Critique of PISA: research that either conceptually evaluates the developed constructs and those elaborated in the background questionnaire or that focus on statistical approaches and procedures applied to scales and constructs.
(2) Impact or policy study: research discussing the role of PISA in monitoring a country's performance in comparison to others or evaluates the impact of PISA national and international educational policies.
    - *Specific themes in secondary data analysis*: variable that distinguishes 16 specific themes: socioeconomic characteristics, teacher characteristics and instruction practices, immigration/language, bullying, systemic characteristics, age, ICT use, affective characteristics, leadership, learning strategies, gender, ability grouping, class size, environment, health, and a mixture of variables. Only articles classified as "secondary data analysis" in the nature variable had a valid value on this variable.
    - *Specific themes in critique of PISA*: variable that distinguishes six specific themes: construct, data processing, questionnaires, scaling procedures/Item Response Theory (IRT), sample, translation/language/interpretation of results and item features. Only articles classified as "critique of PISA" in the nature variable had a valid value on this variable.
    - *Specific themes in impact or policy study*: variable that distinguishes five specific themes: country performance, curriculum, national assessment, policy, and governance and media. Only articles classified as "impact or policy study" in the nature variable had a valid value on this variable.
    - *Cycle*: variable distinguishing whether an article uses data from or refers to the 2000, 2003, 2006, 2009, 2012 study or a combination of these.

2.4. Reliability Checks

We conducted a number of checks on the coding of the articles. First, we cross-checked for plausibility. For example, valid values on "specific topic in critique of PISA" must reflect the number of cases for the "critique of PISA" category on the nature variable. Second, we conducted inter-coder reliability checks. Of the total number of articles, 10% was randomly selected and given to three of our team members who were asked to classify articles on the elaborated variables on the basis of a codebook. Based on this check, 8.3, 1.1, and 0.7% of the values were corrected, which represented a high inter-coder reliability rate of over 90%, which is regarded as substantial for research purposes according to Shrout (1998) and Shrout and Fleiss (1979).

3. Results

3.1. Descriptive Results

The final EndNote library consists of 654 journal articles that matched the inclusion criteria. Domínguez et al. (2012) found 322 articles between 2002 and 2010, but our timeframe is broader and publication rates seem to be increasing in the past years. Even though Lindblad et al.'s (2015) review of 248 peer-reviewed PISA articles was published recently, the review surveyed used different key search terms compared to the present review. For example, the authors narrowed the PISA articles to those that had "PISA" and the word "education" mentioned in the abstract and, in addition, they used *three* search engine Scopus, JSTOR, and Discovery (p. 122). The present review has used five databases, ERIC, PsycINFO, Scopus, Web of Science, and Zetoc, and therefore the reviews might differ in the number of articles detected.

3.1.1. Frequency of published PISA articles over time, disciplines, and journals.

The number of publications of PISA-related research per year has been on the rise since 2002, with the largest number of publications (103) occurring in 2014 (see Figure 1). Articles published between January and September 2015 are excluded here as they are not representative of the whole year and would give the impression that the number of publications fell in 2015. While only one article per year was published from 1999 to 2001, seven articles were published in 2002 following the release of PISA's first cycle results in December 2001. The overall rise in the number of publications and the initial rise in 2002 both mirror the findings of Domínguez et al. (2012), which show both an overall rise and concomitant rises in publications following the release of each cycle's results.

The rise of published PISA articles is also reflected in the number of journals detected. Our sample covers a total of 251 journals, which were included within our criteria, (please see supplemental

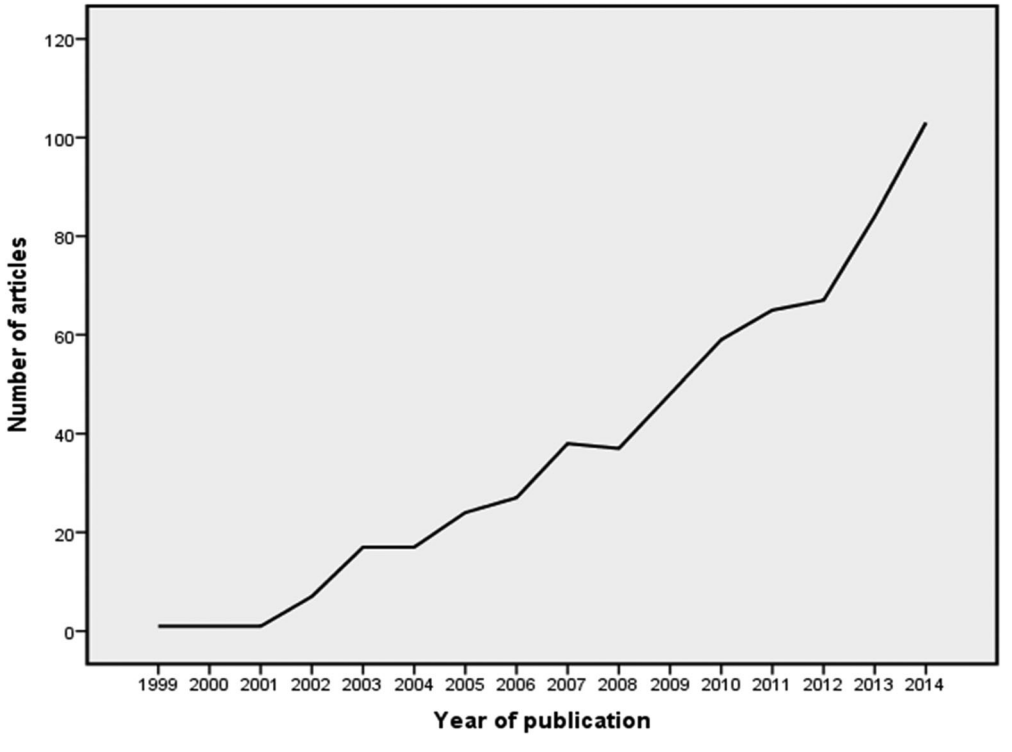

Figure 1. Publications per year.

material Appendix B for a full list of journals reviewed and Appendix C for a list of special issues). One of the noticeable findings is that only 17 journals feature more than 10 articles of PISA-related research (see Figure 2), 61 journals feature more than three articles, and an additional 45 journals contain only two articles. In comparison, 145 journals contain only one article.

The journal containing the most articles is *Educational Research and Evaluation* (21), followed by the *International Journal of Science Education* (20), the *European Educational Research Journal* (19), the *International Journal of Science and Mathematics Education* (18), and *Teachers College Record* (17).

With each journal having been coded to a corresponding discipline, articles are most frequently published in journals that are allocated to the expansive discipline of *Educational Research* (229). Significantly, the next most frequent discipline is that of *Economics* (59 articles), although the disciplines of *Assessment and testing* and of *Science research* are close in number, with 54 articles each. As the OECD is an organisation centred upon economic concerns, the prevalence of articles in the *Economics* discipline does not come as a complete surprise; however, it is noteworthy that even the discipline of *Assessment and Testing* has been surpassed by publications of an economic nature. Further disciplines include *Sociological Research* (38), *Comparative Research* (35), *Policy Studies* (33), *Teaching and Instruction* (31), *Psychology* (24), *Mathematics Research* (23), *Interdisciplinary Research* (15), *Information Systems and Technology* (13), *Educational Psychology* (12), *Statistics* (10), *Literacy/Reading Research* (10), *Educational Administration* (6), *Health* (2), *Law* (1), and *Other* (5).

Overall, publications most frequently referred to more than one cycle of PISA at a time. In publications referring to only one cycle, the 2006 and 2000 cycles were most prevalent, with 120 and 112 articles respectively. In contrast, only 9 articles refer to the PISA 2012 cycle alone; however, since the

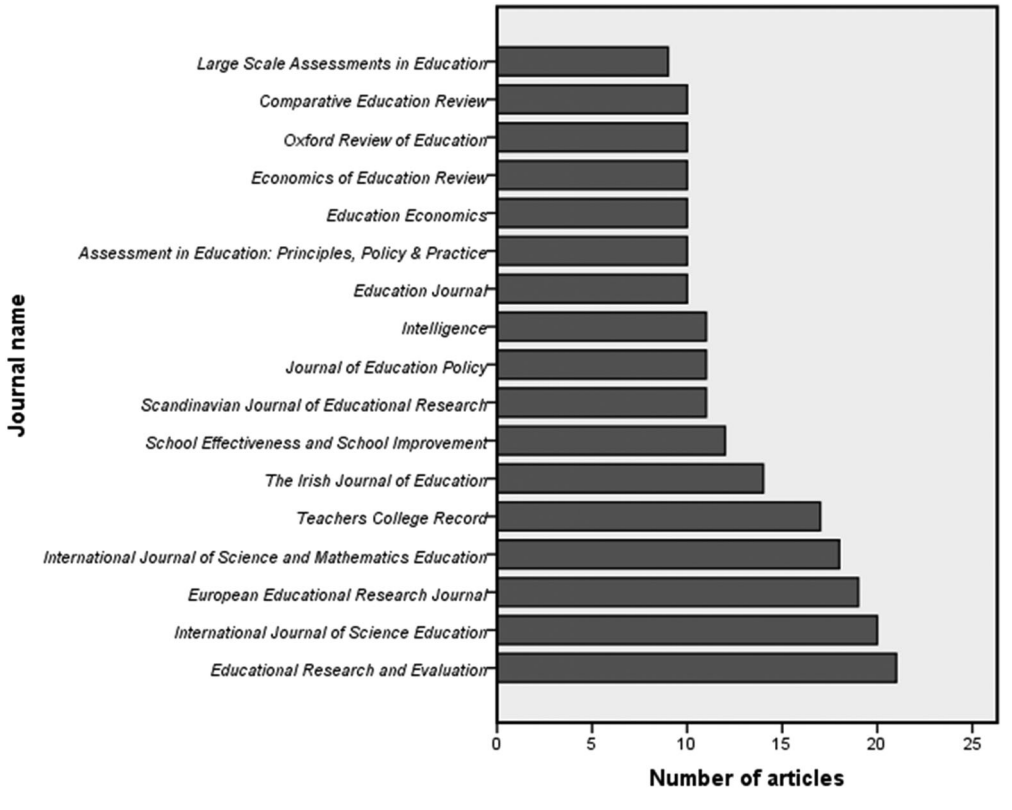

Figure 2. Publications with more than 10 articles.

results from the 2012 cycle were released in December 2013, future articles may be referring to the 2012 cycle more often. Additionally, the majority of articles (547) referred specifically to PISA alone, with only 107 articles referring to PISA in combination with another large-scale assessment such as TIMSS or PIRLS.

3.1.2. PISA articles across countries and disciplines

The five countries that have published the most articles on PISA are the USA (114), Australia (72), Germany (69), the UK (52), and Ireland (31). Regarding the greater number of publications from the USA, it is significant that over half of these publications (58 articles) appear between January 2013 and September 2015, suggesting that there has been a marked increase in attention towards PISA from the USA since the release of results from PISA 2012 (Sellar & Lingard, 2014) (Table 1).

The majority of publications from the five most frequently publishing countries are located in the discipline of *Educational Research*, a trend that can be observed across countries in the review. *Teaching and Instruction* and *Assessment and Testing* are the next two most frequently appearing disciplines in both the USA and Australia.

In Germany, however, the most frequently occurring discipline apart from general research that publishes PISA-related research in English is that of *Economics*, followed by *Policy Studies* and *Assessment and Testing* in equal numbers for third place. It should be noted, though, that economists tend to publish more frequently in English than the scholars in social sciences (particularly education science) and that an initial search in the German education-focused literature database yielded almost 1,200 German language articles. The UK also has *Economics* as its third most frequently occurring discipline, although the category of *Comparative Research* surpasses *Economics* by one article.

3.1.3. PISA articles and the three main research categories

As described in the methodology section, articles were classified into three categories: secondary data analysis, critique and impact/policy. The largest proportion of articles falls into the category of secondary data analysis (61.8%; 404 articles) (see Figure 3).

Articles in the critique (106) and impact/policy (144) categories are fewer in number, comprising less than half of total publications. The number of publications within each category rises on average over the years, with the most prevalent increase for publications categorised as secondary data analysis. Figure 3 shows how this leads to a widening gap between the frequencies of publications within the three categories.

Within the category of *Secondary data analysis*, approximately a quarter of publications (100 articles) address a mixture of topics including, for example, socioeconomic, school, class or affective variables such as students' self-efficacy and self-concept. Of *critique* publications, 40% of the articles

Table 1. Publication frequency: top five countries.

Discipline	USA	Australia	Germany	UK	Ireland
Educational research	29	23	18	18	22
Educational psychology	1	1	3	0	0
Sociology/social sciences	6	4	3	3	2
Comparative research	8	6	2	6	0
Literacy/reading	2	1	3	0	1
Mathematics research	5	2	2	1	1
Science research	10	9	2	4	0
Economics	5	3	11	5	2
Teaching and instruction	14	6	4	1	0
Psychology	6	1	3	2	3
Policy studies	5	1	6	3	0
Assessment and testing	13	10	6	4	0
Total number of articles	114	72	69	52	31

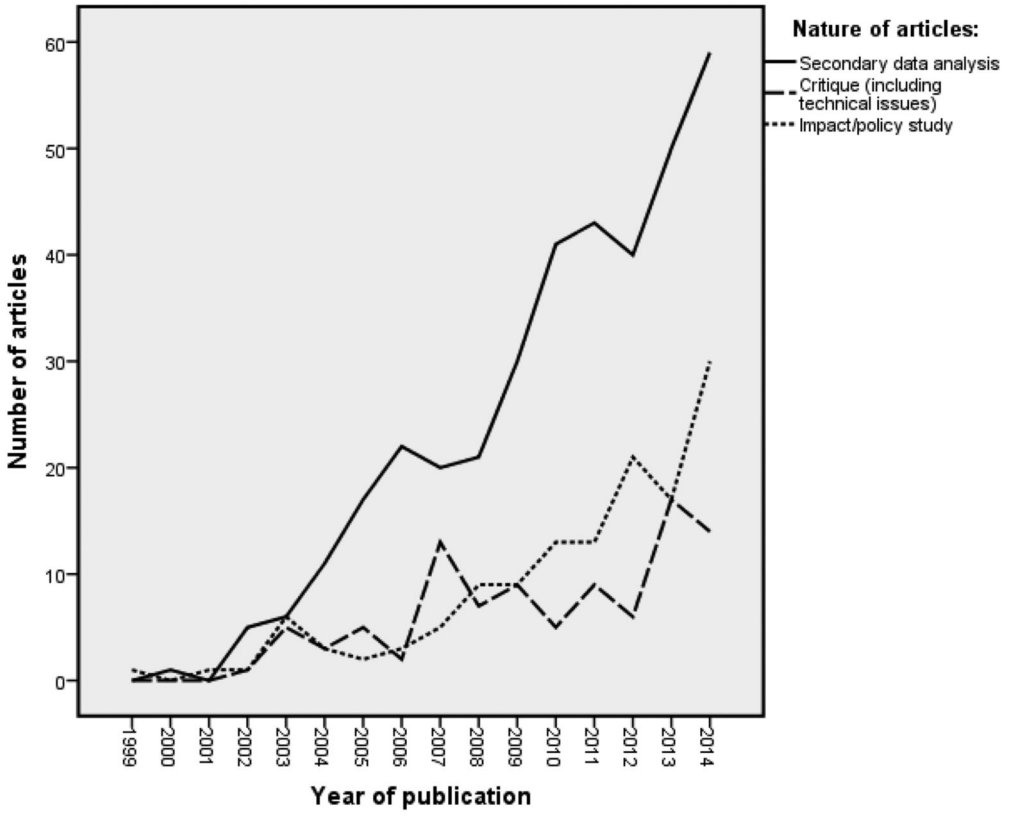

Figure 3. Nature of articles per year.

concern issues surrounding the nature of constructs, followed by 27% of articles upon topics such as scaling and IRT. Finally, 66% of the articles in the *impact/policy* category address issues of policy and governance, far outranking other topics such as curriculum and national assessment.

The following sections review a number of PISA-related research on topics in the categories of *secondary data analysis*, *critique* and *impact/policy studies*. Because the *secondary data analysis* category is so extensive, only the specific theme of inequalities related to socioeconomic background is reviewed. The choice seems justified as the theme represents one of the major concerns in educational research (Sirin, 2005) and receives substantial attention in PISA's design (see economic, social and cultural status [ESCS] construct) and reporting of findings. Themes in the *critique* and *impact/policy studies* categories are reviewed more comprehensively.

3.2. Results from the secondary data analysis category

3.2.1. Inequalities related to socio-economic status (SES)

A substantial number (109) of the articles classified as secondary data analysis (430 articles in this category in total) explored educational inequalities related to SES (Marks, 2006), migration or language use (Shapira, 2012), gender (Lietz, 2006) or ability grouping (Chmielewski, Dumont, & Trautwein, 2013). Of those, 52 specifically investigated inequalities related to SES disparities at student, school, community or system level. Many more publications, of course, use measures of SES as a control variable in their analyses, but articles discussed in this section specifically have inequalities related to SES as their thematic focus. Authors are particularly interested in the following aspects of SES-related inequalities.

3.2.1.1. SES gaps. Jerrim (2012) investigates the association of SES (represented by parental occupation) and reading achievement for different performance levels across Australia, Canada, Finland, Germany, the UK and the USA. He found that the association was stronger in the USA and the UK than in most other countries and that it is particularly strong at high performance levels. In similar research for Australia McConney and Perry (2010) and Perry and McConney (2010) asked whether the compositional SES effects (as measured by PISA's index of ESCS) were similar for students with different SES levels and found that while the compositional effect was substantial for Australian students, it was equally so across different levels of individual SES and subjects.

Cross-country research by Martins and Veiga (2010) and Oppedisano and Turati (2015) used a multidimensional SES measure including, for example, parents' education, books and home possessions. They, too, showed that SES gaps in maths achievement were particularly large in the UK, Germany, Belgium, Greece and Portugal but much lower in Sweden and Finland. They added that inequalities decreased in Germany and Spain but increased in France and Italy. Results also indicated that compositional SES effects were larger in some countries (e.g., Austria, Germany, Italy) than in others (e.g., Sweden, Ireland, Spain). In addition, Turmo (2004) found that the SES dimension drove inequalities in Scandinavia, while Polidano, Hanel and Buddelmeyer (2013) investigated the role of SES for school completion in Australia.

3.2.1.2. Systemic and institutional parameters. Many authors drew on systemic and institutional parameters in an attempt to enhance our understanding of the differences in SES-related inequalities across countries. In mainly EU cross-country comparisons they find that early forms of selection such as tracking, transition odds related to SES, grade repetition and choice in school selection distinguish countries with lower and higher levels of SES-related inequalities (Duru-Bellat & Suchaut, 2005; Gorard & Smith, 2004; Strakova, 2007). Le Donne (2014) added to these findings by identifying different forms of educational freedoms that magnified the compositional SES effect: early selection with numerous tracks, public selective schools and private schools with fees.

3.2.1.3. Urban-rural locale. More regional factors may also be related to SES gaps across different countries. Williams (2005), for example, investigated how rural and urban contexts related to achievement differences across industrialised nations. In most countries, a linear relationship between community size and achievement was found, which disappeared when accounting for SES, suggesting an association between locale and average SES. In Australia, too, students in urban schools achieved significantly higher outcomes than those in rural schools, but discussions by Pegg and Panizzon (2007) as well as Sullivan, Perry and McConney (2013) indicate that achievement differences persist even after accounting for SES measures due to, for example, differences in school resources and teacher recruitment to remote areas.

3.2.1.4. Family cultural capital. A number of articles investigated SES gaps, particularly focusing on the dimension of the family's cultural capital and evaluating its effects across a variety of countries. They found, for example, that parents' attitudes towards science were positively associated with science achievement across a variety of countries, but that those attitudes do not mediate the SES-achievement relationship (Perera, 2014). Others found that indicators of social and cultural communication practices between parents and students as well as cultural activities and possessions were consistently positively related to achievement, even after accounting for other SES characteristics (Hampden-Thompson, Guzman, & Lippman, 2013; Xu & Hampden-Thompson, 2012).

3.2.1.5. Family structure. Yet another topic related to SES gaps and frequently picked up in the literature is the family structure, often linked to economic, social and emotional deprivation. Comparing Western-European countries, Hampden-Thompson (2009) found that the achievement gap between students from single- and two-parent families was accounted for by the availability of economic resources. The effect of economic resources on this gap differs, however, across countries and

welfare regimes, with smaller effects in Scandinavian countries. No gap between students from single- and two-parent families was found, for example, in Greece. In this context De Lange, Dronkers and Wolbers (2014) added evidence that the percentage of single-parent families per school was negatively related to individual student achievement and that this association can only partly be explained by the school's SES composition.

Overall, PISA provides rich information to investigate inequalities related to SES from various perspectives – systemic characteristics, locale, family structure and characteristics – and many authors use this data to test theories related to family SES and academic achievement across different countries. Nevertheless, it is worth noting that authors used a range of indicators to represent SES in their research and thereby limit comparisons of findings on similar aspects. As Nonoyama-Tarumi (2008) points out, using PISA data, different composites of the SES measure will have different explanatory power within and across countries and studies thus might over- or underestimate SES effects in their analyses. Moreover, as research by Caro, Sandoval-Hernández and Lüdtke (2013) has shown, comparability of SES indicators is limited across different countries and thus points to careful interpretations in cross-national research.

3.3. Results From the PISA Critique Category

3.3.1. The cognitive test constructs

The effect of a curriculum-unrelated, literacy approach to defining the constructs in mathematics (De Lange, 2006; Ferrini-Mundy & Schmidt, 2005; Linneweber-Lammerskitten & Walti, 2005) and science (Bybee, McCrae, & Laurie, 2009; Huang, Wilson, & Wang, 2016; Kind, 2013; Lau, 2009; Olsen, 2004; Osborne, 2013) has been the subject of a small number of articles. More broadly, whether the PISA cognitive tests are measuring intelligence or academic literacy has been highly contested, with 54 articles published on this topic. Certainly, the tests have been shown to correlate highly at student level ($r = .88$) with intelligence tests. However, there is evidence of divergent validity for the cognitive tests, but there is also a high level of convergence with intelligence (Baumert, Ludtke, Trautwein, & Brunner, 2009). The OECD's definition of cognitive literacy is similar to Spearman's definition of general intelligence (Nyborg, 2007).

As PISA is really a test targeted at national level, rather than at student level, some have tried to draw conclusions about the abilities, in terms of cognitive literacy and intelligence, at the level of nations (Rindermann, 2007). Others have pointed out that the meaning of variables shifts when you aggregate to different levels; a conceptual, methodological point that is well-established in the field of multi-level modelling (e.g., Bosker, 2007; Flynn, 2007). In Brunner and Martin's (2007) terms, nations do not cognise.

3.3.2. Design, data processing and questionnaires

All PISA surveys are administered following standardised procedures to foster strong validity and high reliability of the assessments (see OECD, 2009a). The systematic training of a large number of test administrators and the systematic procedures followed ensure that students in all participating countries are subjected to the same testing conditions thereby providing solid grounds for within- and between-country comparisons. Evidence from Germany suggests that the standardising procedures followed by PISA are effective (Lüdtke, Robitzsch, Trautwein, Kreuter, & Ihme Marten, 2007).

PISA student and school questionnaires allow the contextualisation of student performances in participating countries (Rutkowski & Rutkowski, 2010). The questionnaire data allow the study of the impact of various cultural, social and economic factors on student achievement (Caro et al., 2013). Unfortunately however, up until 2012, PISA did not gather information about teachers, which is often seen as a limitation to investigating classroom-level characteristics. Since the 2015 cycle, however, countries can opt to administer a teacher questionnaire and therewith gather information, for example about teachers' qualifications and professional knowledge and teaching

practices (OECD, 2013). Nevertheless, Kaplan and McCarty (2013) used data fusion techniques to combine data from PISA and another large-scale assessment (Teaching and Learning International Survey) in the same country to synthesise a single data file with student-, school- and teacher-level information. Despite their recognised value amongst researchers, PISA questionnaires have been criticised for various reasons. Caro et al. (2013) raised concerns about the differential meaning of the cultural, social and economic constructs across the countries making cross-cultural comparisons difficult to establish and to some degree invalid. In addition, Hopfenbeck and Maul (2011) used cognitive interviews and statistical methods to find that the scales of self-report questionnaires in Norway led to invalid responses largely due to poor questionnaire design and language ambiguity. Rutkowski and Rutkowski (2010) argued that background questionnaires, namely the ones administered in non-OECD countries, suffered a number of weaknesses, such as missing data, possible respondent misinterpretation and low reliability on some of the scales. Poor questionnaire design has great implications for the validity of results given that background questionnaires are an integral part of development of achievement scores of the assessed samples. Policy makers should be aware of the limited interpretation that can be made based on flawed data resulting from weak questionnaires (Ercikan, Roth, & Asil, 2015; Rutkowski & Rutkowski, 2010).

Since the PISA 2006 cycle, student surveys have included scales assessing the levels of students' subject-specific motivation and interest. Baumert and Demmrich (2001) provided evidence that student performance in Germany is less affected by the level of stakes of the test than by the student interest in taking part in an international test and the value they attach to this involvement. The level of interest in engaging with assessments might be cultural and may vary across countries constituting thereby a source of bias (Hambleton, Merenda, & Spielberger, 2005). Contrarily, Butler and Adams (2007) reported that the effort students spent on PISA assessments is fairly stable across most countries and therefore support the validity of cross-cultural comparisons. Similar results have also been reported in the Nordic countries (Eklöf, Hopfenbeck, & Kjærnsli, 2012).

3.3.3. Technical issues

3.3.3.1. Sampling. The sampling frame includes criteria of eligibility and exclusion of students and schools to ensure optimal representation of the 15-year-old population within each country and limit bias. The OECD set a high threshold of response rate at 85% per country. Criticism of the PISA sampling frame adopted has been expressed since the early cycles of PISA. Scholars have debated OECD's age criterion as opposed to the grade level criterion adopted in TIMSS (McGaw, 2008; Prais, 2004; Wagemaker, 2008). Prais (2004) considers that using an age criterion for selecting participants is problematical because the sample would include a sizeable proportion of students who had either repeated a class or skipped one. That is, the sample would include students who are at various grade levels with different exposure to the curriculum compared to an average 15-year-old student in a particular country. Further, Wagemaker (2008) pointed out that the sampling frame adopted by PISA misses opportunities of investigating the impact of classroom- and teacher-specific variables on performance. Goldstein (Goldstein, Bonnet, & Rocher, 2007; Goldstein & Thomas, 2008) proposed the incorporation of longitudinal data to account for variability in grade retention and school starting age.

Other scholars commented on participation rates. Micklewright, Schnepf and Skinner's (2012) empirical study suggested that response rates are not a good proxy for data quality. Bias in the data was higher in the 2000 cycle in England when response rates were acceptable compared with the 2003 cycle where response rates were less acceptable. Prais (2003, 2004, 2007) cast serious doubts on the interpretation of results of PISA 2003 in England following a very low rate of participation (~60%) and a lack of sample representativeness. Another criticism of the sampling frame of PISA concerns the inclusion and exclusion criteria applied by PISA. Schuelka (2013) commented on the exclusion of students with disability from participating in international tests, including PISA, and warned that this procedure marginalised disabled students further and prevented them from taking part in any policy that related to educational equity.

3.3.3.2. Scaling/IRT. PISA items are scaled using a complex version of the Rasch model, which assumes a unidimensional trait underlying the scores. Kreiner and Christensen (2014) provided evidence that PISA reading data did not fit the model. Goldstein (2004) and Goldstein et al. (2007) showed that PISA data fit a multidimensional model with two factors better. In addition, other quantitative techniques have been suggested (e.g., Frey & Seitz, 2011).

In an incomplete design such as the one adopted in PISA, scores on items not sat by particular students are imputed based on the model. These data are called "plausible values" (Wu, 2005) and have been the centre of much controversy. Monseur and Adams (2009) showed that a hierarchical model created plausible values that led to a better estimation of between- and within-school variances than did the single-dimension model adopted in PISA. Kreiner and Christensen (2014) further raised concerns about imputing data for students in a particular subject when students have not taken a single item in that subject.

3.3.3.3. Measurement invariance and bias in PISA. PISA applies a very rigorous procedure of double translation and adaptation of tests from two source versions, English and French (Grisay, de Jong, Gebhardt, Breezier, & Halleux-Monseur, 2007). McQueen and Mendelovits (2003) described the thorough process of ensuring linguistic equivalence and cultural relevance of PISA reading assessment materials. Nevertheless, one of the main concerns about PISA is the extent to which the adapted forms are comparable to the source versions (English and French). Bias could emerge because of poor translation but could also be due to differences in language, culture, curriculum coverage and so on (Grisay et al., 2007; Nardi, 2008). Measurement invariance in different language versions of the same test could have serious implications on the validity of cross-lingual comparisons (Benítez & Padilla, 2014; Huang et al., 2016).

A considerable number of studies examined measurement invariance in cognitive and non-cognitive PISA assessments (Akour, Sabah, & Hammouri, 2015; Benítez & Padilla, 2014; Çetin, 2010; Chen & Jiao, 2014; Çıkrıkçı Demirtaşlı & Uluştaş, 2015; Grisay et al., 2007; Grisay & Monseur, 2007; Huang et al., 2016; Segeritz & Pant, 2013; Xie & Wilson, 2008; Yildirim & Berberoğlu, 2009; Yildirim, Yildirim, & Verhelst, 2014; Yildirim & Yildirim, 2011). Only three papers investigated measurement invariance in student self-report questionnaires (Benítez & Padilla, 2014; Çetin, 2010; Segeritz & Pant, 2013) compared to 11 publications examining measurement invariance in PISA cognitive surveys.

Bias was investigated using different approaches to differential item functioning (DIF): (1) factor analysis techniques (Çetin, 2010; Grisay et al., 2007; Grisay & Monseur, 2007; Segeritz & Pant, 2013; Yildirim & Berberoğlu, 2009) and (2) logistic regression (Oliveri & Ercikan, 2011; Xie & Wilson, 2008) and IRT (Chen & Jiao, 2014; Huang et al., 2016). Most of the studies reported a substantial amount of DIF when comparing different language versions. However, Oliveri conducted a number of studies on PISA problem solving tests in which she and her colleagues showed that, although DIF favours a language group at item level, at test level DIF balances out and results in high levels of comparability at test level (Oliveri & Ercikan, 2011; Oliveri, Olson, Ercikan, & Zumbo, 2012). This finding could be considered reassuring: despite significant DIF, the broad picture depicted in PISA results may be accurate.

Regardless of the method chosen to detect DIF, these techniques are unable to identify the source of bias in items. A number of studies attempted to understand the sources of DIF by adopting a mixed-method approach,; that is, by analysing more qualitatively items flagged as behaving differently across different groups (Benítez & Padilla, 2014; Huang et al., 2016; Yildirim & Berberoğlu, 2009). For instance, Benítez and Padilla (2014) carried out cognitive interviews with students (aged 15–16) from Spain and from the USA while Huang et al. (2016) conducted a content analysis of DIF items of PISA 2006 science tests.

3.3.3.4. Translation and language effects. Evidence supporting the differential effect of language on item difficulty in PISA has been provided by a number of studies (Arffman, 2010; Eivers, 2010;

Grisay et al., 2007; Grisay & Monseur, 2007; Roth, Ercikan, Simon, & Fola, 2015). Arffman (2010) reported mistranslations, lack of stylistic cues in and more boring Finnish adaptations of three PISA texts. Hatzinikita, Dimopoulos and Christidou (2008) argued that the low performance of Greek students in science could be due to a discrepancy in the linguistic modes (formal versus non-formal) employed in stimulus materials (texts, tables, graphs, figures, etc.) and items of PISA science assessments compared with Greek science textbooks.

Eivers (2010) reported different lengths of PISA test adaptations compared with the English source version: Finnish 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 to complete the test. Test speededness becomes a serious concern for some language groups such as German-speaking participants.

PISA instruments are more comparable across Western countries than they are across Middle Eastern or Asian countries (Grisay & Monseur, 2007; Grisay et al., 2007; Grisay, Gonzalez, & Monseur, 2009; Kankaraš & Moors, 2014). Empirical studies suggest that as target languages become less related to the source languages, the probability of bias in the adapted versions increases (Grisay et al., 2009). Hence, the German version would result in less bias than would the Arabic or Chinese version because of the proximity of the German language to the English language. Moreover, in the same study, Grisay et al. (2009) suggested that bias was higher in countries with a low gross domestic product. These relationships have been criticised by some scholars as the imposition of the Western model of education and test construction on the rest of the world (Bonnet, 2002).

3.3.3.5. Curriculum and culture curriculum fairness. The OECD claim that the scope of the national curriculum of participating countries would not represent a source of cultural or curriculum bias on the premise that PISA is not curriculum-based (OECD, 2009b). This claim is only valid in cases where the PISA framework and the national curriculum are commensurate, that is, they overlap completely (Nardi, 2008). Indeed, Huang et al. (2016) found that curriculum coverage constituted the main source of DIF in their study.

In addition to curriculum bias, Feniger and Lefstein (2014) argued in a study comparing the performance on PISA of Chinese pupils in Shanghai to that of their counterpart immigrants in New Zealand and Australia that cultural factors had a more significant effect on performance than curriculum exposure. Moreover, Zamir and Sabo (2012) point out the issue of cultural fairness in PISA when using texts that may be offensive to some religious groups such as ultra-orthodox pupils in Israel.

3.4. Results From the Policy and Impact Category

The 144 articles comprising the category of *policy and impact* focus predominantly upon the effect that PISA has had on policy and governance in a number of international contexts. In this category are also those publications analysing the potential factors, mechanisms, networks and dynamics driving PISA's influence on policy and governance. While a majority of articles in the category (94) addressed issues of policy and governance, a smaller number discussed PISA as related to curriculum (19), country performance (18), economics (7), the media (4) and national assessment (2). Issues of *policy and governance* were the most far-reaching across publications. These articles frequently consider the implications of policy borrowing, shifts in accountability structures and increasing demands for standardisation as a result of PISA.

While there are many publications comparing the effect of PISA upon a range of countries (e.g., Grek, 2009; Sellar & Lingard, 2013, 2014), there are also several articles presenting analyses upon individual country case studies. In Germany, for example, many articles following the 'PISA Shock' (Ertl, 2006) surrounding Germany's performance in PISA's first cycle focused on the ramifications of PISA upon Germany alone. Discussion concerning reforms occurring in Germany post-

PISA (Di Fuccia, Witteck, Markic, & Eilks, 2012; Hartong, 2012; Klemm, 2003) have also included debate regarding the nature of educational standards previously upheld by the German educational system and those standards now being introduced via PISA-driven reform (Neumann et al., 2012; Sünker, 2003).

Although the impact of PISA may differ across national and even subnational contexts, there are overarching elements identified within the literature as potential mechanisms behind PISA's ability to generate impact. A significant point of consensus occurs as multiple authors (e.g., Afonso & Costa, 2009; Bieber & Martens, 2011) suggest that the OECD (through PISA) employs mechanisms of *soft policy*, whereby the status of the OECD as an international organisation allows for powerful yet indirect governance to influence policy-making at the national level. Further, the OECD's promotion of the usage of PISA data upholds the desire for data-driven policy-making at the national level, encouraging systems of weak governance relying upon external authorities such as the OECD for knowledge production and policy guidance (Grek, 2010; Meyer, 2014).

Some nascent discussion exists regarding PISA's influence on curricular standards in science (e.g., Dall, 2011; Duit, 2007), which is particularly intriguing given PISA's non-curricular and literacy-oriented nature of assessment. A similar discussion of PISA's impact on mathematics and reading standards is not yet extant. Similarly, only two publications (Hendrickson, 2014; O'Mara, 2014) wholly discuss evidence of impact from PISA upon national assessment, although there is sometimes difficulty in extricating the topic of national assessment from the wider expanses of policy shifts elaborated in many articles. In countries such as Switzerland (Bieber & Martens, 2011), Germany (Ertl, 2006), and Norway (Møller & Skedsmo, 2013) where assessment systems have been introduced since the onset of PISA, very few articles discuss the significance of such changes in regions previously lacking country-wide standardised assessment. The discussion of PISA's policy implications potentially extending to the level of national assessment ranges widely, but the literature lacks a similarly extensive exploration of changes induced across national assessment systems as a result of PISA. Respective discussions might, however, be held in national-language contexts that are not captured in this review.

4. Discussion of Results

In this article we intended to provide a comprehensive bibliographical synthesis of PISA-related research in academic journals that provides information on how it evolved over the years as well as across disciplines and countries. We also aimed to provide summaries of the main aspects that PISA-related research is concerned with.

Despite the fact that our review concentrated upon articles published in English, Germany is among the top three countries with highest publication frequency, with only Australia and the USA publishing more research on PISA. England and Ireland follow Germany as the fourth and fifth countries. These results echo the findings on publication patterns from the PIRLS review (Lenkeit et al., 2015), where Germany published most research, followed by the USA. One possible explanation of the high publication frequency of German researchers could be the development and support of research in Germany after the 'PISA shock' (Ertl, 2006). Several policy initiatives were introduced, including the 2004 creation of the Institute for Educational Progress based at the Humboldt University in Berlin with the aim to:

> … provide the infrastructure and scientific capacity needed to support the development of the standards and assessments the new monitoring system would need, and to gather, analyse and disseminate the resulting information. (OECD, 2011, p. 211)

More recently, in 2010 the Centre for International Student Assessment (Zentrum für Internationale Bildungsvergleichsstudien, ZIB) was established to unite three prominent institutions of German educational research[2] with the explicit aim to "represent the excellent German presence on the international stage of educational research in the area of large-scale assessments" (http://zib.education/en/home.html). This importance of such a research infrastructure and capacity is apparent in all

three of the top publishing countries. The USA, Germany, and Australia have all been involved in the development of PISA through their different research centres and research companies. In the USA both the Educational Testing Service and Westat have been involved since PISA's inception, while in Australia the Australian Council of Educational Research, which led the first consortium for PISA 2000, has continued to play a major role in the subsequent PISA cycles. Much of the expertise supporting the PISA test is therefore naturally found in these countries.

Lenkeit et al. (2015) suggested that differences among countries in publication frequencies in PIRLS articles could be linked to the different research traditions. Norway has historically not focused upon quantitative data analysis in educational research, while the countries such as USA and England have a longer research tradition in testing and evaluation (Lenkeit et al. 2015), and that could be one of many explanations for the publication frequency in these countries.

Thematically, the review shows that most articles fall into the secondary data-analysis category, in which socioeconomic-related inequalities present one major theme. The respective summary indicates that PISA provides a range of information to study educational inequalities within and across countries from various perspectives and that researchers are making good use of this data to identify the structures of inequalities. Less discussed seem to be, however, the processes behind those inequalities or how they differ across different cultures. Instead, some scholars cast doubt upon the cross-national comparability of SES indicators and therewith caution against the use of data provided in PISA to investigate those inequalities further.

This review also highlights main controversies around PISA test design, result analysis, and score interpretation. Given the strong impact of PISA results on many educational systems across the participating economies, the strong critique of PISA instruments ensures the OECD is kept under the scrutiny of academic expertise and provides substantial feedback for improving test design and data analysis techniques. While a large number of studies investigated measurement invariance in cognitive PISA assessments, there has been less interest in examining measurement invariance in PISA student questionnaires. In many of these studies, language was claimed to be a potential source of bias disadvantaging some countries, particularly countries in which the national language is very different from PISA source languages (e.g., Arabic, Chinese, Japanese, etc.). Future research should focus on providing evidence supporting such claims.

There exists a somewhat unsettling divide across the articles in the three different categories of the review : the majority of articles, those in *secondary data analysis*, use PISA data as a foundation from which to build additional levels of newly constructed knowledge to bolster even further the already vast amount of information generated by PISA alone. Simultaneously, a substantial number of articles in both the *critique* and *impact/policy* categories are warning policy-makers and researchers alike to be cautious about using PISA data as a means for valid comparison or for informed policy-making. As a result, the surveyed literature appears conflicted: the authors of *secondary data analysis* publications are often building upon PISA data, and the *critique* and *impact/policy* authors pointing out structural weaknesses and cracks in the foundations of ongoing PISA constructions.

There is a general consensus on the informative value of PISA data at the national and international levels (e.g., Ercikan et al., 2015; McGaw, 2008; Shiel & Eivers, 2009; Sireci, 2015). With the sheer number of related publications PISA has undoubtedly contributed to the advancement of educational research, if not within national boundaries, surely in the field of cross-national comparative educational research. Nevertheless, major criticism has been expressed in relation to the way results have been interpreted and used to introduce educational reforms (Dancis, 2014; Ercikan et al., 2015; Leung, 2014; Shiel & Eivers, 2009). Every cycle, league tables published by OECD reduced PISA to a "horse-race" with "winners and losers" (De Lange, 2006, p. 29). Sweeping conclusions about the quality and effectiveness of countries' education systems based on country ranking have been rejected (Ercikan et al., 2015; Leung, 2014; Sireci, 2015; Tienken, 2014).

[2]ZIB is a network of the Deutsches Institut für Internationale Pädagogische Forschung, Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik and the School of Education der Technischen Universität München.

No doubt, PISA provides us with a rich and varied dataset, but it must be remembered that the results we derive will always be dependent on the methodology used and the assessment design adopted (Lafontaine & Monseur, 2009). Learning from PISA to draw valid inferences will continue to necessitate exercising a great amount of wisdom in our research (Leung, 2014; Sireci, 2015). Thus, with the upcoming release of the PISA 2015 results by the end of 2016, it will be worthwhile to reflect upon the knowledge that PISA has already provided the research community with as well as the persistent critiques identified around how this knowledge is technically generated.

4.1. Limitations and Future Research

There are several limitations in our research review that need to be addressed. First of all, we accept the limitation of only reviewing articles in English. The research team is aware of publications in languages such as German, Norwegian, Swedish, Turkish, and Chinese that would have offered a broader perspective on PISA research. We initially considered including non-English language articles, that within our team we would be able to understand. This would, however, be an arbitrary choice and there is no good justification for including, for example, articles written in Spanish but not those written in Russian.[3] Thus, including articles in English seems justified as it is the most far-reaching language in research (Galloway & Rose, 2015; Mauranen, 2003). A future research project could be to co-operate with language experts to publish a PISA review including more languages. Secondly, we did not include "grey literature" on PISA, such as policy documents, OECD reports, briefings, books, and book chapters. We are aware of countries that publish research on PISA data such as the Nordic countries (Lie, Linnakylä, & Roe, 2003), and we acknowledge that PISA may have more impact on countries educational research than our review has been able to detect (Breakspear, 2012). A further limitation is that we did not conduct a detailed inspection of all the articles in the secondary analysis category, but limited the readings to the article on SES. A more in-depth analysis of the remaining articles in this category would offer a more comprehensive overview of how PISA data is used by quantitative analysts, for example to investigate themes on school leadership and motivational aspects. Finally, we also acknowledge that our review does not specifically discuss the optional PISA domains such as financial literacy.

A further limitation is that the categorisation of articles' features into the variables remains to some extent a matter of judgement. We addressed this challenge by conducting reliability checks but different scholars will look at articles from different perspectives depending on their specific research focus and approaches.

Nevertheless, the paper offers an introduction to the PISA study for those who are new to it, as well as an overview of English peer-reviewed papers for researchers already engaging with the PISA data. In addition to the descriptive results, we have outlined the main research categories, which may be helpful for researchers in the field and which we hope will motivate researchers to take advantage of the broad databases offered and explore topics such as school climate, students' approaches to learning, and assessment cultures in schools, all of which have been less explored by PISA researchers.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Kunnskapssenter for utdanning, Norway.

[3]An initial search in the German education focused, literature catalogue (*FIS Bildung*) identified almost 1,200 German-language articles between 1999 and 2015, when the search was conducted with PISA entered in the keyword section. This indicates, not only for Germany, the extent to which PISA-related research is discussed within national-linguistic contexts.

References

Afonso, N., & Costa, E. (2009). A influência do PISA na decisão política em Portugal: o caso das políticas educativas do XVII Governo Constitucional Português [The influence of PISA on the political decision in Portugal: the case of the educational policies of the XVII Portuguese Constitutional Government]. *Sísifo. Revista de Ciências da Educação, 10*, 53–64.

Akour, M., Sabah, S., & Hammouri, H. (2015). Net and global differential item functioning in PISA polytomously scored science items: Application of the differential step functioning framework. *Journal of Psychoeducational Assessment, 33*(2), 166–176. Retrieved from http://jpa.sagepub.com/cgi/content/abstract/33/2/166

Arffman, I. (2010). Equivalence of translations in international Reading literacy studies. *Scandinavian Journal of Educational Research, 54* (1), 37–59.

Baird, J., Johnson, S., Hopfenbeck, T. N., Isaacs, T., Sprague, T., Stobart, G., & Yu, G. (2016). On the supranational spell of PISA in policy. *Educational Research, 58*(2), 121–138. Special Issue: International Policy Borrowing and Evidence-based Educational Policy Making: Relationships and Tensions.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*, 441–462.

Baumert, J., Ludtke, 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*, 165–176.

Benítez, I., & Padilla, J.-L. (2014). Analysis of nonequivalent assessments across different linguistic groups using a mixed methods approach: Understanding the causes of differential item functioning by cognitive interviewing. *Journal of Mixed Methods Research, 8*(1), 52–68. Retrieved from http://mmr.sagepub.com/cgi/content/abstract/8/1/52

Bieber, T., & Martens, K. (2011). The OECD PISA study as a soft power in education? Lessons from Switzerland and the USA. *European Journal of Education, Research, Development and Policy, 46*(1), 101–116.

Bonnet, G. (2002). Reflections in a critical eye: On the pitfalls of international assessment. *Assessment in Education: Principles, Policy & Practice, 9*(3), 387–399. http://dx.doi.org/10.1080/0969594022000027690a

Bosker, R. J. (2007). The different levels of the *g*-factor. *European Journal of Personality, 21*, 712–714.

Breakspear, S. (2012). The policy impact of PISA: An exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers*. doi:10.1787/5k9fdfqffr28-en.

Brunner, M., & Martin, R. (2007). Not every *g* is *g*. *European Journal of Personality*, *21*, 714–716.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement, 8* (3), 279–304.

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

Caro, D. H., Sandoval-Hernández, A., & Lüdtke, O. (2013). Cultural, social, and economic capital constructs in international assessments: An evaluation using exploratory structural equation modeling. *School Effectiveness and School Improvement*, *25*(3), 433–450. doi:10.1080/09243453.2013.812568

Çetin, B. (2010). Cross-cultural structural parameter invariance on PISA 2006 student questionnaires. *Eurasian Journal of Educational Research, 38*, 71–89.

Chen, Y.-F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 Reading assessment. *Educational Assessment, 19*(2), 77–96. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/10627197.2014.903650

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50*(5), 925–957. doi:10.3102/0002831213489843

Çıkrıkçı Demirtaşlı, N., & Uluştaş, S. (2015). A study on detecting of differential item functioning of PISA 2006 science literacy items in Turkish and American samples. *Eurasian Journal of Educational Research, 58*, 41–60.

Dall, A. (2011). Is PISA counter-productive to building successful educational systems? *Social Alternatives, 30*(4), 10–14.

Dancis, J. (2014). What does the international PISA math test really tell Us? *AASA Journal of Scholarship and Practice, 10*(4), 31–42.

De Lange, J. (2006). Mathematical literacy for living from OECD-PIA perspective. *Tsukuba Journal of Educational Study in Mathematics, 25*, 13–35.

De Lange, M., Dronkers, J., & Wolbers, M. H. J. (2014). Single-parent family forms and children's educational performance in a comparative perspective: Effects of school's share of single-parent families. *School Effectiveness and School Improvement,* 25(3), 329–350. doi:10.1080/09243453.2013.809773

Di Fuccia, D., Witteck, T., Markic, S., & Eilks, I. (2012). Trends in practical work in German science education. *Eurasia Journal of Mathematics, Science and Technology Education, 8*(1), 59–72.

Domínguez, M., Vieira, M.-J., & Vidal, J. (2012). The impact of the programme for international student assessment on academic journals. *Assessment in Education: Principles, Policy and Practice, 19*(4), 393–409.

Duit, R. (2007). Science education research internationally: Conceptions, research methods, domains of research. *Eurasia Journal of Mathematics Science and Technology Education, 3*(1), 3–15.
Duru-Bellat, M., & Suchaut, B. (2005). Organisation and context, efficiency and equity of educational systems: What PISA tells us. *European Educational Research Journal, 4*(3), 181–194.
Eivers, E. (2010). PISA: Issues in implementation and interpretation. *Irish Journal of Education/Iris Eireannach an Oideachais, 38*, 94–118. Retrieved from http://www.jstor.org/stable/20789130
Eklöf, H., Hopfenbeck, T. N., & Kjærnsli, M. (2012). Hva vet vi om elevers testmotivasjon? Erfaringer fra internasjonale og nasjonale undersøkelser i Norge og Sverige, p. 84–96 [What do we know about students' test motivation? Experiences of international and national tests in Norway and Sweden]. In T. N. Hopfenbeck, M. Kjaernsli, & R. V. Olsen (Eds.), *Kvalitet i norsk skole. Internasjonale og nasjonale undersøkelser av læringsutbytte og undervisning* [Quality in the Norwegian school. International and national tests of learning outcomes and teaching] (pp. 84–96). Oslo: Universitetsforlaget.
Ercikan, K., Roth, W.-M., & Asil, M. (2015). Cautions about inferences from international assessments: The case of PISA 2009. *Teachers College Record, 117*, 1–28.
Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education, 32* (5), 619–634.
Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: An ironic investigation. *Journal of Education Policy, 29* (6), 845–855.
Ferrini-Mundy, J., & Schmidt, W. H. (2005). International comparative studies in mathematics education: Opportunities for collaboration and challenges for researchers. *Journal for Research in Mathematics Education, 36*(3), 164–175.
Flynn, J. (2007). What lies behind *g*(I) and *g*(ID). *European Journal of Personality, 21*, 722–724.
Frey, A., & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment. *Educational and Psychological Measurement, 71*(3), 503–522. Retrieved from http://epm.sagepub.com/cgi/doi/10.1177/0013164410381521
Galloway, N., & Rose, H. (2015). *Introducing global Englishes*. New York: Routledge.
Goldstein, H. (2004). Measuring educational standards. *Significance*, *1*, 103–105. doi:10.1111/j.1740-9713.2004.00039.x
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.
Goldstein, H., & Thomas, S. M. (2008). Reflections on the international comparative surveys debate. *Assessment in Education: Principles, Policy & Practice, 15*(3), 215–222.
Gorard, S., & Smith, E. (2004). An international comparison of equity in education systems. *Comparative Education, 40*(1), 15–28. doi:10.1080/0305006042000184863
Gough, D., Oliver, S., & Thomas, J. (2012). *An introduction to systematic reviews*. London: Sage.
Grek, S. (2009). Governing by the numbers: The PISA "effect" in Europe. *Journal of Education Policy, 24*(1), 23–37.
Grek, S. (2010). International organisations and the shared construction of policy "problems': Problematisation and change in education governance in Europe. *European Educational Research Journal, 9*(3), 396–406.
Grisay, A., Gonzalez, E., & Monseur, C. (2009). Equivalence of item difficulties across national versions of the PIRLs and PISA Reading assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 2*, 63–83.
Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Breezier, 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.
Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Associates.
Hampden-Thompson, G. (2009). Are two better than one? A comparative study of achievement gaps and family structure. *Compare: A Journal of Comparative and International Education,* 39(4), 517–534. doi:10.1080/03057920802366372
Hampden-Thompson, G., Guzman, L., & Lippman, L. (2013). A cross-national analysis of parental involvement and student literacy. *International Journal of Comparative Sociology,* 54(3), 246–266. doi:10.1177/0020715213501183
Hartong, S. (2012). Overcoming resistance to change: PISA, school reform in Germany and the example of lower Saxony. *Journal of Education Policy, 27* (6), 747–760.
Hatzinikita, V., Dimopoulos, K., & Christidou, V. (2008). PISA test items and school textbooks related to science: A textual comparison. *Science Education, 92*(4), 664–687.
Hendrickson, K. A. (2012). Learning from Finland: Formative assessment. *The Mathematics Teacher, 105*(7), 488–489.
Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*(2), 95–121. doi:10.1080/15305058.2010.529977
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: An International Journal of Experimental Educational Psychology*, *36*(2), 378–390. doi:10.1080/01443410.2014.946890

Jerrim, J. (2012). The socio-economic gradient in teenagers' reading skills: How does England compare with other countries? *Fiscal Studies, 33*(2), 159–184.
Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-scale Assessments in Education, 1*(6). Retrieved from http://www.largescaleassessmentsineducation.com/content/1/1/6
Kankaraš, M., & Moors, G. (2014). Analysis of cross-cultural comparability of PISA 2009 scores. *Journal of Cross-Cultural Psychology, 45*(3), 381–399. Retrieved from http://jcc.sagepub.com/cgi/doi/10.1177/0022022113511297
Kind, P. M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education,* 97(5), 671–694.
Klemm, K. (2003). If no reform then at least no rollback. *European Education, 35*(4), 34–43.
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.
Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal, 8*(1), 69–79.
Lau, K.-C. (2009). A critical examination of PISA's assessment on scientific literacy. *International Journal of Science and Mathematics Education, 7*, 1061–1088.
Le Donne, N. (2014). European variations in socioeconomic inequalities in students' cognitive achievement: The role of educational policies. *European Sociological Review, 30*(3), 329–343. doi:10.1093/esr/jcu040
Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J.-A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review, 16*, 102–115.
Leung, F. K. S. (2014). What can and should we learn from international studies of mathematics achievement? *Mathematics Education Research Journal, 26*, 579–605.
Lie, S., Linnakylä, P., & Roe, A. (Eds.). (2003). *Northern lights on PISA: Unity and diversity in the Nordic countries*. Oslo: Department of Teacher Education and School Development, University of Oslo.
Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation*, 32(4), 317–344. doi:10.1016/j.stueduc.2006.10.002
Lindblad, S., Pettersson, D., & Popkewitz, T. S. (2015). *International comparisons of school results: A systematic review of research on large scale assessments in education*. Stockholm: Swedish Research Council.
Linneweber-Lammerskitten, H., & Walti, B. (2005). Is the definition of mathematics as used in the PISA assessment framework applicable to the HarmoS project? *ZDM, 37*(5), 402–407.
Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F., & Ihme Marten, J. (2007). Are there test administrator effects in large-scale educational assessments? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(4), 149–159.
Marks, G. N. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research, 48*(1), 21–40. doi:10.1080/00131880500498396
Martins, L., & Veiga, P. (2010). Do inequalities in parents' education play an important role in PISA student mathematics achievement test score disparities? *Economics of Education Review, 29*(6), 1016–1033. doi:10.1016/j.econedurev.2010.05.001
Mauranen, A. (2003). The corpus of English as *lingua franca* in academic settings. *TESOL Quarterly, 37* (3), 513–527.
McConney, A., & Perry, L. B. (2010). Science and mathematics achievement in Australia: The role of school socioeconomic composition in educational equity and effectiveness. *International Journal of Science and Mathematics Education, 8*(3), 429–452.
McGaw, B. (2008). The role of the OECD in international comparative studies of achievement. *Assessment in Education: Principles, Policy & Practice, 15*(3), 223–243. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/09695940802417384
McQueen, J., & Mendelovits, J. (2003). PISA Reading: Cultural equivalence in a cross-cultural study. *Language Testing, 20* (2), 208–224.
Meyer, H. D. (2014). The OECD as pivot of the emerging global educational accountability regime: How accountable are the accountants?. *Teachers College Record*, *116*(9), 1–20.
Meyer, H.-D., & Benavot, A. (2013). *PISA, power, and policy: The emergence of global educational governance*. Didcot, UK: Symposium Books.
Micklewright, J., Schnepf, S. V., & Skinner, C. (2012). Non-response biases in surveys of schoolchildren: The case of the English programme for international student assessment (PISA) samples. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(4), 915–938.
Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 320–334.
Møller, J., & Skedsmo, G. (2013). Modernising education: New public management reform in the Norwegian education system. *Journal of Educational Administration and History, 45*(4), 336–353.
Nardi, E. (2008). Cultural biases: A non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice, 15*(3), 259–266.

Neumann, E., Kiss, A., Fejes, I., Bajomi, I., Berenyi, E., Biro, Z. A., & Vida, J. (2012). The hard work of interpretation: The national politics of PISA reception in Hungary and Romania. *European Educational Research Journal, 11*(2), 227–242.

Nonoyama-Tarumi, Y. (2008). Cross-national estimates of the effects of family background on student achievement: A sensitivity analysis. *International Review of Education,* 54(1), 57–82. doi:10.1007/s11159-007-9069-5

Nyborg, H. (2007). Do recent large-scale cross-national student assessment studies neglect general intelligence *g* for political reasons? *European Journal of Personality, 21*, 739–741.

OECD. (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris: OECD.

OECD. (2000). *Measuring student knowledge and skills: The PISA 2000 assessment of Reading, mathematical and scientific literacy*. Paris: OECD.

OECD. (2009a). *PISA 2006 technical report*. Paris: OECD.

OECD. (2009b). *PISA data analysis manual*. Paris: OECD.

OECD. (2011). *Strong performers and successful reformers in education lessons from PISA for the United States*. Paris: OECD.

OECD. (2013). *PISA*

2015 draft questionnaire framework. Paris: OECD.

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

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203–223.

Olsen, R. V. (2004). The search for descriptions of students' thinking and knowledge: Exploring nominal cognitive variables by correspondence and homogeneity analysis. *Scandinavian Journal of Educational Research, 48*(3), 325–341. doi:10.1080/00313830410001695763

O'Mara, J. (2014). Closing the emergency facility: Moving schools from literacy triage to better literacy outcomes. *English Teaching: Practice and Critique, 13*(1), 8–23.

Oppedisano, V., & Turati, G. (2015). What are the causes of educational inequality and of its evolution over time in Europe? Evidence from PISA. *Education Economics,* 23(1), 3–24. doi:10.1080/09645292.2012.736475

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

Ozga, J. (2012). Introduction. Assessing PISA. *European Educational Research Journal, 11*(2), 166–171.

Pegg, J., & Panizzon, D. (2007). Inequities in student achievement for literacy: Metropolitan versus rural comparisons. *Australian Journal of Language and Literacy, 30*(3), 177–190.

Perera, L. D. H. (2014). Parents' attitudes towards science and their children's science achievement. *International Journal of Science Education,* 36(18), 3021–3041. doi:10.1080/09500693.2014.949900

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record, 112*(4), 1137–1162.

Polidano, C., Hanel, B., & Buddelmeyer, H. (2013). Explaining the socio-economic status school completion Gap. *Education Economics, 21*(3), 230–247. doi:10.1080/09645292.2013.789482

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

Prais, S. J. (2004). Cautions on OECD's recent educational survey (PISA): Rejoinder to OECD's response. *Oxford Review of Education, 30*(4), 569–573.

Prais, S. J. (2007). Two recent (2003) international surveys of schooling attainments in mathematics: England's problems. *Oxford Review of Education, 33*(1), 33–46.

Rindermann, H. (2007). The *g*-factor of international cognitive ability comparisons: The homogeneity of results in PISA, TIMSS, PIRLS and IQ tests across nations. *European Journal of Personality, 21*, 667–706.

Roth, W., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *The Journal of Mathematical Behavior*, *40*, 88–105.

Rutkowski, L., & Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies, 42*(3), 411–430.

Schuelka, M. J. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy, 28*(2), 216–230.

Segeritz, M., & Pant, H. A. (2013). Do they feel the same way about math? Testing measurement invariance of the PISA "students' approaches to learning" instrument across immigrant groups within Germany. *Educational and Psychological Measurement, 73*(4), 601–630. Retrieved from http://epm.sagepub.com/cgi/doi/10.1177/0013164413481802

Sellar, S., & Lingard, B. (2013). The OECD and global governance in education. *Journal of Education Policy, 28*(5), 710–725.

Sellar, S., & Lingard, B. (2014). The OECD ad the expansion of PISA: New global modes of governance in education. *British Educational Research Journal, 40*(6), 917–936.

Shapira, M. (2012). An exploration of differences in mathematics attainment among immigrant pupils in 18 OECD countries. *European Educational Research Journal, 11*(1), 68. doi:10.2304/eerj.2012.11.1.68

Shiel, G., & Eivers, E. (2009). International comparisons of reading literacy: What can they tell us? *Cambridge Journal of Education, 39*(3), 345–360.
Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*(3), 301–317.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.
Sireci, S. (2015). Beyond ranking of nations: Innovative research on PISA. *Teachers College Record, 117*, 1–8.
Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research,* 75(3), 417–453.
Strakova, J. (2007). The impact of the structure of the education system on the development of educational inequalities in the Czech Republic. *Sociologicky Casopis-Czech Sociological Review, 43*(3), 589–610.
Sullivan, K., Perry, L. B., & McConney, A. (2013). How do school resources and academic performance differ across Australia's rural, regional and metropolitan communities? *Australian Educational Researcher*, *40*(3), 353–372.
Sünker, H. (2004). Education and reproduction of social inequality: German politics and sociology of education. *Policy Futures in Education*, *2*(3-4), 593–606.
Tienken, C. H. (2014). PISA problems. *AASA Journal of Scholarship and Practice, 10*(4), 4–18.
Turmo, A. (2004). Scientific literacy and socio-economic background among 15-year-olds: A nordic perspective. *Scandinavian Journal of Educational Research,* 48(3), 287–305. doi:10.1080/00313830410001695745
Vidal-Abarca, E., Gilabert, R., Ferrer, A., Ávila,V., Martínez, T., Mañá, A., Llorens, A.-C., Gil, L., Cerdán, R., Ramos, L., & Serrano, M.-A., (2014). TuinLEC, an intelligent tutoring system to improve reading literacy skills / TuinLEC, un tutor inteligente para mejorar la competence lector. *Journal for the Study of Education and Development*, *37*(1), 25–56.
Wagemaker, H. (2008). Choices and trade-offs: Reply to McGaw. *Assessment in Education: Principles, Policy & Practice, 15*(3), 267–278. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/09695940802417491
Williams, J. H. (2005). Cross-national variations in rural mathematics achievement: A descriptive overview. *Journal of Research in Rural Education, 20*(5), 1–18.
Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*, 114–128.
Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly, 50*(3), 403–416.
Xu, J., & Hampden-Thompson, G. (2012). Cultural reproduction, cultural mobility, cultural resources, or trivial effect? A comparative approach to cultural capital and educational performance. *Comparative Education Review, 56*(1), 98–124.
Yildirim, H. H., & Berberoğlu, G. (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9*(2), 108–121.
Yildirim, H. H., & Yildirim, S. (2011). Correlates of communalities as matching variables in differential item functioning analyses. *Hacettepe University Journal of Education, 40*, 386–396.
Yildirim, H. H., Yildirim, S., & Verhelst, N. (2014). Profile analysis as a generalized differential item functioning analysis method. *Education and Science, 39*(172), 49–64.
Zamir, S., & Sabo, H. (2012). PISA assessment: The problematic issue of administrating PISA science literacy survey to ultra-orthodox pupils in Israel, 2006. *Acta Didactica Napocensia, 5*(3), 43–48.