Trend Analysis with International Large-Scale Assessments

30

Past Practice, Current Issues, and Future Directions

David Kaplan and Nina Jude

Contents

Abstract

Of critical importance to education policy is the monitoring of trends in educational outcomes over time. With six cycles of TIMSS, three cycles of PIRLS, and seven cycles of PISA, these data enable cross-time comparisons at the country level. A careful analysis of trend data can enable researchers and policymakers to assess cross-country progress and forecasts toward agreed-upon international education goals. The focus of this chapter is on the methodology and presentation of trend results in PISA, TIMSS, and PIRLS. As a motivating example, we focus on how trends in the gender gap for literacy and numeracy outcome are reported by the official organizations in charge of producing these international large-scale assessments. We also examine how relevant stakeholders have used trend analyses to inform national education policy. We conclude with discussion of current issues and future directions in trend analysis, wherein we argue for a predictive

D. Kaplan (✉)
University of Wisconsin–Madison, Madison, WI, USA
e-mail: david.kaplan@wisc.edu

N. Jude
University of Heidelberg, Heidelberg, Germany
e-mail: jude@ibw.uni-heidelberg.de

T. Nilsen et al. (eds.), *International Handbook of Comparative Large-Scale Studies in Education*, Springer International Handbooks of Education,
https://doi.org/10.1007/978-3-030-88178-8_57

model-based view of trend analysis and reporting that more fully leverages the purposes for which international large-scale assessments were intended.

Keywords

Large-scale assessments · Trend analysis · Forecasting

Introduction

Of critical importance to education policy is the monitoring of trends in educational outcomes over time. Policymakers and school practitioners need to follow the development of core indicators to adapt their policies and – ideally – plan ahead. At the global level, the United Nations Sustainable Development Goals focus on achieving equity in literacy and numeracy for men and women (Goal 4.6). Stakeholders of all UN nations have signed on to meet these goals by 2030, hence development of these indicators needs to be closely monitored. To this end, international large-scale assessment (ILSA) programs such as TIMSS, PIRLS, and PISA provide population-level trend data on literacy outcomes of relevance to these goals. The purpose of this chapter is to provide a review of the methodology of trend analysis for TIMSS, PIRLS, and PISA. We will outline how trend analyses are used by researchers and reported to policymakers. We will focus on the methodology and reporting of trends in the gender gap for math, science, and reading. Our focus on the gender gap is motivated by the fact that this is a critical equity outcome and an important focus of the United Nations Sustainable Development Goals.

The organization of this chapter is as follows. In the next section we will provide a brief overview of the background and designs of TIMSS, PIRLS, and PISA as they pertain to trend reporting. This will be followed by a description of how trends in the gender gap are reported in TIMSS, PIRLS, and PISA, respectively. The chapter concludes with discussion of future directions in trend analysis, wherein we argue for a predictive model-based view of trend analysis and reporting that more fully leverages the major purpose for which international large-scale assessments were intended – namely monitoring population trends in literacy and numeracy outcomes.

Brief Background and Design of TIMSS, PIRLS, and PISA

The primary sources for our discussion will be the most recent technical reports and results of TIMSS, PIRLS, and PISA published in Martin, Mullis, and Hooper (2016, 2017), Mullis, Martin, Foy, and Hooper (2016a, b, 2017), and OECD (2016, 2017). For simplicity we discuss the methodology of TIMSS and PIRLS together as their sampling design and item development frameworks are very similar. TIMSS is a quadrennial survey of the mathematics and science skills of fourth and eighth grade students. As of 2015, 57 countries and 7 benchmark countries participated in TIMSS, and, as of 2019, TIMSS is entering its seventh assessment cycle. TIMSS

defines its fourth grade target population as all students enrolled in the grade that represents 4 years of schooling counting from the first year of ISCED Level 1, provided that the mean age of the time of testing is at least 9.5 years. The eighth grade TIMSS target population is counted from the first year of ISCED Level 1 as long as the mean age of testing is at least 13.5 years.

In contrast to TIMSS, PIRLS is a quinquennial survey of the reading skills of fourth graders. As of 2016, 50 countries and 11 benchmark countries participated in PIRLS, and as of 2021, PIRLS will enter its fifth assessment cycle. PIRLS defines its target population in the same manner as TIMSS.

Both TIMSS and PIRLS employ a two-stage random sampling design, with a sample of schools drawn in the first stage followed by one or more intact classes of students selected from each of the sampled schools at the second stage. Intact classes of students are sampled rather than individuals from across the grade level or of a certain age because TIMSS and PIRLS focus on students' curricular and instructional experiences, and these are typically organized at the classroom level.

In contrast to TIMSS and PIRLS, PISA is a triennial survey that began in 2000 and is arguably the most important policy-relevant international educational survey currently operating.[1] With the results of PISA 2018 published in December 2019, PISA has now reached seven cycles of assessment. Unlike TIMSS and PIRLS, PISA is an age-based sample with the target population consisting of in-school 15-year-old students from each participating country and economy (OECD, 2002). These students are approaching the end of compulsory schooling in most participating countries, and school enrollment at this level is close to universal in almost all OECD countries.

The sampling framework for PISA follows a two-stage stratified sample design. Each country/economy provides a list of all "PISA-eligible" schools, and this list constitutes the sampling frame. Schools are then sampled from this frame with sampling probabilities that are proportional to the size of the school, with the size being a function of the estimated number of PISA-eligible students in the school. The second-stage of the design requires sampling students within the sampled schools. A target cluster size of 42 students within schools was desired, though for some countries this target cluster size was negotiable.

The PISA assessments take a literacy perspective, "focusing on the extent to which students can apply the knowledge and skills they have learned and practiced at school when confronted with situations and challenges for which that knowledge may be relevant" (OECD, 2017).

In addition to these so-called "cognitive outcomes," policymakers and researchers alike have become increasingly interested in the nonacademic contextual aspects of schooling. Context questionnaires provide important variables for models predicting cognitive outcomes and these variables have become important outcomes in their own right – often referred to as "non-cognitive outcomes" (see, e.g.,

[1] Due to the coronavirus pandemic, the OECD PISA 2021 will now be moved to 2022 (https://www.oecd.org/pisa/), and they will shift all further PISA cycles as well. (2025 instead of 2024).

Heckman & Kautz, 2012). PISA, TIMSS, and PIRLS also assess these noncognitive outcomes via an internationally agreed-upon context questionnaire (see Kuger et al., 2016).

How Are Trends Reported for TIMSS and PIRLS?

The central focus of TIMSS and PIRLS is the reporting of trends in mathematics, science, and (in the case of PIRLS) reading, and so TIMSS and PIRLS abide as closely as possible to the adage attributed to John Tukey and Albert Beaton that "If you want to measure change, don't change the measure." As such, TIMSS and PIRLS attempt to carry forward as many assessment items as possible from previous cycles while at the same time making room for new items. This balance between maintaining trend items and incorporating new items is handled through the assessment design, and particularly through the distribution of the assessment items by content within the cognitive domain.

In reporting trends, TIMSS and PIRLS first track any large demographic changes in student populations over time. TIMSS and PIRLS (Martin et al., 2016, 2017) report on changes in four important demographic characteristics of the assessment population: (a) number of years of formal schooling, (b) average student age, (c) percent of students in the national target population excluded from the assessment, and (d) overall participation rate after using replacements. For TIMSS fourth grade, possible changes were studied for the 1995, 2003, 2007, 2011, and 2015 cycles; for eighth grade every 4 years from 1995 to 2015; for PIRLS every 5 years from 2001 to 2016. The TIMSS and PIRLS reports (Martin et al., 2016, 2017) indicate good consistency across countries in these measures across the cycles, grades, and assessments. Exceptions included the Russian Federation and Slovenia that exhibited structural changes in the age at which students entered school.

For TIMSS and PIRLS, trend differences in mathematics, science, and reading are reported descriptively and separately for each participating country. In addition, the results are broken down in terms of short-term trends (2011–2015) and long-term trends (1995–2015). However, it should be noted that trends are discussed in terms of the differences between the end points of these time frames, and thus the entire trend line (especially for the long-term trends) are not considered. By presenting the results in this fashion, the potential nonlinearity in changes over time, evident from the trend plots, are not addressed.

With regard to our motivating example of the gender gap, the TIMSS international report describes short-term trends (41 participating countries), and long-term trends (17 participating countries). The short-term trend results show that boys had higher mathematics achievement than girls in 11 countries, compared to 2 countries for girls, and 16 countries had no difference in average mathematics achievement for boys and girls. In 1995, it was found that for 7 countries boys had higher mathematics achievement than girls, and in 2015 it was found that in 9 countries boys had higher mathematics achievement than girls. The international report concluded that

there had been little change in fourth grade mathematics achievement trends by gender. These descriptive statements are accompanied by long-term trend lines broken down by gender. Statistical significance notation is placed on the trend lines to indicate for each year when there was a statistically significant difference between boys and girls. Though not directly stated, it is assumed that the nominal 5% significance level is used.

The manner in which the trend results for TIMSS eighth grade mathematics are reported are the same as for the fourth grade results. For TIMSS eighth grade mathematics, the international report suggests very little short-term change in the mathematics achievement gender gap. Similarly, there has been little long-term change in the gender gap. As with the fourth grade results, long-term trend plots broken down by gender are presented for each participating country.

The trend results for fourth and eighth grade science are reported similarly to the results in mathematics. The international report indicates that the long-term trend shows a reduction in boys' advantage in science achievement. For eighth grade science, the short-term trends show an increasing advantage for girls in science achievement, while the long-term trend shows a reduction in boys' historical advantage in science. As with mathematics, trend plots over time accompany the descriptive results.

Finally, with regard to PIRLS, the results are not displayed in quite the same format as with TIMSS. However, the overall findings are that there are more countries displaying improvements in reading over both the short and long term. The findings also show that girls are continuing to outperform boys in reading.

How Are Trends Reported for PISA?

In contrast to TIMSS and PIRLS, PISA data enables trend reporting over different time spans: While major indicators on learning contexts and outcomes are assessed and reported every 3 years, in-depth trend comparisons focusing on specific subjects are provided every 9 years (Kuger & Klieme, 2016). This is due to the way PISA currently develops its frameworks – focusing on one major domain every 3 years, and also how PISA develops and implements its assessment instruments (OECD, 2017). To enable trend analysis, the PISA assessment design uses trend items (also called link items) for every domain, i.e., test items that have been used in previous waves and have not been published. For every wave of assessment, new test items are developed for the major domain, thus every 3 years new testing material is included into the secure item pool. Due to the equating procedures used in the scaling process, a comparison of outcome scores is possible over time (OECD, 2017).

PISA has been developing its assessment design to improve the statistical foundations for analyzing trend. As examples, up until the 2012 cycle, test scores were generated by equating new and trend items for each cycle separately with a link error reflecting the uncertainty associated with the equating process. In addition, PISA 2015 introduced an integrated assessment design by enhancing the number of test

booklets and test items assessing the minor domain aiming at reducing the measurement error for these domains over time. Moreover, PISA 2015 used an IRT item calibration that included all previous cycles simultaneously to estimate a common scale (OECD, 2017; Mazzeo & von Davier, 2014).

PISA reports differences in achievement and noncognitive indicators between boys and girls in each cycle, usually using tables or box-plots. Means and variances on the PISA scale are compared across countries, as well as the share of boys and girls on the different levels of competence. The PISA international reports also present the change over a 9-year cycle for each cognitive domain by gender, both for the mean scores and in the percentage of low achievers and top performers (OECD, 2016).

With regard to the gender gap reported by PISA, a narrowing gap between boys' and girls' overall performance has been observed. However, when looking at the different levels of competence, the share of low-performing girls declined over the years whereas this has been not the case for the boys. Regarding the share of top performers over time, an opposite trend has emerged with an increase in the percentage of top performing girls (OECD, 2015a). In PISA 2018, an overall decline in the share of top-performing boys in the area of science was noted, leading to a decline in the gender gap (OECD, 2019). In fields like digitalization, the number of school dropouts, and the areas of interest for tertiary education, new challenges for equity can be seen when analyzing attitudes and motivation of boys and girls (OECD, 2015b).

The PISA international reports conclude that gender differences in achievement could be influenced through the learning environment, focusing on students' attitudes and learning approaches (OECD, 2016). This includes the task of addressing gender-related stereotypes which need to be challenged by schools and society alike.

Reporting of Trend Results and Policy Reactions

Policy relevance is a key feature of large-scale assessments, and, in particular, the PISA studies. The PISA design offers international comparisons of educational outcomes and allows tracking developments in single indicators or in the relationship between, for example, school-level factors and outcomes, over time. The prominent reporting by the OECD secretariat (OECD, 2019) is usually cross-country comparisons on average performance, distributions on different competence levels, as well as changes in country ranking over time. Moreover, the reports try to identify major factors associated with educational outcomes across all participating countries. These include, among others, admittance and tracking policies, school autonomy, and assessment policies. How do policymakers and educational researchers use the data? Are educational policies shaped based on change in results over time?

PISA data itself offers several indicators that allow tracking changes in educational policies over time. Teltemann and Jude (2019) analyzed the implementation of assessment systems for accountability purposes between PISA 2000 and PISA 2015 and found country-specific trends in areas like standardization of school-level

assessment, marketization, and teacher accountability. Though it is not possible to attribute these trends directly to a countries' participation in PISA, the authors describe an international trend toward more accountability in education where PISA can be seen as one measure which might have triggered the implementation of national accountability procedures.

In many participating countries, PISA results are debated publicly and have led to a so-called "shock reaction" for some of them. Especially in those countries that performed lower than expected, policymakers saw the need to discuss the quality of their education system. While in some countries, like Germany, the reaction was immanent already after the first round of PISA, others reacted only when changes in their performance over time became visible (Breakspear, 2014).

German policymakers initiated extensive educational reforms both after the publications of the TIMSS results in 1997 and again after the rather poor performance based on country rankings and equity measures in PISA 2000 (Niemann, 2015). Almost 10 years later, with PISA 2009, Germany saw an above-average performance in all three assessment areas, while equity measures were still below average. While there is no causal explanation for this change, Klieme, Jude, Baumert, and Prenzel (2010) list a number of policy measures that had been implemented following the PISA shock, including changes in tracking policies and overall quality assurance in the German school system which could be related to visible changes toward more equity.

Trends in equity indicators has also been analyzed for Great Britain. Jerrim et al. (2018) show that while the overall gap in equality grew smaller over time, this was mainly due to a decrease in the achievement scores of high-performing students rather than improvement on the lower end of the proficiency scale in most of the states. For England itself, only little progress in equality has been made over time and no significant improvement could be seen for the group of low-achieving students despite existing policy interventions.

Inequality in achievement based on PISA data has also been discussed in Australia, even though the country's overall performance revealed a high-performing school system. However, long-term trend analyses in Australia raised concerns because of slowly declining average performance in all competence domains over the last 20 years, with no evidence of a closing gap between students from different socioeconomic backgrounds (Masters, 2018).

For France, PISA scores showed a decline in literacy between 2000 and 2009 and an increase in the share of low-performing students. Dobbins and Martens (2011) analyzed French policy reactions and noted that strong arguments were made in the beginning to discredit the accuracy of measurement for the national context. The findings of a continuous decline of results then led to a stronger policy interest in international comparisons and support for broader educational reforms. A similar phenomenon has been described in Japan, where declining rankings in international comparisons between 2000 and 2003 sparked a public debate about ongoing curricular reform (Takayama, 2008).

Martens and Niemann (2013) compared policy reactions on PISA in 21 OECD countries and concluded that an impact was visible only when evaluation of the

educational system was seen as a relevant part of policy and when empirical results did not match the national self-perception; for example, when other countries are ranking higher in the league tables. In their review of the representation of PIRLS-related research in scientific journals, Lenkeit, Chan, Hopfenbeck, and Baird (2015) noted that less than 10% of articles discuss the impact of PIRLS on educational policy and governance. The authors highlight that while this study was intended to inform and guide policy, still their review could not detect any articles analyzing changes to curriculum or assessment which might explain changes in the outcome.

In their analysis of the influence of ILSAs on national policies, Fischman, Topper, Silova, Goebel, and Holloway (2018) noted a pressure felt by educational authorities to present improving trends in core indicators that might intuitively be met by attempts to emulate policies of high-ranking countries. Nevertheless, explanations of "what works" in which context, especially when it comes to predicting development in a multifaceted environment, are still missing.

To summarize, when it comes to interpreting trend data from international large-scale assessments, policymakers tend to analyze: (i) changes in ranking over time in comparison to other countries, (ii) changes in mean scores across cycles, and – rather less often – (iii) changes in subgroup or indicators of inequality over time. To our knowledge, no publications so far have tried to estimate growth in different areas for purposes of forecasting outcomes in future cycles, nor are we aware of any attempt to develop formal statistical models for predicting change in equity. We discuss this issue in the section on Future Directions and Opportunities.

Current Issues in Trend Reporting

Measuring and reporting trend is usually based on the assumption of a stable measurement design, i.e., that the same indicators are administered at several measurement intervals to enable comparisons over time. In the context of educational assessment, it can be argued that measures must adapt to developments in, for example, educational policies, and thus changes in indicators are required. This is the case, for example, when test items become outdated or there is a need to consider innovations in curriculum or assessment methods. In these cases, indicators used in ILSAs are developed anew or adapted based on the respective assessment frameworks (see, e.g., OECD, 2019). In PISA, for example, requirements for a more efficient test implementation have led to: (i) changes in the booklet designs (L. Rutkowski, 2016), (ii) change from a paper-based to a computer-based assessment mode over several cycles, and (iii) adaptive testing approaches to the cognitive assessment (Yamamoto et al., 2018). Although the constructs being assessed might still be given the same label, e.g., "reading competence" or "interest in science," and so-called link items are implemented to assure comparability across time points, critics point out that these changes could endanger the validity of reporting trend across different cycles.

Especially when comparing means in indicators over time based on a cross-country ranking, caution needs to be taken. Gillis, Polesel, and Wu (2016) provide

several examples which show that trends in performance might not be due to a real change in students' performance, but could be attributed to a systematic change in the measures implemented in the study. These effects could result from differences in translation, differential item functioning of link items in specific language groups, and, of course, national context issues influencing the implementation or sampling in a country (see, e.g., Cosgrove & Cartwright, 2014). The assessment mode also seems to be related to lower scores on average across countries when transitioning to computer-based tests (Jerrim et al., 2018). Moreover, differential effects for specific competence areas in individual countries are being discussed as well as gender differences where girls are seen to be performing lower on average in computer-based tests (for an overview see, for example, Davidsson et al., 2012).

In addition to the cognitive tests, context questionnaire indicators in ILSAs also underwent significant changes over time. Jude and Kuger (2018) summarize the changes and resulting challenges: Although questionnaires deliver extensive information on context of learning to help explain changes in achievement, they are also influenced by the need to balance new content with trend indicators within the given testing time. Accordingly, changes in the assessment frameworks have led to changes in the questionnaires. Teltemann and Jude (2019) give an overview of these changes regarding assessment and accountability practices on the school level addressed in PISA between 2000 and 2015. They show that none of the indicators have been present in all six cycles; still a trend in these indicators can be analyzed when allowing for gaps in specific years of assessment. Recently, a prominent socioeconomic indicator, the PISA indicator of economic, social, and cultural status (ESCS), was revised again for PISA 2018 (Avvasati, 2020). While this indicator has been changed several times in the past, it is still being debated as to how to identify valid measures that allow for comparing countries over time (Rutkowski & Rutkowski, 2013; Pokropek et al., 2017; O'Connell, 2019). It can be summarized that trend analyses with international large-scale data face constant challenges, and these challenges have to be considered when interpreting change and drawing lessons for future policy directives.

Future Directions and Opportunities

The question for this final section of the chapter concerns the future of trend analysis using ILSA data, and the opportunities that accrue from innovations in trend analysis and reporting. At the beginning of this chapter, we argued that monitoring trends in educational outcomes was of critical importance to local, national, and global education policy. However, an inspection of trend reporting for PISA, PIRLS, and TIMSS reveals informative but relatively simple displays of changes in averages or percentages across time for populations and subpopulations of interest. Although these displays are important for communicating trends to stakeholders, we believe that more detail can be gleaned from ILSA trend data by adopting a predictive model-based view of changes in trends over time. We further argue that a predictive model-based view of changes in trend over time can lead to the development of

forecasting models which can supplement discussions of how countries are moving toward (or away from) internationally agreed-upon aims such as the UN Sustainable Development Goals.

We situate our proposed model-based forecasting approach in similar work conducted in economics looking at cross-national trends in economic growth (see Fernández et al., 2001). First, perhaps obviously, we recognize that data must be longitudinal in nature in order to study changes in trends over time. Clearly, ILSA data are longitudinal at the country level and thus, across the participating countries, constitute a panel data structure. Second, we follow the work of Fernández et al. (2001) by advocating for an approach toward forecasting that accounts for uncertainty in the parameters of change over time by implementing a fully Bayesian methodology (Kaplan, 2014). Third, we argue that to be policy-relevant, it is necessary to identify predictors of change over time in educational outcomes of interest while at the same time recognizing the uncertainty in choosing any specific set of predictors as the "true" predictor set. Recognizing this uncertainty in forecasting model choice is also handled in a fully Bayesian framework.

A key feature of our proposed predictive model-based approach is the use of methods of data linking across different data sources (see, e.g., Kaplan & McCarty, 2013; Rässler, 2002). For example, to obtain predictors of change in the mathematics or science gender gap, we can draw on the questionnaires from the surveys themselves, or bring in information from supplementary data sources. In addition to competency outcomes, PISA, TIMSS, and PIRLS include school-level resource, accountability, and leadership indicators that can be aggregated to the country level. (However, it is important to exercise caution when interpreting aggregated variables as their meaning might change.) In addition to PISA, the OECD also provides data on country-level economic indicators such as gross domestic product and government spending on education (see https://data.oecd.org/education.htm). Additional data sources from the OECD can be obtained from their annual "Education-at-a-Glance" volumes (e.g., OECD, 2021). Also, many of the OECD education indicators are also made available to the World Bank through its "EdStats All Indicator Query" system. This system offers more than 4000 internationally comparable indicators covering different aspects of system-level education and data are available from the year 1970 onward (see The World Bank EdStats All Indicator Query, 2019). Finally, UNESCO offers a considerable amount of data in the area of international education, including data already linked to the SDGs since 2012 (see UNESCO, 2015). UNESCO also has in place a global educational monitoring system for which additional data are readily available (see UNESCO, 2015). Because trend analyses are conducted at the country level, these data sources and perhaps other international indictors can be merged with PISA, TIMSS, and PIRLS.

A recent paper by Kaplan and Stancel-Piątak (2019), presented at the 2019 IEA International Research Conference, presented preliminary results for the proposed approach by using TIMSS for model-based forecasting of girls' mathematics achievement. Utilizing a Bayesian growth curve model (Kaplan, 2014) to estimate the trend over time in girls' mathematics achievement, combined with Bayesian model averaging to address model forecasting model uncertainty, Kaplan and

Stancel-Piątak found that girls' math achievement followed a relatively flat linear trend across the 16 participating countries used in the analysis, with some noticeable variation within countries. Of the predictors of growth accounting for model uncertainty, Kaplan and Stancel-Piątak found that country aggregated shortage of calculators (a rough indicator of resources for mathematics instruction) was the dominant predictor of trend in math achievement for girls across countries. Forecasts of growth were obtained for each of the 16 countries, and it was found that the predicted growth was close to the actual growth for many, but not all, countries.

It must be noted that the paper by Kaplan and Stancel-Piątak (2019) was a "proof-of-concept" and that the variables that were used in their model were never designed with the purposes of probabilistic forecasting in mind. Nevertheless, opportunities for sophisticated modeling of trend are apparent. Specifically, with a well-fitting forecasting model one can engage in a variety of policy-relevant forecasting exercises. First, one might wish to conduct an *ex post* forecast in which one is interested in whether the model can reproduce the known historical trend. This is essentially an assessment of model fit. Second, one can use the trend data and forecasting model to conduct *pseudo-ex ante* forecasting wherein one might build a forecasting model on data from, say, PISA 2003 to PISA 2015, and use the data to predict the known results from PISA 2018. Given that the actual PISA 2018 would be available, the pseudo-ex post forecast exercise would allow one to examine forecasted values against actual values and calibrate the model accordingly. Finally, one could conduct a true *ex ante* forecast of, say, PISA 2022 and wait to see if the PISA 2022 results match the prediction. Of course, there will be forecasting error, but this would naturally lead to further calibration of the forecasting model.

The significance of adopting a predictive model-based view of trend analysis is threefold. First, this viewpoint can advance the policy and educational monitoring purposes of ILSAs. A recent paper by Braun and Singer (2019) pointed out the problems associated with common uses of ILSAs. In particular, Braun and Singer (2019, p. 82) noted that the use of ILSAs for evaluating curricular, instructional, and/or educational policies were could be conducted but only with "extreme caution" and that using ILSAs for causal inference was "generally impossible." Braun and Singer (2019) do note, however, that ILSAs are particularly useful for purposes of "transparency," to "spur educational reforms" (e.g., the German "PISA shock"), to "describe and compare student achievement and contextual factors" (with caveats), and, of relevance to this paper, "[t]o track changes over time" (again with some caveats). We agree with many of the issues raised in Braun and Singer (2019) and argue that ILSAs have not been sufficiently leveraged for one of the purposes for which they were originally intended – namely, monitoring population trends. The predictive model-based framework we are proposing can demonstrate the richness of policy information that can be obtained when using Bayesian prediction models to study educational trends at the population level. Indeed, a recent paper by Kaplan and Huang (2021) extended the work of Kaplan and Stancel-Piątak (2019) by developing a workflow that permits Bayesian probabilistic forecasting to be applied to large-scale assessments, with data from the the National Assessment of Educational Progress (NAEP) being used as an example. Results from the Kaplan and

Huang (2021) study also demonstrate the opportunities for this type of trend analysis.

Second, the predictive model-based approach advocated here and demonstrated by Kaplan and Huang (2021) goes beyond the simple presentations of trend. This approach argues for models and methods in order to explicitly forecast changes in literacy and numeracy outcomes over time. In this regard, we borrow from a long tradition of work on demographic forecasting and specifically consider the population gender gap in mathematics and science literacy as population demographic trends worthy of the same attention as aging, mortality, fertility, and migration – trends that are of primary focus in the field of demography.

Finally, we view the significance of our proposed predictive model-based framework toward trend analysis as contributing to the goals of international organizations such as the United Nations, UNESCO, the OECD, the World Bank, and the World Education Forum in monitoring progress toward realizing inclusive and equitable quality education for all.

References

Avvisati, F. (2020) The measure of socio-economic status in PISA: a review and some suggested improvements. *Large-scale Assessments in Education, 8*. https://doi.org/10.1186/s40536-020-00086-x.

Braun, H. I., & Singer, J. D. (2019). Assessment for monitoring of education systems: International comparisons. *The Annals of the American Academy of Political and Social Science*. https://doi.org/10.1177/0002716219843804

Breakspear, S. (2014). *How does PISA shape education policy making? Why how we measure learning determines what counts in education (technical report)*. Centre for Strategic Education, Seminar Series Paper No. 240.

Cosgrove, J., & Cartwright, F. (2014). Changes in achievement on PISA: The case of Ireland and implications for international assessment practice. *Large-scale Assessments in Education, 2*. https://doi.org/10.1186/2196-0739-2-2

Davidsson, E., Sorensen, H., & Allerup, P. (2012). Assessing scientific literacy through computer-based tests – Consequences related to content and gender. *Nordic Studies in Science Education, 8*. https://doi.org/10.5617/nordina.533

Dobbins, M., & Martens, K. (2011). Towards an education approach a la finlandaise? French education policy after PISA. *Journal of Education Policy, 27*, 23–43.

Fernández, C., Ley, E., & Steele, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics, 16*, 563–576.

Fischman, G., Topper, A., Silova, I., Goebel, J., & Holloway, J. L. (2018). Examining the influence of international large-scale assessments on national education policies. *Journal of Education Policy, 30*, 1–10.

Gillis, S., Polesel, J., & Wu, M. (2016). PISA data: Raising concerns with its use in policy settings. *Australian Educational Researcher, 43*, 131–146.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics, 19*, 451–464.

Jerrim, J., Micklewright, J., Heine, J., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the "mode effect" and what has been done about it? *Oxford Review of Education, 44*, 476–493.

Jude, N., & Kuger, S. (2018). *Questionnaire development and design for International Large-Scale Assessments (ILSAs): Current practice, challenges, and recommendations*. Workshop series on methods and policy uses of International Large-Scale Assessments (ILSA).

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. Guilford Press.

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*. Retrieved from http://www.largescaleassessmentsineducation.com/content/1/1/6

Kaplan, D., & Stancel-Piątak, A. (2019). *Optimally predictive cross-country growth models with applications to TIMSS*. IEA Research Conference.

Kaplan, D. & Huang, M. (2021). Bayesian probabilistic forecasting with large-scale educational trend data: A case study using NAEP. *Large-Scale Assessments in Education, 9*. https://doi.org/10.1186/s40536-021-00108-2.

Klieme, E., Jude, N., Baumert, J., & Prenzel, M. (2010). PISA 2000–2009. Bilanz der Veranderungen im Schulsystem. In E. Klieme, C. Artelt, J. Hartig, O. Köller, N. Jude, W. Schneider, M. Prenzel, & P. Stanat (Eds.), *PISA 2009. Bilanz nach einem Jahrzehnt*. Waxmann.

Kuger, S., & Klieme, E. (2016). Dimensions in context assessment. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 3–38). Springer.

Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (2016). *Assessing contexts of learning: An international perspective*. Springer.

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.

Martens, K., & Niemann, D. (2013). When do numbers count? The differential impact of the PISA rating and ranking on education policy in Germany and the US. *German Politics, 22*, 314–332.

Martin, M. O., Mullis, I., & Hooper, M. (2016). *Methods and procedures in TMISS 2015*. TIMSS and PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I., & Hooper, M. (2017). *Methods and procedures in PIRLS 2016*. TIMSS and PIRLS International Study Center, Boston College.

Masters, G. (2018). Using PISA to monitor trends and evaluate reforms in Australia. In L. Volante (Ed.), *The PISA effect on global educational governance* (pp. 175–188). Routledge.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 229–258). CRC Press.

Mullis, I., Martin, M. O., Foy, P., & Hooper, M. (2016a). *TIMSS 2015: International results in mathematics: Eighth grade mathematics*. TIMSS and PIRLS International Study Center Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I., Martin, M. O., Foy, P., & Hooper, M. (2016b). *TIMSS 2015: International results in mathematics: Fourth grade mathematics*. TIMSS and PIRLS International Study Center Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016: International results in reading*. TIMSS and PIRLS International Study Center Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Niemann, D. (2015). Pisa in Deutschland: Effekte auf Politikgestaltung und organisation. *Die Deutsche Schule, 107*, 141–157.

O'Connell, M. (2019). Is the impact of SES on educational performance overestimated? Evidence from the PISA survey. *Intelligence, 75*, 41–47.

OECD. (2002). *PISA 2000 technical report (technical report)*. OECD Publishing.

OECD. (2015a). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD Publishing.

OECD. (2015b). *What lies behind gender inequality in education? PISA in focus (49)*. OECD Publishing.

OECD. (2016). *PISA 2015 results (volume 1): Excellence and equity in education*. OECD Publishing.

OECD. (2017). *PISA 2015 technical report (technical report)*. OECD Publishing.

OECD. (2019). *Pisa 2018 assessment and analytical framework*. OECD Publishing.
OECD. (2021). *Education at a glance 2021*. OECD Publishing.
Pokropek, A., Borgonovi, F., & McCormick, C. (2017). On the cross-country comparability of indicators of socioeconomic resources in PISA. *Applied Measurement in Education, 30*, 243–258.
Rässler, S. (2002). *Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches*. Springer.
Rutkowski, L. (2016). *A look at the most pressing design issues in international large-scale assessments*. Workshop series on methods and policy uses of International Large-Scale Assessments. National Academy of Education.
Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*, 259–278.
Takayama, K. (2008). The politics of international league tables: PISA in Japan's achievement crisis debate. *Comparative Education, 44*, 387–407.
Teltemann, J., & Jude, N. (2019). Assessments and accountability in secondary education: International trends. *Research in Comparative and International Education, 14*, 249–271.
The World Bank EdStats All Indicator Query. (2019). Retrieved accessed: 15 June 2019, from https://datacatalog.worldbank.org/dataset/education-statistics
UNESCO. (2015). Education for all global monitoring report: Achievements and challenges. Paris, UNESCO Publishing.
Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage adaptive testing design in international large-scale assessments. *Educational Measurement, 37*, 16–27.