PISA 2021 Context Questionnaire Framework (Field Trial Version)

Doc.: CY8_201901_QuestionnaireFramework_FieldTrial.pdf

January 2019

Produced by ETS, Core B2

Contents

The PISA 2021 Context Questionnaires – Balancing Continuity with Efficiency and Innovation

Jonas Bertling & Jan Alegre, Educational Testing Service

Glossary

1. Throughout the PISA 2021 Context Questionnaire Framework, several specific terms are used. To ensure consistent understanding of these terms, Table 1 below lists key terms used throughout the framework, along with brief definitions of the terms.

Table 1. Glossary of Key Terms (thematically grouped)

Term	Definition
Construct	A theoretically defined conceptualization (i.e. something constructed) of an aspect of human behaviour or an empirical phenomenon; a construct has empirical indicators, but may not be completely observable due to deficits of existing measures. Two broad content categories of constructs are distinguished in the framework, those that are specific to a PISA 2021 content domain (i.e. mathematics, reading, science, creative thinking) and those that are general (i.e. not specific to a PISA 2021 content domain) and may or may not be related to achievement.
Module	Grouping of two or more related constructs that mark a key topic or theme measured with the PISA 2021 questionnaires.
Question	The parts of a questionnaire designed to elicit information from a respondent. When presented in the digital platform, the question appears on a single screen. In PISA, the question can take the form of a stand-alone discrete question or a matrix question.
Matrix Question	A question that consists of a question stem and several items with the same response options.
Item	The unit(s) of a question that a respondent answers. In case of a stand-alone discrete question, the item is the same as the question. In case of a matrix question, one question includes several items.
Response Options	A typically verbally labelled set of answer choices provided to respondents for close-ended multiple-choice questions.
Scaled Index	An index or measure based on the scaling (using item response theory) of multiple items that all are indicators of an underlying construct.
Questionnaire Matrix Sampling	A questionnaire design where each respondent receives only a subset of items in the entire questionnaire. In a *within-construct* matrix sampling design, a respondent answers items for all constructs but only receives a subset of items for each construct. In contrast, in a *construct-level* matrix sampling design entire constructs are rotated across questionnaire booklets.

Table 2. Glossary of Acronyms (ordered alphabetically)

Acronym	Term
CIPO	Context-Input-Process-Output Model
CT	Creative Thinking
ESCS	Economic, Social, and Cultural Status
FT	Field Trial
ICT	Information and Communication Technology
IRT	Item Response Theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
LSA	Large-scale Assessment
MEG	Mathematics Expert Group
MS	Main Survey
NAEP	National Assessment of Educational Progress
OECD	Organisation for Economic Co-operation and Development
OTL	Opportunity to Learn
PGB	PISA Governing Board
PISA	Programme for International Student Assessment
PISA-D	PISA for Development
QEG	Questionnaire Expert Group
SCQ	School Questionnaire
SES	Socioeconomic Status
SSES	Study of Social and Emotional Skills
STQ	Student Questionnaire
TAG	Technical Advisory Group
TALIS	Teaching and Learning International Survey
WBQ	Well-being Questionnaire

1. Introduction

1.1. Aims of the PISA Questionnaires

2. National and international Large-scale Assessments (LSAs) play an important role in evaluating education systems in terms of their capacity to develop human potential, advance progress and the quality of life of individuals across the globe, and prepare future workforces for 21st century demands. Since its inception in the late 1990s, the *Programme for International Student Assessment* (PISA) has been known for its important contribution to education policy discussions within the Organisation for Economic Co-operation and Development (OECD) and partner countries and economies.

3. The main features of PISA are as follows:

- PISA is a *system-level assessment*, representing a commitment by governments to monitor the outcomes of education systems.
- PISA is *policy-oriented*, linking data on students' learning outcomes with data on key factors that shape learning in and out of school.
- PISA is *carried out regularly*, enabling countries to monitor their progress in meeting key learning objectives.
- PISA *assesses both subject matter knowledge*, on the one hand, and *the capacity of individuals to apply that knowledge creatively*, including in unfamiliar contexts, on the other.
- PISA focuses on *knowledge and skills towards the end of compulsory schooling*. In most countries, the end of compulsory education is around the age of 15, where students are supposed to have mastered the basic skills and knowledge to continue on to higher education or in the workforce.
- PISA is designed to *provide comparable data across a wide range of countries.* Considerable efforts are devoted to achieving cultural and linguistic breadth and balance in assessment materials.
- PISA is a *collaborative effort* involving multiple parties.

4. PISA continues to yield indicators on effectiveness, efficiency, and equity of educational systems, setting benchmarks for international comparison and monitoring trends over time. PISA also builds a sustainable database that allows researchers worldwide to study basic as well as policy-oriented questions on education, including those related to society and economy. The OECD and the PISA Governing Board (PGB) continue to look for ways to increase the scientific quality and policy relevance of the PISA context questionnaires to meet these needs.

5. Since the first cycle of PISA in 2000, the student and school context questionnaires have performed two interrelated purposes in service of the broader goal of evaluating educational systems:

- First, the questionnaires provide a context for interpreting the PISA results both within and between education systems.
- Second, the questionnaires aim to provide reliable and valid measurement of *additional educational constructs*, which can inform policy and research in their own right.

6. Over the seven cycles of PISA to date, education policy discussions have shifted from a heavy focus on the first objective to increased focus on the second aim as well. This development corresponds to a shift in policymakers' views of the core goals for education systems in the 21st century, away from primarily teaching clearly defined subject knowledge and skills, to fostering broader skills, such as

creativity, communication, collaboration, or learning to learn, that help individuals face the demands of a technology-rich and truly global society (United Nations, 2015). There is now a growing recognition that other factors and competencies aside from subject-specific knowledge play a vital role in fostering students' success in school and beyond. In order to understand and guide policy decisions regarding student development, the PISA 2021 context questionnaires will strengthen the measurement of the contexts that promote learning in these areas, as well as an array of general constructs of policy relevance.

1.2. Outline of the PISA 2021 Context Questionnaires Framework

7. The PISA 2021 context questionnaires framework explains the goals and rationale for selecting specific questionnaire content for the eighth cycle of PISA, guiding both instrument development and subsequent reporting for students and school administrators. It draws from current literature on educational effectiveness to describe relevant constructs and modules. Similar to prior frameworks, the present framework touches upon how measured constructs theoretically relate to one another and to student achievement. Additionally, the framework outlines a set of survey design principles and methodologies to consider during the questionnaire development and Field Trial (FT) stages with the aim of improving measurement, efficiency, and continuity of PISA in the long term. To achieve these goals, the framework is structured as follows:

- *Section 2.* describes a set of general considerations that led to the development of this framework and will guide instrument development moving forward. These considerations include priorities for re-administration of questions from previous PISA cycles, changes to the mathematics framework since PISA 2012 that need to be considered when prioritizing questionnaire constructs, country-specific needs across the range of participating countries, directions taken with the PISA 2021 innovative domain of creative thinking, and plans for optional questionnaires.
- *Section 3.* presents the PISA 2021 two-dimensional framework taxonomy. The first dimension classifies proposed constructs into the two overarching categories distinguished by the PGB (domain-specific constructs and general constructs, with the latter including Economic, Social, and Cultural Status [ESCS]). The second dimension classifies proposed constructs into five categories based on key areas of educational policy setting at different levels of aggregation (Student Background; Student Beliefs, Attitudes, Feelings, and Behaviours; Teaching Practices and Learning Opportunities; School Practices, Policies, and Infrastructure; and Governance, System-Level Policies and Practices). Linkages between the 2021 approach to the overarching cross-cycle structure developed across the PISA 2000 – 2018 questionnaire frameworks are highlighted, with a focus specifically on the three past PISA cycles, i.e. 2012, 2015, and 2018 (Klieme et al., 2013; Klieme & Kuger, 2014; OECD, 2013).
- *Section 4.* gives a detailed overview of the PISA 2021 FT questionnaire modules and constructs, and explores in depth the breadth of policy issues within the previously laid out content areas. Based on analysis of FT data and discussion of priorities among experts and policy makers (including the PGB), a final set of constructs and items will be selected for inclusion in the PISA 2021 Main Survey (MS) after FT data has been collected.
- *Section 5.* summarizes recommended survey design principles to guide the PISA 2021 questionnaire development process, subsequent FT administration, and post-FT analyses and item selection for the MS.

2. Balancing Re-administration of Questions from Previous Cycles with New Development

8. For PISA 2021, the PGB recommended re-balancing questionnaire content in the direction of a larger focus on general constructs and a slightly reduced focus on domain-specific constructs. Specifically, the PGB suggested that 40% of the content be devoted to domain-specific constructs. The remaining 60% focused on general constructs would be split between 20% devoted to measuring ESCS and 40% focused on other general constructs, including additional outcomes (PISA Governing Board, 2017). By contrast, the balance of questionnaire content across domain-specific constructs, ESCS, and general constructs in 2018 was 50%, 17%, and 33%, respectively.

9. It is suggested that percentages be allocated based on estimated questionnaire administration time. For the 2021 MS, the upper limit of testing time for the student questionnaire (STQ) is 35 minutes.[1] That is, approximately 7 minutes of the STQ will be devoted to ESCS and 14 minutes each will be devoted to domain-specific and general constructs. Within the boundaries of these overall strategic priorities, two key areas of consideration guide the development of this draft of the PISA 2021 questionnaire framework: (1.) re-administration of questions from previous PISA cycles and (2.) new development.

2.1. Guidelines for Re-administration of Questions from Previous Years

10. A key force driving the PISA design in general is the cyclical change of focus in the cognitive assessment. Mathematics was the major domain of cognitive assessment in PISA 2003 and 2012, and will be the major domain again in 2021. Reading was the major domain of assessment in PISA 2000, 2009, and 2018. Science was the focus of PISA 2006 and 2015. The major domain serves as the primary focus of domain-specific content in the associated PISA context questionnaire (e.g. various mathematics-related constructs marked the focus of the 2003 and 2012 questionnaires).

11. In order to describe educational constructs of interest over time at the country level, it is desirable to maintain a stable set of questionnaire measures that can be used as major reporting variables across PISA cycles. Given the cyclical nature of PISA, measurement stability can be considered at two levels:

- First, there is the issue of stability of measures across cycles of three years (i.e. administration of items for constructs that may appear in every cycle, e.g. ESCS).
- Second, stability is desirable in measuring domain-specific constructs across cycles of nine years (i.e. mathematics-specific constructs assessed in the 2012 and/or 2003 cycles).

A priority of PISA 2021 is to retain a reasonable number of questions that have been administered in previous PISA questionnaires.

[1] Please note, more questionnaire content will be administered for the international FT. Percentages across the three broad content categories may not exactly match the envisioned percentages for the MS because some areas require more new development than other areas, which can already be mostly represented based on questions from previous PISA cycles.

12. Table 3 summarizes guidelines for considering the retention or deletion of previously administered PISA items in the PISA 2021 FT.

Table 3. Guidelines for Retention or Deletion of PISA Questions from Previous Cycles

Guidelines for Retention or Deletion of PISA Questions from Previous Cycles
1. Retain questions that best explain variations in academic achievement within and across countries; 2. Retain questions that are of highest policy relevance and/or necessary to establish or extend trend lines, which can inform policy and research; 3. Where possible and sensible, carry forward constructs intact, or with only minor changes that improve measurement precision; 4. Delete or revise questions that are outdated (e.g. questions that make reference to resources or technologies that are no longer in use); 5. Delete or revise questions that do not meet PISA 2021 psychometric criteria established by the Technical Advisory Group (TAG; e.g. internal consistencies of <.70 or issues with scalability in a substantial number of countries); and 6. Delete or shorten questions that provide information that is redundant with other questions or items within a matrix question.

2.2. Guidelines for New Development

13. The goal of continuity must be carefully balanced with innovations that take into account new developments and emerging priorities in educational systems, as well as ongoing assessment designs in PISA. These include: introduction of new technologies (e.g. computer-based assessment [CBA]); introduction of new innovative domains of assessment (e.g. collaborative problem solving in 2015, global competency in 2018, or creative thinking in 2021); continuous updates of the assessment and analytical frameworks for each content area across cycles (e.g. expansion of the PISA 2021 mathematics framework); and the emergence of new policy priorities (e.g. measuring student health and well-being as well as other social and emotional characteristics).

14. For PISA 2021, the scope of the mathematics framework will be expanded to evaluate students' mathematical reasoning grounded in six core concepts or "big mathematical ideas" that undergird the specific content, skills, and algorithms of school mathematics (PISA Governing Board, 2017):

- Quantity, number systems and their algebraic properties;
- Mathematics as a system based on abstraction and symbolic representation;
- Mathematical structure and its irregularities;
- Functional relationships between quantities;
- Mathematical modelling as a lens onto the real world (e.g. those arising in the physical, biological, social, economic, and behavioural sciences); and
- Variation as the heart of statistics.

15. Students will also be assessed in their familiarity with, or prior classroom exposure to, four emerging areas of mathematics content in which reasoning skills need to be applied: computer simulations, growth phenomena, conditional decision making, and geometric approximation. The questionnaire framework accordingly recommends updates aimed at better understanding students' opportunities to learn these concepts, as well as the extent to which 21st century skills are emphasized in mathematics instruction.

16. Additionally, creative thinking will be assessed as the innovative domain in PISA 2021, and a defined section of the PISA 2021 context questionnaires will be devoted to constructs that contribute to the understanding of students' performance in this innovative domain. Please note that information regarding the specific creative thinking related contextual constructs is provided in the separate Creative Thinking framework for PISA 2021.

17. A number of new educational systems will participate in PISA beginning in 2021, many of which belong to lower- and middle-income countries. In order to maximize the value of PISA to these participants, the context questionnaires may consider the integration of constructs related to student background and additional constructs that have previously been described in the PISA for Development (PISA-D) framework (OECD, 2018).

18. New development should make use of informed practices in survey methodology (e.g. principles regarding item types, response options, balancing of scales, length of matrix questions) and technological capabilities (e.g. routing, matrix sampling) to the extent that they will enhance measurement. Section 5 of this framework elaborates in detail on suggested survey design principles for PISA 2021.

19. While this framework focuses on the conceptual underpinnings of the PISA questionnaires for students and schools, additional frameworks that are not part of this document provide in-depth theoretical foundation for additional questionnaires included in PISA 2021 as part of international options (i.e. frameworks for Financial Literacy, Information and Communication Technology [ICT] Literacy, Student Well-being, Teacher Well-being).

20. Table 4 summarizes guidelines for considering the addition of new items for existing constructs as well as entirely new constructs in PISA 2021.

Table 4. Guidelines for New Development

Guidelines for Addition of New Items for Existing Constructs as well as Entirely New Constructs
1. Develop questions for new constructs that are central to the educational research literature and the PGB priorities; 2. Develop new questions that are relevant to changes to the PISA mathematics framework (e.g. addition of mathematical reasoning); 3. Develop new questions for constructs that are related to the innovative domain assessed in PISA 2021 (i.e. creative thinking); 4. Develop new questions to replace previously used questions that do not comply with PISA 2021 psychometric criteria, substantially violate PISA 2021 survey design principles, and/or require updates to accurately describe students' living and learning realities; and 5. Develop new questions to replace previously used questions that do not offer sufficient flexibility to meet country- or region-specific needs of all participating education systems.

3. PISA 2021 Context Questionnaire Framework Taxonomy

21. Beginning with the questionnaire framework used for the PISA 2009 assessment, questionnaire content was explicitly linked to different levels of the education system: the student level, level of instruction in the classroom, school level, and system level (Jude, 2016). The questionnaire framework used for PISA 2012, and subsequently refined for PISA 2015 and 2018, further underscored the importance of collecting information on learning contexts for comparative system monitoring. These frameworks outlined an overarching two-dimensional structure of high-level questionnaire content areas to be measured and kept comparable across assessment cycles (OECD, 2013).

22. The theoretical foundation of the 2012 overarching framework is based on Purves' (1987) *Context-Input-Process-Outcome* (CIPO) model. In the CIPO model, contextual variables for understanding education systems are conceptualized as a series of inputs (i.e. student background), processes (i.e. teaching and learning, school policies, governance), and outcomes (i.e. performance and non-cognitive outcomes) shaped at the student, classroom, school, and country levels. Starting with PISA 2015 and 2018, an additional dimension further classified questions more explicitly into domain-specific and domain-general modules. Domain-specific modules represent the set of constructs with strong expected relationships to student experiences, outcomes, and teaching and learning factors tied to a specific content area (e.g. reading, mathematics, or science). Domain-general modules represent the set of constructs that are important for understanding differences in achievement that are not tied to a specific subject-area. Figure 1 illustrates the high-level structures of the context questionnaire frameworks from 2012, 2015, and 2018.

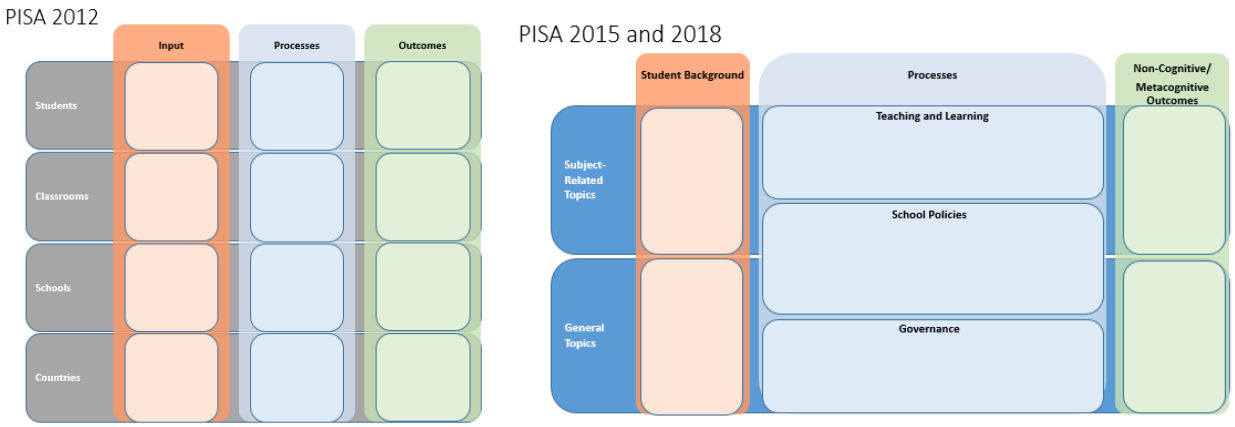
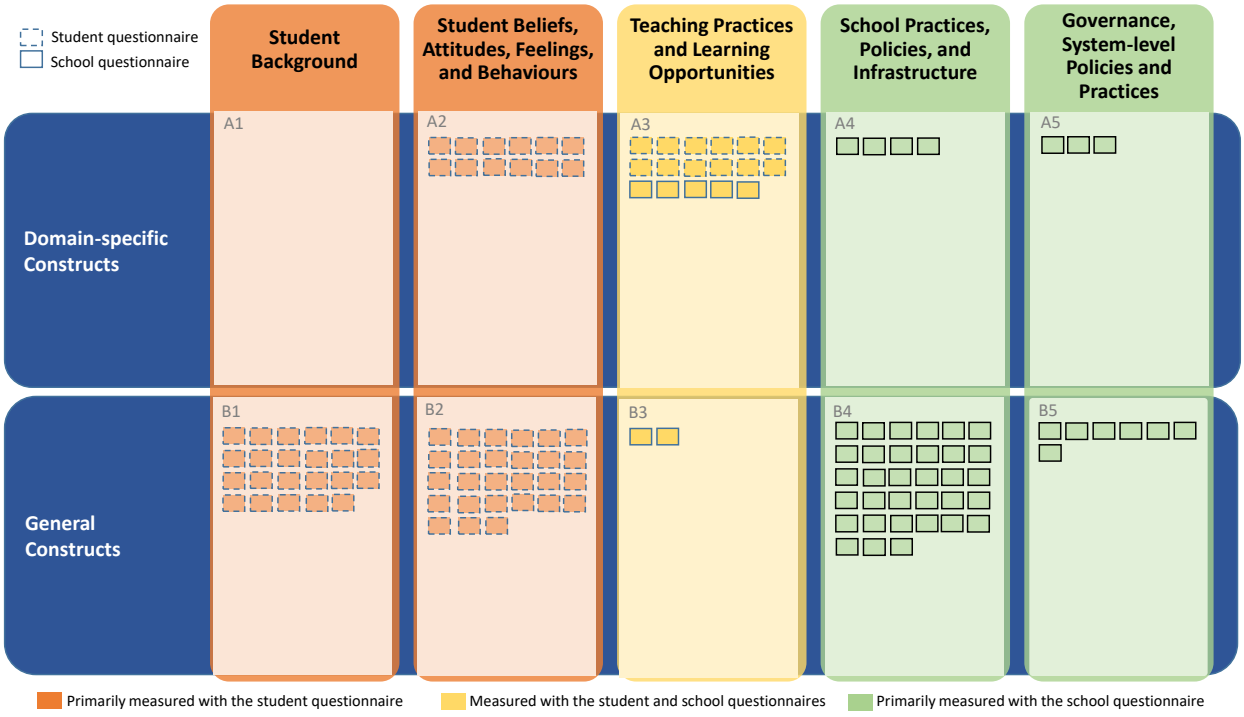


Figure 1. Framework Structures of PISA 2012, 2015, and 2018

23. In keeping with the long-term goal of balancing continuity with innovation, the PISA 2021 context questionnaires framework retains key framework elements from previous cycles as a foundation, and introduces refinements that will facilitate the strategic development of new constructs and move toward improved measurement. This updated framework structure is illustrated in Figure 2 below. Please note, while performance and contextual variables have been classified as "outcomes" in previous PISA frameworks per the CIPO model, both variables also constitute possible inputs (OECD, 2013). For instance, a student's prior achievement and his/her curiosity, perseverance, achievement motivation, or confidence will likely impact the student's future achievement, as well as his/her future development of social and emotional characteristics. Due to the cross-sectional nature of PISA, variables collected through

the questionnaires cannot be clearly assigned a single "role". While the CIPO model remains useful to describe an actionable policy perspective and serve as a helpful theoretical perspective for researchers on the variables measured with the PISA questionnaires, it seems less useful as a guide to classify and prioritize variables for instrument development. Due to the ambiguity in classifying variables, constructs will *not* be classified as inputs, processes, or outcomes in the PISA 2021 framework taxonomy. Instead, we will allude to the possible roles each variable might play in the detailed descriptions of each module. Further description of the framework dimensions and the modules is provided in subsequent sections of this framework.

24. Across the two overarching (vertical) framework dimensions and of the five (horizontal) policy focus areas as shown in Figure 2, a total of 20 modules are specified (see Section 4. of this document), which each consist of groupings of two or more related constructs that mark a key topic or theme relevant for education research, practice, and policy making. In addition to these 20 modules, a separate module focused on contextual questions specific to Creative Thinking (CT) will be included in the PISA 2021 FT. Constructs in this module are described in the CT framework. The small boxes in the taxonomy below indicate the relative distribution of constructs proposed for the PISA 2021 FT across all 20 modules described in this framework.



Note. Domain-specific FT constructs related to Creative Thinking are not reflected in this chart.

Figure 2. PISA 2021 Two-Dimensional Draft Context Questionnaire Framework Taxonomy

3.1. Classification based on Relationships to PISA Content Domains

25. As outlined above, the PISA 2021 student and school questionnaires will serve two interrelated purposes (i.e. provide contextual information and provide additional measures) in service of the broader goal of evaluating the effectiveness of all educational systems participating in the 2021 International FT and MS.

The two categories along the vertical dimension of the taxonomy in Figure 2 represent the primary types of content in the student and school questionnaires:

A. *Domain-specific* Constructs; and
B. *General* Constructs (including ESCS).

26. Both categories of constructs may represent questions that are included in PISA primarily to report their relationships with academic achievement and provide a context for interpreting the PISA results within and between education systems, as well as constructs that are included in PISA primarily to report additional variables that describe educational systems beyond academic achievement to inform policy and research in their own right.

3.1.1. Domain-specific constructs

27. *Domain-specific constructs* include constructs that demonstrate a relationship to students' academic achievement in the major domain of the current cycle (i.e. mathematics for PISA 2021) or hold power to explain broader outcomes in the major domain, such as students' educational career and post-secondary decisions (e.g. course enrolment, career decisions). Examples of indicators may include mathematics-related school curricula, the value attributed to mathematics within the school community, or students' interest and motivation to learn mathematics topics. Constructs that are included primarily to better understand differences in achievement in the PISA 2021 mathematics achievement scores should be evaluated empirically after the FT according to their relationship with mathematics achievement to determine inclusion in the PISA 2021 MS. In addition to constructs related to the major domain (i.e. mathematics), a smaller number of contextual variables specific to all three domains (including the two minor domains of this assessment cycle, Reading and Science) will also be included in the PISA 2021 FT questionnaires to provide relevant contextual information for student achievement (e.g. cross-subject analysis with student fixed effects models). Lastly, the category of domain-specific constructs includes constructs included to contextualize achievement results in the innovative domain for PISA 2021 (i.e. creative thinking).

3.1.2. General constructs

28. *General constructs* include constructs that demonstrate relationships to students' academic achievement across multiple domains, such as students' feelings towards school (e.g. student-teacher relationships, bullying experiences), school infrastructure (e.g. availability of digital technology for learning), or constructs that complement traditional indicators of educational effectiveness (e.g. subjective well-being, social and emotional characteristics). General constructs also include ESCS to assess students' socioeconomic status (SES) and the equity of educational opportunities within and across educational systems.

3.2. Classification based on Educational Policy Areas

29. The horizontal dimension of the taxonomy distinguishes five categories of educational policy focus that correspond to different aggregate levels for the collected survey responses, from individual-level variables to highly aggregated system-level indicators:

1. Student Background;
2. Student Beliefs, Attitudes, Feelings, and Behaviours;

3. Teaching Practices and Learning Opportunities;
4. School Practices, Policies, and Infrastructure; and
5. Governance, System-Level Policies and Practices.

3.2.1. Student Background

30. The first educational policy area of interest relates to *Student Background.* In order to understand students' education pathways and to study equity within and across educational systems, basic demographic variables (e.g. gender, age, or grade), constructs related to ESCS, migration and language background, as well as information about students' early years must be taken into account. The distribution of educational opportunities and outcomes correlated with these background constructs may provide data about whether countries succeed in providing equity in educational opportunities.

3.2.2. Student Beliefs, Attitudes, Feelings, and Behaviours

31. The second educational policy area of interest focuses on *Student Beliefs, Attitudes, Feelings, and Behaviours.* In addition to measuring 15-year-olds' academic achievement in reading, mathematics, science, and creative thinking, measures of students' subjective attitudes and feelings, as well as their behavioural choices may provide important indicators for an education system's success in fostering productive members of society.

32. Beliefs may include constructs such as beliefs about learning or student's mindsets. Attitudes may include constructs such as attitudes towards mathematics, or attitudinal aspects of social and emotional characteristics. Feelings may concern feelings about their school or about specific subject-areas, and emotional aspects of social and emotional characteristics. Behaviours may include participation in activities outside of school or behavioural aspects of social and emotional characteristics. Constructs such as respecting and understanding others, being motivated to learn and collaborate, or being able to regulate one's own behaviour may play a role as prerequisites of acquiring subject-area knowledge and skills. In addition, such characteristics may also be judged as goals of education in their own right (Almlund, Duckworth, Heckman, & Kautz, 2011; Bertling, Marksteiner, & Kyllonen, 2016; Heckman, Stixrud, & Urzua, 2006; Rychen & Salgnik, 2003).

33. Each of the past seven PISA cycles included a significant number of questions tapping into students' beliefs, attitudes, feelings, and behaviours with regard to the major domain. In addition, recent PISA cycles have increased their focus on general constructs (e.g. "Noncognitive outcomes" modules in PISA 2015 and PISA 2018). PISA 2021 will carry these developments forward and include several modules addressing a range of constructs such as students' effort on the PISA test and questionnaires (Module 5), students' general school-related attitudes and feelings associated with school climate (Module 6), attitudes towards each of the PISA content domains (Module 7), and students' general social and emotional characteristics (Module 8). A broad range of student behaviours will further be assessed via a module focused on out-of-school experiences (Module 11). In addition, students' subjective views on their socioeconomic standing, as well as their well-being and future aspirations, are captured in modules 2, 9, and 10, respectively.

3.2.3. Teaching Practices and Learning Opportunities

34. The third educational policy area of interest pertains to *Teaching Practices and Learning Opportunities.* Classroom-based instruction is the immediate and core setting of formal, systematic education. Therefore, policy makers need information on the organization of classrooms and the teaching and learning experiences that occur within them. The knowledge base of educational effectiveness research (e.g. Scheerens & Bosker, 1997; Creemers & Kyriakides, 2008) allows for the identification of core

variables with an expected bearing on mathematics and student achievement in general, for example, teachers' qualifications, teaching practices and classroom climate, learning time, and learning opportunities provided in and outside of school. As such, this policy area closely links to the idea of *opportunity to learn* (OTL), which was first introduced by Carroll (1963) to indicate whether students have had sufficient time and received adequate instruction to learn (Abedi, Courtney, Leon, Kao, & Azzam, 2006). Though the meaning of OTL has since broadened, it has been an important concept in international student assessments (e.g. Schmidt et al., 2001) and shown to be strongly related to student performance in cross-country comparisons (Schmidt & Maier, 2009).

35. Researchers have suggested defining OTL not only based on subject-specific teacher instruction (Callahan, 2005; McDonnell, 1995) and have stressed the importance of the quality of instruction in addition to mere quantity (Duncan & Murnane, 2011; Minor, Desimone, Spencer, & Phillips, 2015), though the measurement of instructional quality poses challenges as self-reports are likely to be affected by social desirability (Little, Goe, & Bell, 2009; van de Vijver & He, 2014). Researchers have also pointed out the importance of informal learning opportunities and experiences in the home (Lareau & Weininger, 2003) and highlighted the need to evaluate OTL in country-specific contexts (Cogan & Schmidt, 2015). Accounting for these broader directions, OTL could be defined as all contextual factors that capture the cumulative learning opportunities a student has been exposed to at the time of the assessment (Bertling et al., 2016). These contextual factors may comprise both learning opportunities at school and informal and formal learning opportunities outside of school. In this framework, several aspects of OTL are captured across different modules, including modules capturing opportunities provided through the ways in which student learning is organized (Module 15), opportunities defined based on the mathematics content students are exposed to (Module 16), and opportunities created based on the behaviours teachers show in the classroom (Module 17).

3.2.4. School Practices, Policies, and Infrastructure

36. The fourth educational policy area of interest examines *School Practices, Policies, and Infrastructure.* As policymakers have limited direct impact on teaching and learning processes, information on school-level factors (e.g. practices, policies, and infrastructure) that help to improve schools, and thus indirectly improve student learning, are a priority. In addition to individual student demographics and structural factors, such as school location, school type, and school size, the social, ethnic, and academic composition of the school influences students' learning processes and outcomes. Therefore, PISA uses aggregated student data to characterize demographic and other contextual factors at the level of the school community.

37. Similar to the *Teaching Practices and Learning Opportunities* modules and constructs, school effectiveness research has reported that "essential supports" are associated with school effectiveness (Bryk, Sebring, Allensworth, Luppescu, & Easton, 2010). These essential supports include leadership and school management; well-organized curriculum, instructional, and enrolment policies; tangible resources; positive school climate; and parent or guardian involvement. Many of these factors have been previously addressed in the PISA questionnaires as domain-general processes on the school level. Also covered is school-level support for teaching the major domain, such as the provision of learning resources and space, information and communication technology (ICT), and a school curriculum for mathematics education.

3.2.5. Governance, System-level Policies and Practices

38. Finally, the fifth educational policy area of interest focuses on *Governance, System Level Policies and Practices.* To meet policy requests directly, PISA also needs to address issues related to governance at the system level (Hanushek & Wöβmann, 2011). For instance, assessment and evaluation are basic

processes that policy makers and/or school administrators use to control school quality, and to monitor and foster school improvement. These issues have been previously examined in the PISA questionnaires as domain-general context variables on the system level; domain-specific system-level context variables are also recommended for consideration in PISA 2021. Some of this information can be collected through the PISA school questionnaire (SCQ) and system level measurements, while some information can potentially be acquired from other sources. During instrument development, the possibility of using the OECD system-level data collection as well as administrative records may be explored to collect complementary information to the PISA questionnaires and/or reduce respondent burden by using system-level data instead of individual-level questionnaire data.

4. Detailed Overview of PISA 2021 FT Modules

4.1. Basic Demographics

39. PISA questionnaires have routinely included questions on students' gender and age, as well as their grade. Please note, while it is recommended to include these questions again in the STQ during the FT, data for these variables may also be obtained from the respective school a student attends. If feasible in all countries and unless needed for quality control purposes, these variables might be candidates for collection via school records only in the MS in order to create room for additional items in the questionnaire.

40. In addition to these questions, PISA 2021 will update questions on home composition to better reflect modern living realities in traditional as well as non-traditional homes and establish a foundation for potential routing throughout the questionnaire based on, for instance, the number of parents or guardians a student has. This will help support the goal of assessing ESCS in a less intrusive and more valid way. PISA 2021 FT instruments will aim to collect data on students' broader living circumstances, specifically whether students live in a traditional nuclear family with two parents or guardians, a single parent or guardian home, or whether they occasionally or regularly reside in homes between multiple parents or guardians. Figure 3 illustrates how all proposed constructs in this module map on the taxonomy.

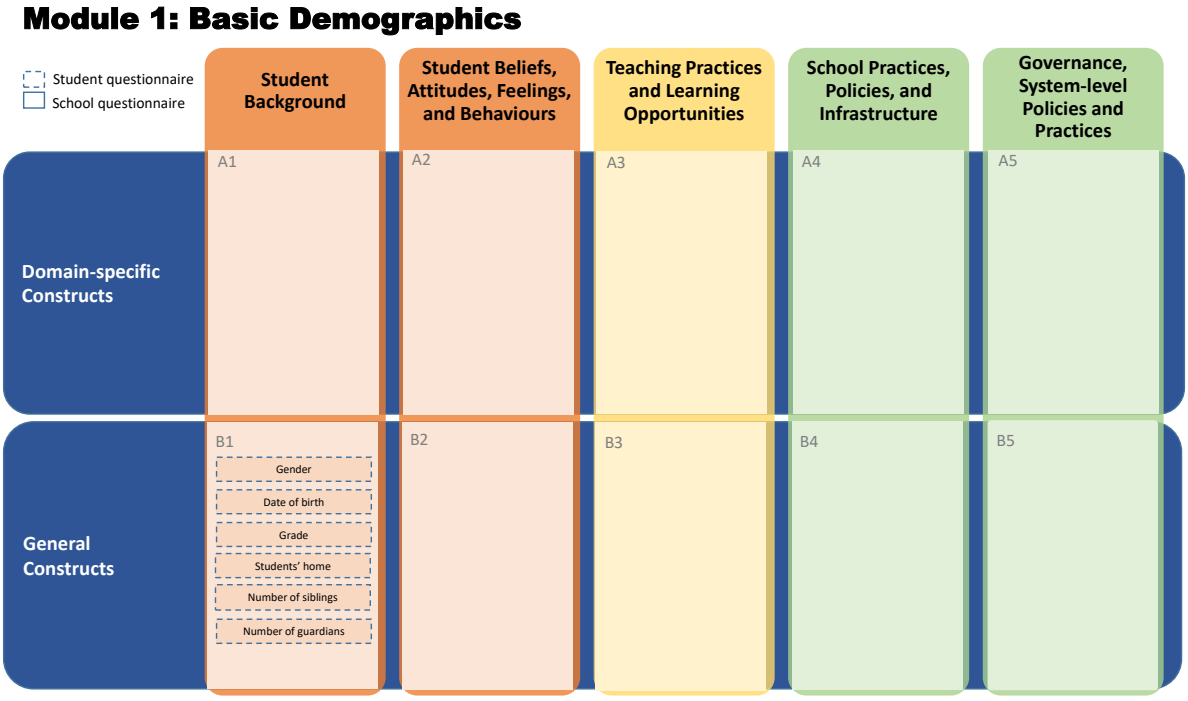


Figure 3. Constructs in Basic Demographics Module

4.2. Economic, Social, and Cultural Status (ESCS)

41. Over the past seven PISA cycles, significant efforts went into the definition and operationalization of individual student background indicators, leading to the establishment of an integrated indicator for

students' ESCS (Willms, 2006). Figure 4 displays how ESCS was created in the two most recent PISA cycles.

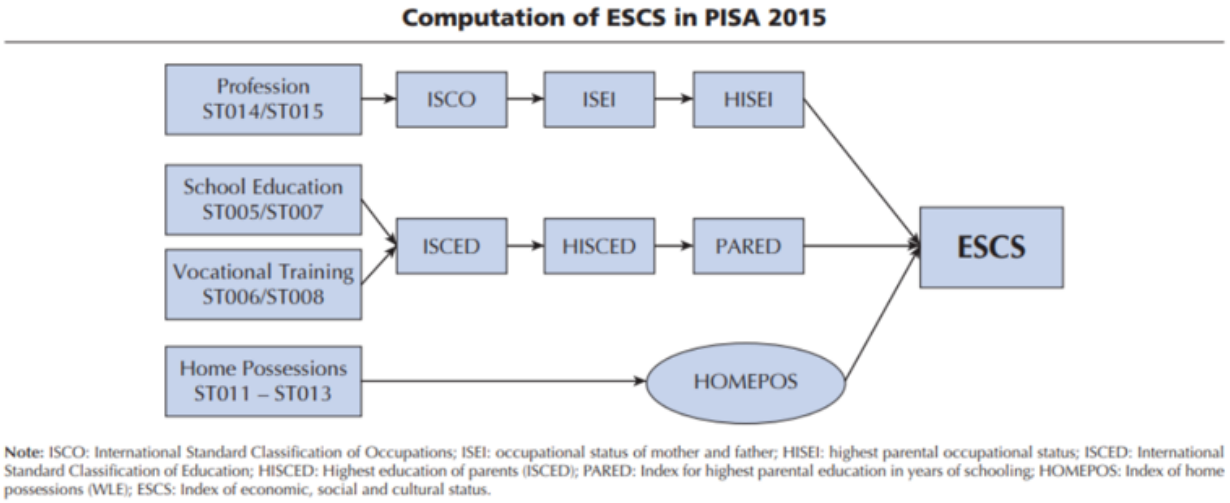


Figure 4. Computation of ESCS Index in PISA 2015 and 2018 (From PISA 2015 Technical Report)

42. The PISA ESCS index is considered internationally as a gold standard measure of socioeconomic status (SES) in LSAs (e.g. Cowan et al., 2012). To examine trends over time and comparisons with previous PISA data on the ESCS index, it is crucial to establish minimal stability in assessing the three components. While well established, the ESCS index has also been criticized in recent years (e.g. Rutkowski & Rutkowski, 2013), calling for revisions and extensions of the index.

43. Few changes have been made over the years to the measurement of ESCS in PISA, resulting in current approaches only partly accounting for students' living realities within and across the much more diverse PISA population. This issue becomes more pressing with the number of participating countries more than doubling over the past cycles. For instance, the current PISA ESCS questions continue to assume a traditional nuclear family with a mother and father, and give little to no room for students to provide information about their families' income and education levels if they live in non-traditional constellations (e.g. multiple households, same-sex parents, multi-generational households, etc.).

44. While used for several cycles, issues remain with the International Standard Classification of Occupations (ISCO) and International Standard Classification of Education (ISCED) coding of parental educational levels and occupations (Kaplan & Kuger, 2016) that pose challenges when making international comparisons on the respective questions. Recent findings from other studies further suggest that student reports on their parents' occupation tend to be very inaccurate, produce larger proportions of missing values, and that these questions take substantially more time to answer than other survey questions (e.g. Tang et al., 2017).

45. The PGB wishes to increase the benefits of participation in PISA for lower- and middle-income countries. The group further expressed a particular need to incorporate questionnaire items that fully reflect the context found in those countries. The broadening of the PISA population to new countries and the widened socioeconomic divides in some countries call for a better approach of assessing the entire range from low to high socioeconomic circumstances. Having common questions between the PISA-D student and out-of-school youth questionnaires and the PISA STQ could be one way of achieving that linkage.

46. Updates to the Index of ESCS will be suggested for PISA 2021, focusing especially on re-evaluating and updating parental education and occupation as well as home possession questions, and introducing routing independently within each dimension of ESCS.

47. PISA 2021 will explore alternative ways of collecting the information about the highest level of education among parents or guardians to address well-reported flaws of the current questions (e.g. inconsistency between "level of schooling" and "post-secondary qualifications", widespread over-reporting of parental education compared to national statistics, possible redundancy of asking about father and mother separately when the goal is to measure the "highest" level among parents or guardians). Note that in previous PISA cycles, information about education levels among parents or guardians have been based on ISCED 1997 classifications; beginning with PISA 2021, the more recent ISCED 2011 classifications will be used. Table 5 summarizes how the updated ISCED 2011 levels correspond to the ISCED 1997 levels. More detailed information about the correspondence or concordance between levels in the ISCED 2011 classification and the earlier ISCED 1997 framework can be found in the ISCED 2011 Operational Manual: Guidelines for Classifying National Educational Programmes and Related Qualifications (OECD, European Union, UNESCO Institute for Statistics, 2015).

Table 5. Correspondence between ISCED 2011 and ISCED 1997 Levels

ISCED 2011		ISCED 1997	
01	Early childhood educational development		--
02	Pre-primary education	0	Pre-primary education
1	Primary education	1	Primary education or first stage of basic education
2	Lower secondary education	2	Lower secondary education or second stage of basic education
3	Upper secondary education	3	(Upper) secondary education
4	Post-secondary non-tertiary education	4	Post-secondary non-tertiary education
5	Short-cycle tertiary education	5	First stage of tertiary education (not leading directly to an advanced research qualification) (5A, 5B)
6	Bachelor's or equivalent level		
7	Master's or equivalent level		
8	Doctoral or equivalent level	6	Second stage of tertiary education (leading to an advanced research qualification)

48. PISA 2021 will explore replacing open-ended parental occupation questions and developing alternative ways of collecting the information about the highest parent or guardian occupation potentially based on a multiple-choice approach.

49. PISA 2021 will also evaluate the continued relevance of current home possession questions and how additional poverty and wealth indicators used in PISA-D may be integrated with previously established PISA questions.

50. Furthermore, PISA 2021 will explore the feasibility of routing respondents through a more targeted set of questions about parental/guardian education, parental/guardian occupation, and home possessions based on student responses to previous questions in the questionnaire. The digital delivery platform will be utilized to create a more seamless and targeted experience for students from different household types

(e.g. single parent or guardian homes, traditional two-parent or guardian homes, homes with three or more parents or guardians) and different ESCS levels in all participating education systems. The introduction of routing, which aims both at making the STQ more accessible to students from all background and more efficient in terms of the data collection, will be carefully considered to ensure that ESCS data collected meets necessary standards to allow for comparisons with previous cycles.

51. Figure 5 illustrates a potential routing approach for empirical evaluation during the FT. Based on answers to a question in the basic demographics modules (see above) about the number of their parents or guardians (i.e. individuals that take care of the student and provide money or other resources), students would receive tailored questions addressing parental/guardian education and occupation. Questions would be tailored in two different ways:

- *First, depending on the number of their parents or guardians* (e.g. students with only one parent or guardian will only be asked to indicate the educational level and occupation of one that parent or guardian, while students with more parents or guardians will be asked about multiple parents or guardians); and
- *Second, based on previous student responses in each category about the education levels of their parents or guardians* (e.g. if students indicate that none of their parents or guardians completed lower secondary education, they would receive additional questions about more basic educational outcomes).

52. As illustrated in Figure 5, the routing principle may also be applied to the collection of home possession data from students. For instance, questions about the presence of specific types of digital devices with screens (e.g. televisions, computers, tablets, smartphones) or types of books (e.g. dictionaries, books of art and design, technical reference books, classical literature) administered to all students in PISA 2000-2018 may be presented only to students who indicated having access to at least one digital device with a screen or at least one book, respectively. Such routing rules may be established during the FT or based on FT data for the MS.

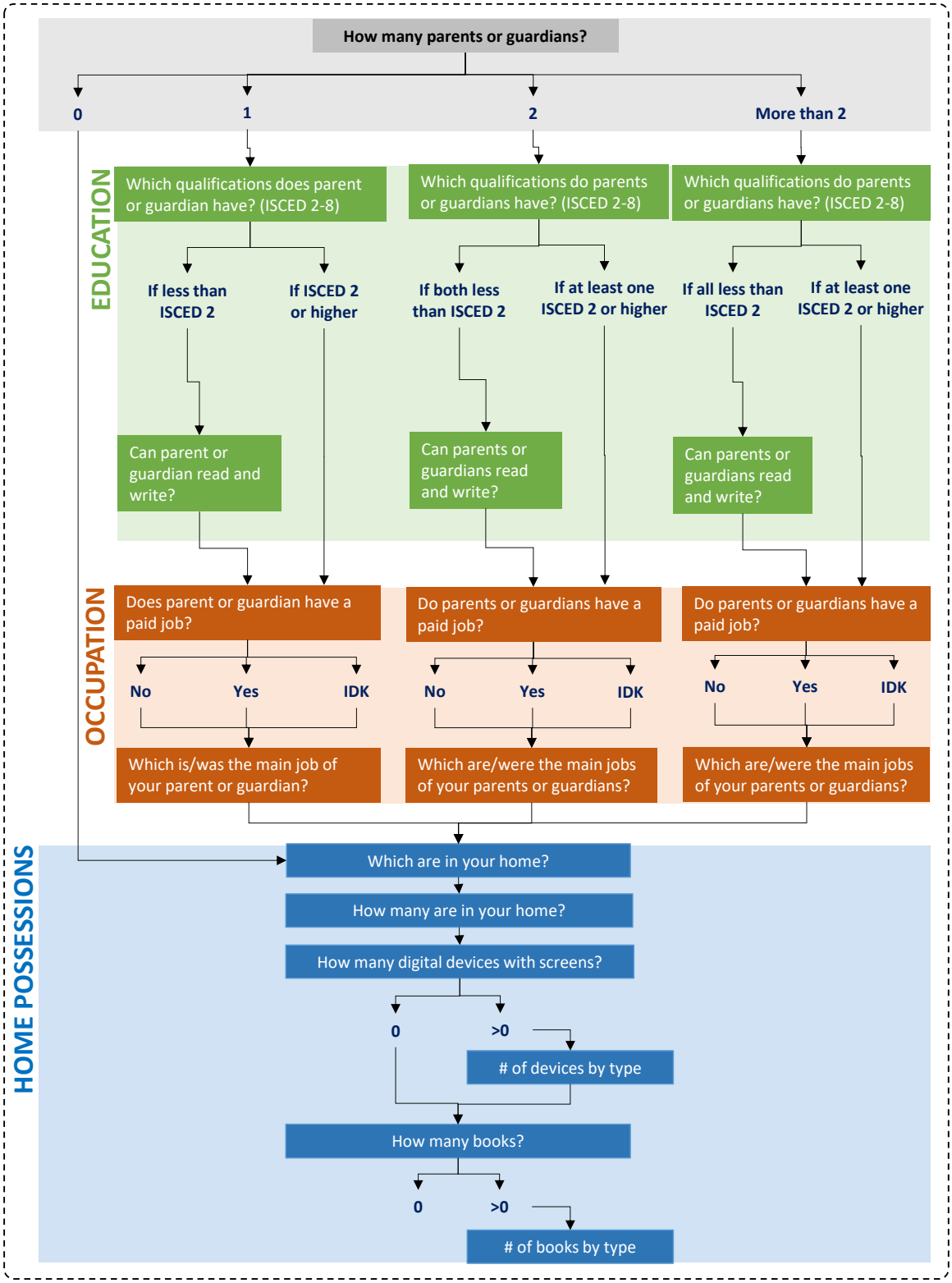


Figure 5. Illustration of Potential Routing Approach for ESCS

53. Beyond these proposed updates to the creation of the Index of ESCS, PISA 2021 also bears a chance to explore the feasibility of measuring extensions of the ESCS construct to gain a broader perspective of students' learning environments and access to educational resources. To reflect on the large number of new lower- and middle-income countries that will join PISA in 2021 and the larger range of ESCS values that PISA 2021 will need to capture, we suggest adding a few additional items focusing on measuring the lower ends of the ESCS continuum. These may include whether parents or guardians can read or write, several updates to the home possessions questions, and a measure of food insecurity based on questions previously validated in PISA-D.

54. In addition, research on subjective SES suggests that student's subjective beliefs about their own and their family's status can be as important as objective SES measures in predicting important outcomes, ranging from achievement and overall future aspirations, to obesity and other health outcomes (e.g. Citro & Michael, 1995; Demakakos, Nazroo, Breeze, & Marmot, 2008; Goodman et al., 2001; Lemeshow et al., 2008; Quon & McGrath, 2014). The most common approach for measuring subjective SES is Cantril's Self-Anchoring Ladder (Cantril, 1965; see Levin & Currie, 2014, for an adaptation for adolescents). It has been used in several variations, including extensions to subjective social status within the school community (Goodman et al., 2001). A subjective SES measure would complement rather than replace the established ESCS indicator in PISA.

55. Figure 14 below illustrates how all proposed constructs in this module map on the taxonomy.

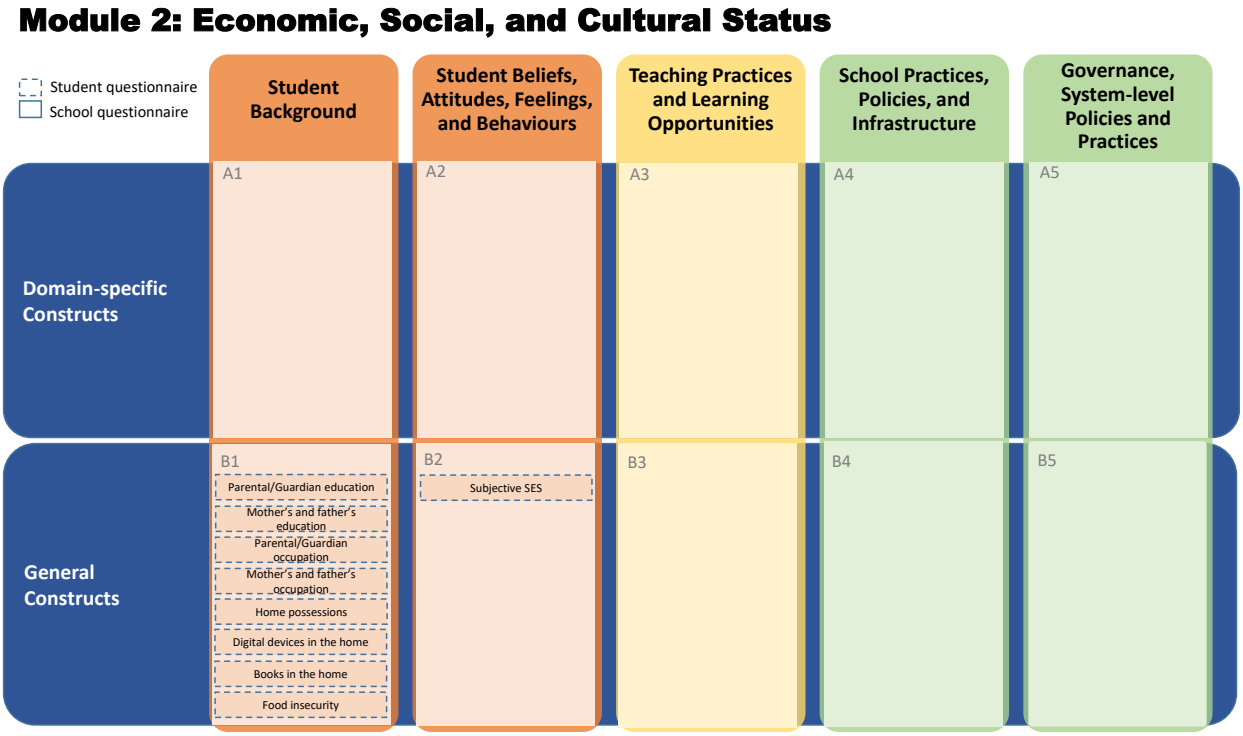


Figure 6. Constructs in ESCS Module

4.3. Educational Career

56. PISA gathers retrospective and prospective information about students' early years, education pathways, and careers across their lifespan. Researchers and public debates in many countries have stressed the importance of early childhood education (Blau & Curie, 2006; Cunha, Heckman, Lochner, & Masterov,

2006). PISA 2021 will continue this tradition to capture some essential information on primary and pre-primary education (bearing in mind that, for the most part, this would be solicited from 15-year-olds or their parents, which may pose validity challenges).

57. Constructs measured in the STQ under this module are considered primarily as general constructs (e.g. attendance of ISCED 0-2; current study programme; history of students repeating a grade; missing, skipping, or arriving late to school).

58. Figure 7 below illustrates how all proposed constructs in this module map on the taxonomy.

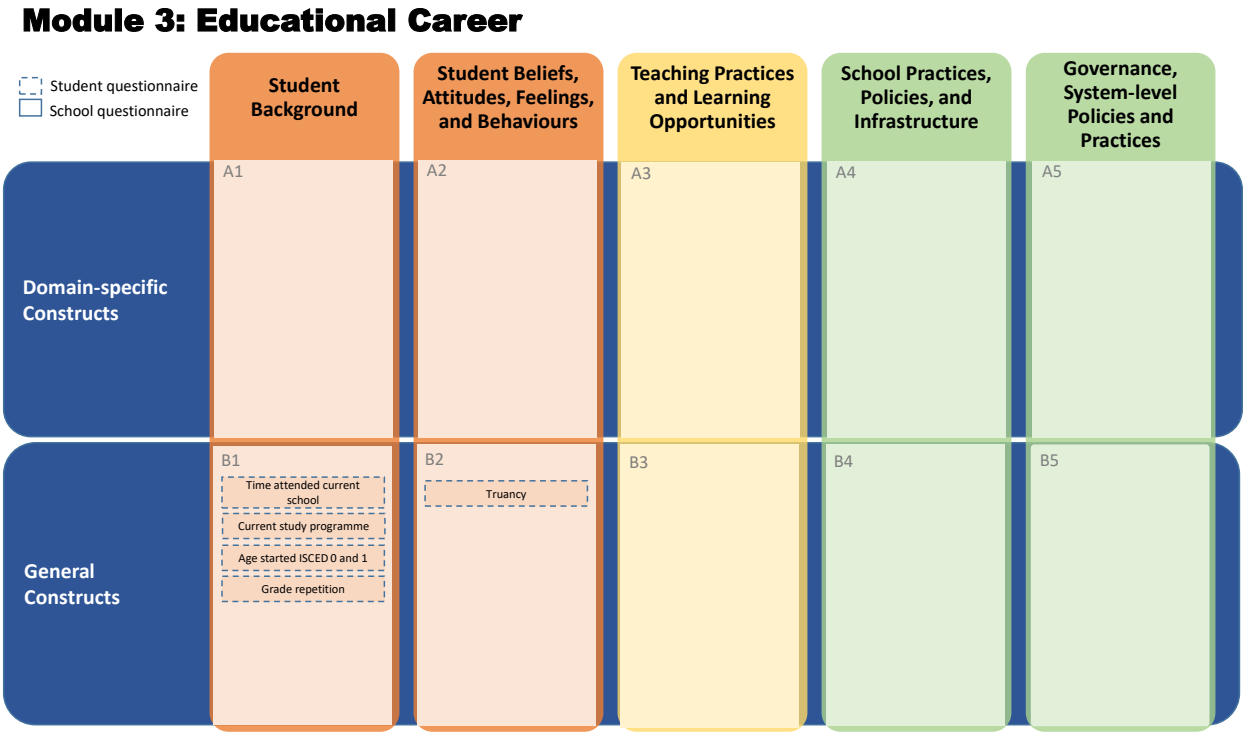


Figure 7. Constructs in Educational Career Module

4.4. Migration and Language Exposure

59. Select aspects of students' migration background and language exposure have been captured in previous PISA STQs, as well as optional questionnaires (e.g. acculturation in the 2012 Educational Career Questionnaire). Immigration is currently a critical topic in many countries, particularly those with traditionally larger immigrant populations (e.g. the United States, Canada) as well as countries facing new challenges due to new populations of refugees (e.g. most central European countries) (Bansak, Hainmueller, & Hangartner, 2016; Wike, Stokes, & Simmons, 2016). Issues regarding the student's experience of a school climate that is accepting of diversity and multiculturalism are relevant to this module and overlap with content examined in the module on *School Culture and Climate.*

60. Student demographic questions in this module will focus on assessing students' migration backgrounds (e.g. country of origin, age of arrival in country), and language backgrounds (e.g. primary language spoken at home, age of learning the test language). Additional constructs are proposed for inclusion in the FT SCQ to gain a deeper understanding of the resources that schools have available to support the learning of immigrant student populations, challenges that schools may encounter in supporting immigrant students, and how factors related to immigration and cultural diversity relate to overall school

climate. General constructs recommended for the FT SCQ include the proportion of students with a migration background (e.g. immigrant or refugee status), number of instructional languages, and instruction of students learning the test language.

61. Figure 8 below illustrates how all proposed constructs in this module map on the taxonomy.

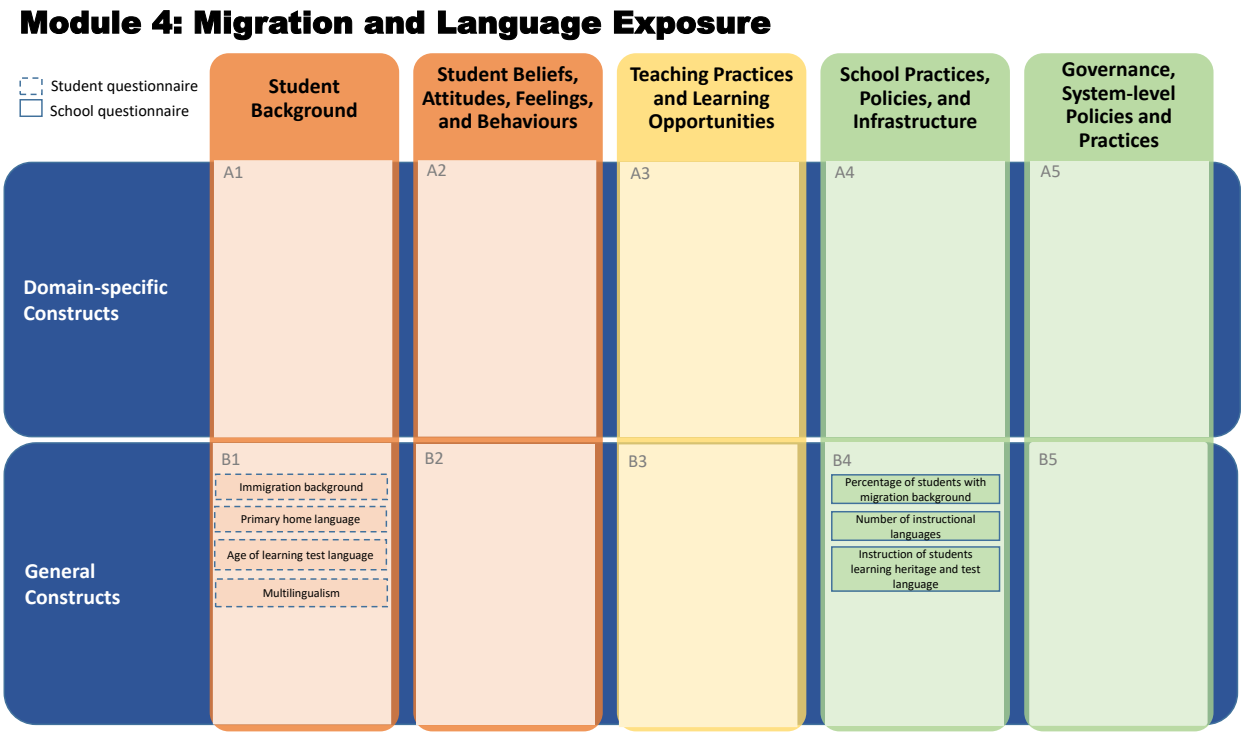


Figure 8. Constructs in Migration and Language Exposure Module

4.5. PISA Preparation and Effort

62. Several researchers have investigated the question whether test-taker effort on low-stakes LSAs may impact achievement results or have asked whether differential effort may play a role in explaining score differences between student groups or educational systems (e.g. Debeer, Buchholz, Hartig, & Janssen, 2014; Eklöf, Pavešič, & Grønmo, 2014; Hopfenbeck & Kjaernsli, 2016; Jerrim, 2015; Penk, 2015).

63. In an effort to inform educational policy with regard to test-taker effort in PISA, this module covers students' subjective perceptions of how much effort they applied when answering the PISA test questions in mathematics, reading, or science, as well as filling out the STQ. Questions will draw on the idea of the "effort thermometer" introduced in PISA 2003 (Butler and Adams, 2007). To complement questions examining students' perceptions of effort, a new school question will examine administrators' support of the PISA test administration; their communication with staff, students, parents or guardians, and teachers about PISA; and their encouragement of students to take the PISA test seriously.

64. Figure 9 below illustrates how all proposed constructs in this module map on the taxonomy.

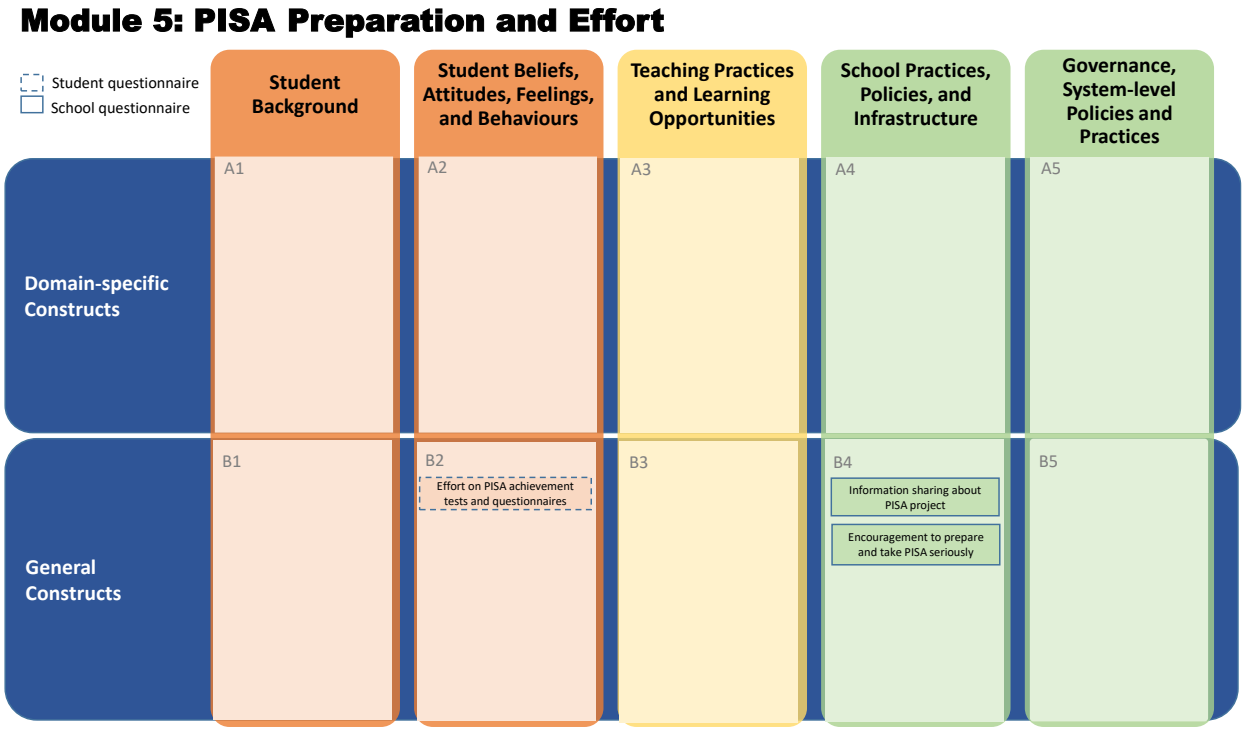


Figure 9. Constructs in PISA Preparation and Effort Module

4.6. School Culture and Climate

65. School climate, safety, and student well-being are important antecedents of academic achievement (Kutsyuruba, Klinger, & Hussain, 2015). School climate encompasses shared norms and values, the quality of relationships, and the general atmosphere of a school (Loukas, 2007) and is often described as the quality and character of school life that sets the tone for all the learning and teaching done in the school environment. An academic focus—that is, a general consensus about the mission of the school and the value of education, shared by school leaders, staff, and parents or guardians—affects the norms in student peer groups and facilitates learning (Opdenakker & Van Damme, 2000; Rumberger & Palardy, 2005). Research shows that positive school climate contributes to immediate student achievement and endures for years (Hoy, Hannum, and Tschannen-Moran, 1998). A positive school climate is associated with student's motivation to learn (Eccles et al., 1993) and has been shown to moderate the impact of socioeconomic context on academic success (Astor, Benbenisty, & Estrada, 2009). Lastly, the relationships that a student encounters at all levels in school (including students' views of the quality of teacher-student support and student-student support) also have an effect on student achievement (e.g., Jia et al., 2009).

66. Closely related to school climate is the safety of the learning environment. An orderly, safe, and supportive learning atmosphere maximizes attendance and the use of learning time. By contrast, a learning environment characterized by disrespect, unruliness, bullying, victimisation, crime, or violence can act as a barrier to students' learning and distract from the school's overall mission and educational goals. In the area of safety, schools without supportive norms, structures, and relationships are more likely to experience violence and victimization, which is often associated with reduced academic achievement (Astor, Guerra, & Van Acker, 2010).

67. Learning in 21st century schools in many countries differs from traditional settings in terms of the diversity of the student population—for instance, diversity in racial/ethnic and cultural backgrounds, as

well as diversity in individual student characteristics and diversity of thought. Experiences with diversity in the classroom may take the form of interpersonal interactions on campus, larger classroom discussions, or diversity-related coursework or workshops. In the United States context, researchers have found that several types of diversity experiences are associated with improvements in students' academic outcomes and cognitive development (e.g. development of critical thinking and problem solving skills). Positive diversity experiences also play an important role in fostering students' social and emotional characteristics, such as tolerance, empathy, and curiosity (e.g. Bowman, 2010; Gurin, Dey, Gurin, & Hurtado, 2004; Gurin, Dey, Hurtado, & Gurin, 2002; Milem, Chang, & Antonio, 2005; Pettigrew & Tropp, 2006).

68. General constructs recommended for the PISA 2021 FT STQ include students' subjective perceptions as well as their values and beliefs about their in-school experiences. Measures may be drawn from previously included constructs (e.g. sense of belonging, bullying experiences, school safety, and teacher support) as well as new constructs (e.g. quality of student-teacher relationships; positive and negative affect at school; students' perception of their subjective standing in the school community). Constructs suggested for inclusion in the SCQ include a school's efforts to promote school diversity/multi-cultural views, school climate-related factors hindering instruction, and negative school climate. Questions in this module show some conceptual overlap with domain-specific questions in other modules (e.g., disciplinary climate in Module 17).

69. Figure 10 below illustrates how all proposed constructs in this module map on the taxonomy.

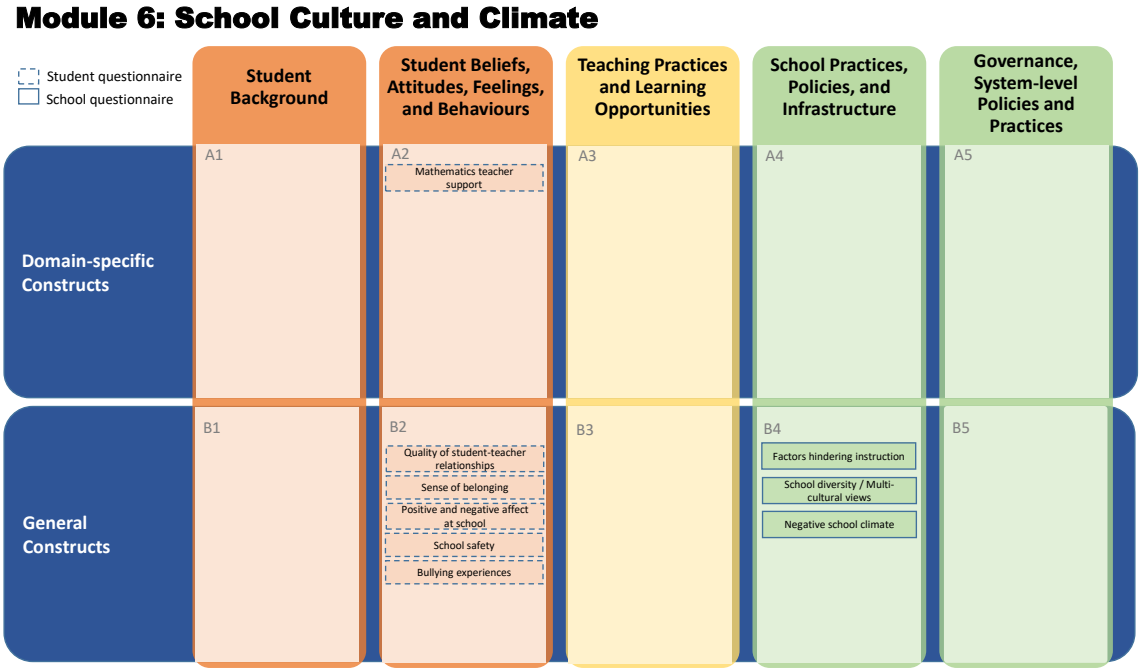


Figure 10. Constructs in School Culture and Climate Module

4.7. Subject-specific Beliefs, Attitudes, Feelings, and Behaviours

70. This module covers students' subjective perceptions as well as their values and beliefs, feelings and behaviours that are specific to mathematics, reading, and science. While a small set of key questions for each content-domain will be included in the PISA 2021 FT, the focus of this module is on mathematics-

related questions. Please note, contextual factors for a module focused on creative thinking are not described in this framework. Instead, information about these constructs is provided in the PISA 2018 Creative Thinking framework.

71. Questions related to all three domains may include students' favourite subjects; whether students are motivated to achieve highly in mathematics, reading, and science; whether they think mathematics, reading, and science are easy for them; and the extent to which students think of skills in the three subjects, as well as their general intelligence, creativity, and social skills, as something malleable through effort (growth mindset) or something largely robust to change (fixed mindset).

72. In addition, a combination of new mathematics-specific questions and questions retained from previous PISA cycles are recommended for this module. PISA 2012, for instance, assessed a large number of mathematics-specific beliefs, attitudes, feelings, and behaviours. Four PISA 2012 scales (mathematics self-efficacy, mathematics anxiety, confidence in knowledge of mathematics concepts, and mathematics self-concept) were among the five constructs with consistently strongest correlational relationship with academic achievement in PISA 2012 (Lee & Stankov, 2018). Based on these findings, measures for these constructs are considered for inclusion also in PISA 2021. Not all constructs, however, should be re-administrated without revisions and adjustments. On a trait level, mathematics self-efficacy, confidence, and self-concept are largely redundant (e.g. Marsh et al., in press), a finding confirmed by PISA 2012 data when looking at joint relationships with achievement of these constructs. For the PISA 2021 FT, the PISA 2012 self-efficacy scale will be retained and expanded by adding additional mathematics-reasoning related skills to the list of knowledge and skills. Self-efficacy will be prioritized due to the concrete nature of the items that allow for clearer, more objective reporting than the agree/disagree type self-concept items used in PISA 2012. This difference in cross-cultural comparability of the two measures is reflected also in the finding that PISA 2012 self-efficacy showed consistently positive relationships with achievement both within and across countries, whereas relationships for self-concept were affected by the so-called "attitude-achievement-paradox" (see Figure 25 in Section 5. of this framework). Rather than creating a second, largely redundant, scale focusing entirely on mathematics self-concept, this construct will be operationalized for all three core PISA domains (mathematics, reading, and science) to allow for new insights based on potentially examining data as a profile across the three domains. Another recommended extension of the range of mathematics-related constructs for PISA 2021 is a greater focus on a balanced set of emotions, not limited primarily to anxiety but also focused on other emotions prevalent in students' affective lives today, such as boredom, excitement, or interest (Pekrun, 2017; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, 2017). Lastly, a new scale targeting students' invested effort and persistence in mathematics work (including homework) will provide actionable data for educators and policymakers that goes beyond the more subjective scales tapping into motivation in previous cycles.

73. Figure 11 below illustrates how all proposed constructs in this module map on the taxonomy.

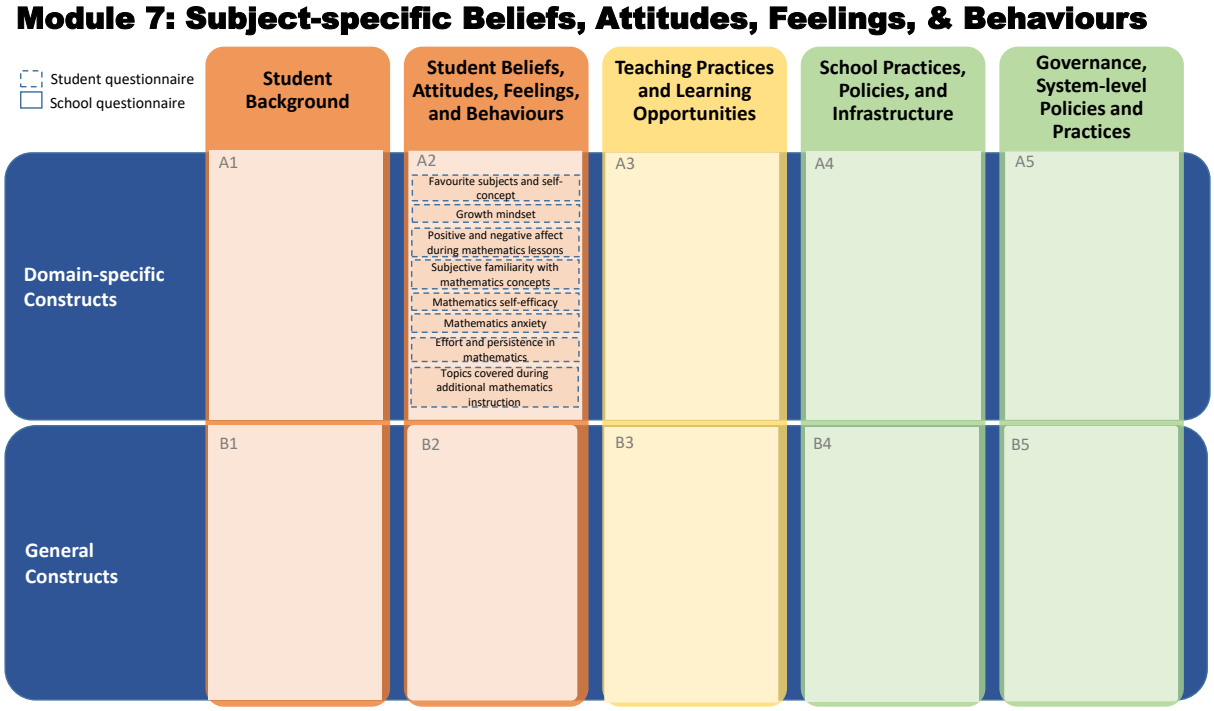


Figure 11. Constructs in Subject-specific Beliefs, Attitudes, Feelings, & Behaviours Module

4.8. General Social and Emotional Characteristics

74. Unlike the constructs listed above, constructs in this module are not primarily school-related, but can be understood more broadly as characteristics indicative of student preparedness and social and emotional characteristics relevant to students' achievement in high school and throughout their lifetime. Two main framework approaches tend to be used to conceptualize social and emotional characteristics: one anchored to the personality psychology literature, which commonly refers to a "Big Five" taxonomy of personality traits; the other anchored to the social psychology literature, which focuses on cognitive constructs like motivations, beliefs, goals, interests, and values. PISA 2021 will expand on these efforts by integrating the PISA framework with OECD's *Study of Social and Emotional Skills* (SSES, OECD, 2017b) to help policymakers and educators better link PISA data with other established frameworks and data sources. Based on the SSES framework, social and emotional characteristics can be defined as individual capacities that (a) are manifested in consistent patterns of thoughts, feelings, and behaviours, (b) can be developed through formal and informal learning experiences, and (c) influence important socioeconomic outcomes throughout individual's life. All general social and emotional characteristics measured in the PISA 2021 FT can be mapped on the OECD SSES taxonomy (OECD, 2017b).

75. *Task performance* describes different aspects of students' conscientiousness and their striving for task performance, including setting high standards for themselves and working hard to meet them, fulfilling commitments and being reliable, being able to avoid distractions and focus attention on tasks, and persevering in the face of difficulty to complete tasks.

76. *Emotional regulation* covers different aspects of students' experienced range of emotions and their emotional regulation, including their ability to handle stress well, and regulate their temper, anger, and irritation in the face of frustrations.

77. *Collaboration* covers different aspects of students' approaches to collaboration, specifically their levels of agreeableness, including being kind and caring for others and valuing and investing in close relationships, building trust with others, as well as students' desire to value interconnections among people in general.

78. *Open-mindedness* covers different aspects of students' open-mindedness and openness to new experiences, including their desire to learn and approach situations with an inquisitive mindset, openness to different points of view and diversity, as well as enjoyment of generating novel ideas or visions.

79. *Engaging with others* covers different aspects of students' extraversion and their engagement with others, including their enjoyment of initiating and maintaining social connections, assertiveness in voicing their own views and exert social influence, as well as their tendency to approach daily life with energy, excitement, and spontaneity.

80. To prioritize constructs and identify items for inclusion in the PISA 2021 FT, draft instruments that have been developed for the pilot administration of OECD's SSES (OECD, 2017b) were considered to make use of synergies between different OECD surveys as also recommended by the PGB (PISA Governing Board, 2017) with the goal of at least 5 items per construct being identical between the SSES FT and the PISA 2021 FT.

81. One goal for the PISA 2021 MS is including at least one construct representing the five clusters of constructs described in the OECD's taxonomy while at the same time giving more weight to constructs that can be described as student characteristics that are malleable and can be taught and fostered in a school context and, moreover, relate positively to general life outcomes and/or dispositions relevant to workforce success. The PISA 2021 FT will therefore include a larger number of constructs in order to select final constructs for the MS based on FT data.

82. Constructs suggested for inclusion in the PISA 2021 FT are Perseverance and Self-control (both representing the *Task performance* cluster), Stress resistance and Emotional control (representing the *Emotional regulation* cluster), Curiosity and Perspective taking (representing the openness cluster), Cooperation, Empathy, and Trust (representing the *Collaboration* cluster), and Assertiveness (representing the *Engaging with others* cluster). Please note, in addition to these constructs, the student well-being questionnaire (SWBQ, not described in this framework) includes a range of constructs related to each of the Big Five factors, and the Creative Thinking focused module (not described in this framework) to be included with the core STQ will aim to capture additional facets of openness.

83. Figure 12 below illustrates how all proposed constructs in this module map on the taxonomy.

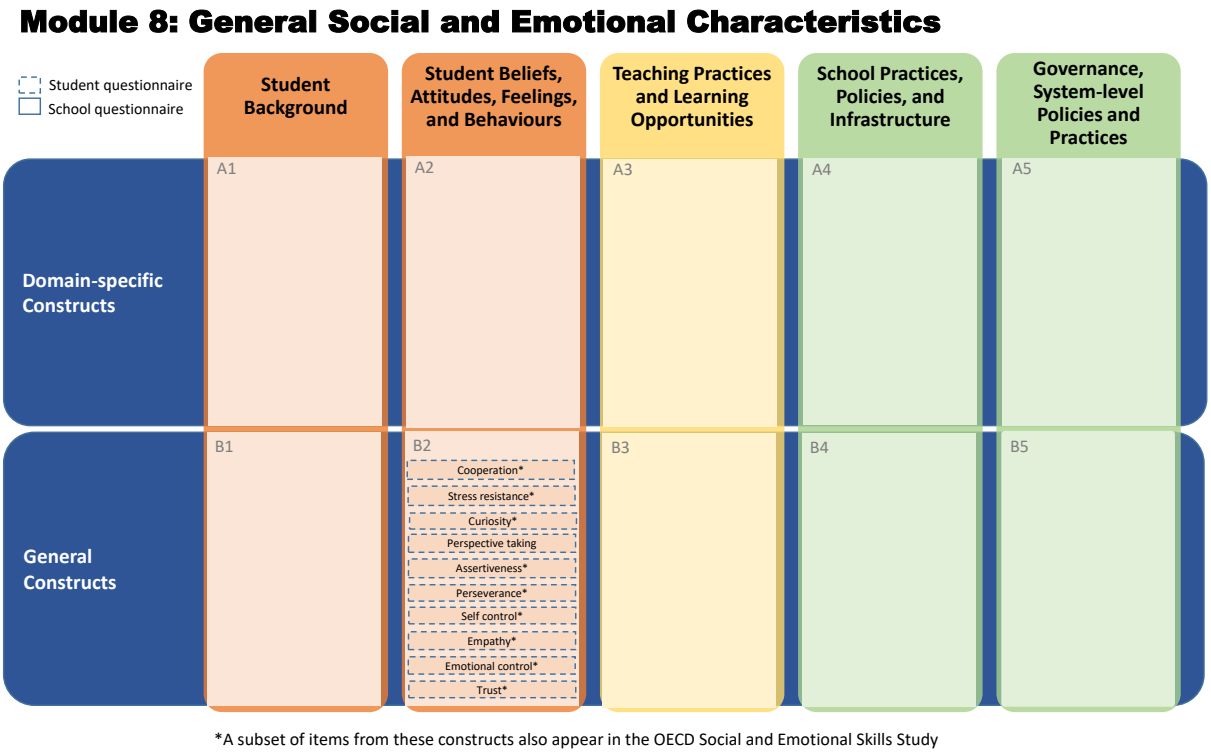


Figure 12. Constructs in General Social and Emotional Characteristics Module

4.9. Health and Well-being

84. PISA 2015 and 2018 started to include questions about health and well-being in the core STQ, and PISA 2018 offered an additional optional student well-being questionnaire (SWBQ) that gathered in-depth data on student well-being in participating countries. PISA 2021 will carry these developments forward and include, in addition to the optional SWBQ, a small module of health- and well-being related questions in the core STQ. Constructs for this module were chosen to avoid any redundancies with the SWBQ and further prioritize well-being related questions that are important to capture student attitudes, feelings, and behaviour in all participating countries. These include students' overall life satisfaction, online activities, and potentially problematic online behaviours (e.g., extensive time spent on social networks and/or video games). In addition, questions in other modules (e.g., school culture and climate, general social and emotional characteristics, out-of-school experiences, physical exercise) will yield data that informs constructs that may be conceptualized also as part of health and well-being (e.g., activities before and after school, sense of belonging, bullying, and student-teacher relationships).

85. Figure 13 below illustrates how all proposed constructs in this module map on the taxonomy.

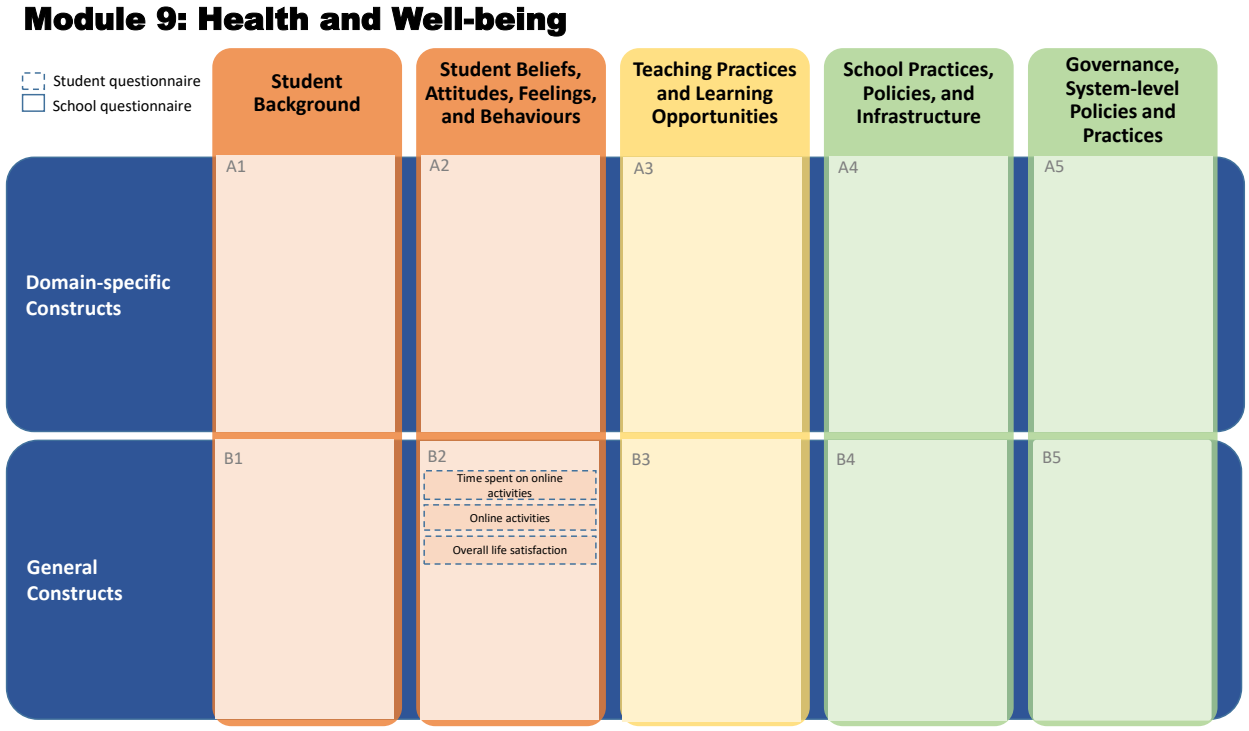


Figure 13. Constructs in Health and Well-being Module

4.10. Post-secondary Preparedness and Aspirations

86. In addition to collecting data on students' early educational careers, previous PISA cycles have gathered prospective information about students' future education pathways and preparation, and their occupational aspirations. While research in the United States has found that interpersonal relationships (e.g. peers, parents or guardians, teachers and staff who provide career guidance) play a significant role in shaping students' educational aspirations, cross-cultural research suggests that these influences may largely depend on the structural features of the educational systems in which they operate. For instance, peers and parents or guardians tend to influence educational aspirations in countries with undifferentiated secondary schooling, but this influence appears to be weaker in countries with more differentiated secondary education (Buchmann & Dalton, 2002). It is possible that in differentiated systems, these effects may be indirect and mediated by early school-related decisions, such as track enrolment. An important factor to consider in understanding students' educational and work aspirations is the role that the school has in shaping these goals—for instance, through students' participation in the curriculum and activities offered by the school, and the provision of additional resources to explore educational and occupational pathways (e.g. Beal & Crockett, 2010).

87. Constructs measured in the STQ (e.g. students' exposure to information about future studies or work; students' education and career expectations) and SCQ (e.g. school's offerings and support in providing information to students about future work and career paths) under this module are considered primarily as general constructs.

88. Figure 14 below illustrates how all proposed constructs in this module map on the taxonomy.

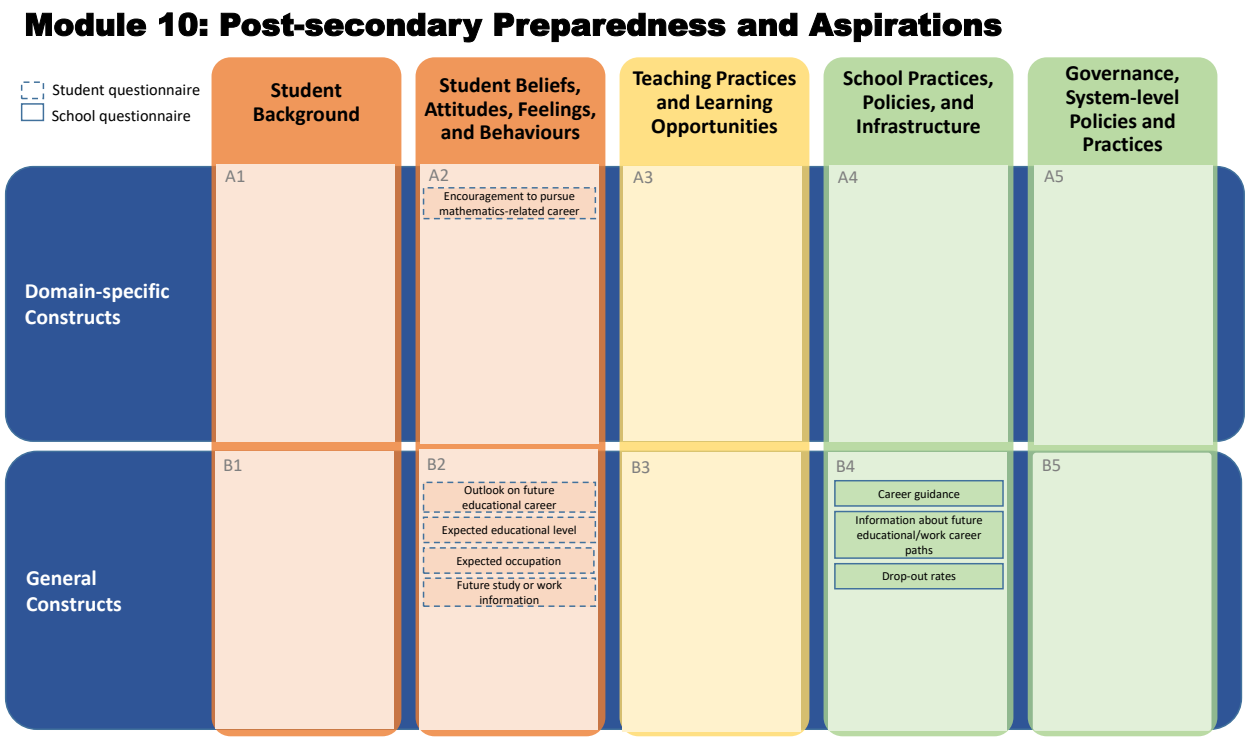


Figure 14. Constructs in Post-secondary Preparedness and Aspirations Module

4.11. Out-of-school Experiences

89. While classrooms serve as important settings for students' engagement in opportunities to learn, student engagement and learning also occur through formal and informal opportunities to learn outside of school. In the 2015 and 2018 questionnaire frameworks, students' out-of-school experiences focused on domain-specific indicators. The PISA 2021 framework takes a broader view on out-of-school experiences including both academic and non-academic experiences that may fall into several of the defined educational policy areas, including student background, student attitudes, feelings, and behaviours, and school practices, policies, and infrastructure.

90. How students spend their time outside of school, and the extent to which they engage in learning-related activities outside of school (e.g. tutoring, extracurricular activities, homework, mathematics-related activities), are important for understanding student achievement. Studies have shown that students' time use outside of school relates to mathematics achievement across several countries (Fuligni & Stevenson, 1995), and engagement in extracurricular activities is associated with lower dropout rates for at-risk students, improved grade point averages, and higher educational aspirations (Broh, 2002; Mahoney & Cairns, 1997). Out-of-school activities can also provide important opportunities to learn, whereby students can apply subject-related content and skills that have been emphasized in class to novel situations. This may be especially true for populations that have less exposure to formal education, as well as countries where structured out-of-school learning activities are prevalent (e.g. after-school tutoring to supplement and enhance in-school learning).

91. Domain-specific constructs recommended for the FT STQ include students' participation in additional mathematics lessons outside of school and tutoring, and time spent on mathematics homework. Domain-specific constructs recommended for the FT SCQ include administrators' reports of the school

offering additional lessons and tutoring arrangements in mathematics. General constructs recommended for the FT STQ include activities before and after school (including physical activities, working for pay, or eating breakfast); general constructs recommended for the FT SCQ include extracurricular activities offered by the school.

92. Figure 15 below illustrates how all proposed constructs in this module map on the taxonomy.

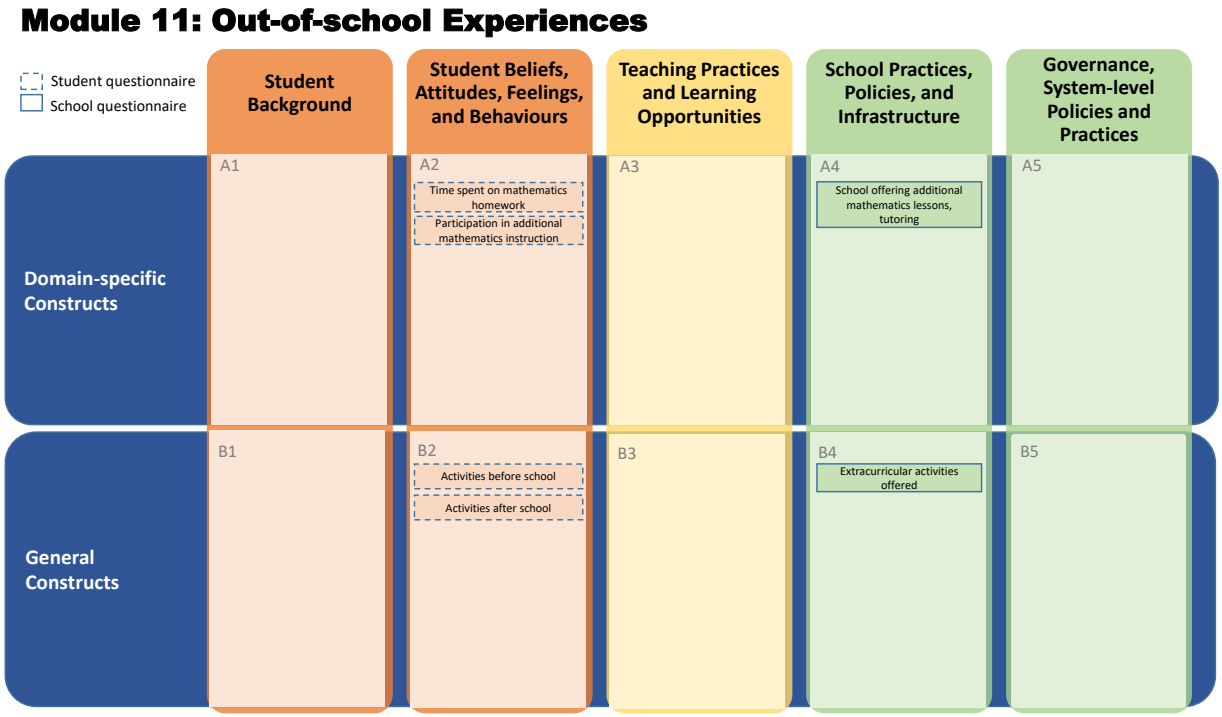


Figure 15. Constructs in Out-of-school Experiences Module

4.12. School Type and Infrastructure

93. This module examines aggregate school-level characteristics of the students' learning environment (e.g. location, type, and size of the school) and school risk factors that may hinder student learning and achievement with regard to the physical set-up of the school, such as deficiencies in school resources and infrastructure. The quality of a school's infrastructure, and the quality and accessibility of digital educational resources (e.g. computers and other digital technology, Internet access) may facilitate or hinder the learning environment's positive impact, and in turn, influence achievement.

94. Conceptually, this module overlaps with other modules measuring the overall characteristics of the school and school population, including those capturing school culture and climate (Module 6), organization of student learning at school (Module 15); assessment, evaluation, and accountability (Module 19); and school autonomy (Module 14). General constructs recommended for the FT SCQ include school size (teachers, students, and non-teaching staff), school type and the type of organization running the school, school location, whether the school hosts visiting teachers from other countries, availability of digital technology, and lack of physical and digital infrastructures.

95. Figure 16 below illustrates how all proposed constructs in this module map on the taxonomy.

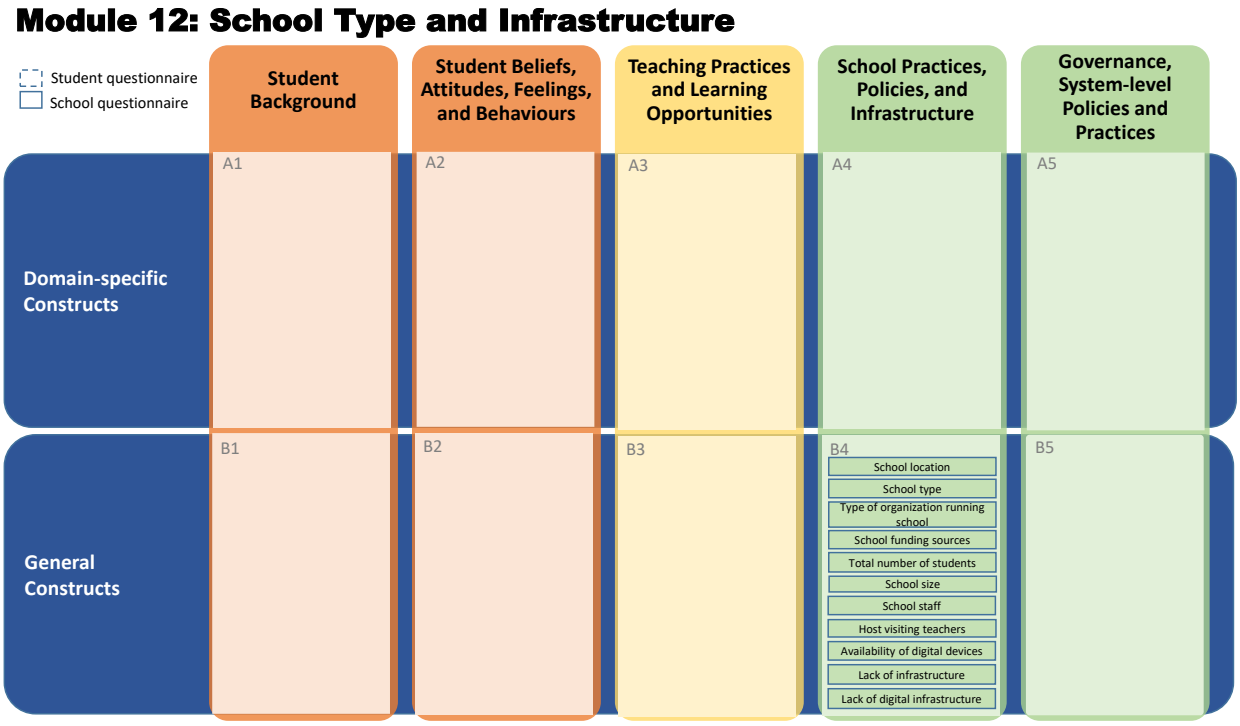


Figure 16. Constructs in School Type and Infrastructure Module

4.13. Selection and Enrolment

96. School principals and administrators play a key role in school management and policy, as they are often seen as the primary agents of change to improve student achievement in their schools. They can shape teachers' professional development, define the school's overall mission and educational goals, ensure that instructional practices and policies within and across subjects are directed towards achieving these goals, suggest modifications to improve teaching practices, and help solve problems that may arise within the classroom or among teachers.

97. The way in which students are channelled into educational pathways, schools, tracks, or courses (also known as stratification, streaming, or tracking) is a core issue of educational governance and is an important aspect of school organization and policy. For instance, highly selective schools provide a learning environment that may differ from the environment offered by schools that are more comprehensive. Some longitudinal studies have demonstrated grade retention harms individual careers and outcomes (e.g. Griffith, Lloyd, Lane, & Tankersley, 2010; Ou & Reynolds, 2010), as well as student behaviour and well-being (e.g. Crothers et al., 2010), while other research finds positive effects (Marsh et al., 2017). Greene and Winters (2009) showed that once a test-based retention policy has been installed, those who were exempted from the policy did worse. Additionally, Babcock and Bedard (2011) showed that a large number of students being retained could have a positive effect on the cohort (i.e. all students, including those who are promoted). Kloosterman and De Graaf (2010) argued that in highly tracked systems, such as in some European countries, grade repetition might serve as a preferred alternative to moving into a lower track. The authors found evidence that this strategy is preferred for students with higher SES. Thus, changing grade repetition policies might be a viable option regarding low-cost interventions (Binder, 2009).

98. General constructs recommended for the SCQ include the school's selection competition, percentage of students who have repeated a grade, academic selectivity, and student transfer policies.

99. Figure 17 below illustrates how all proposed constructs in this module map on the taxonomy.

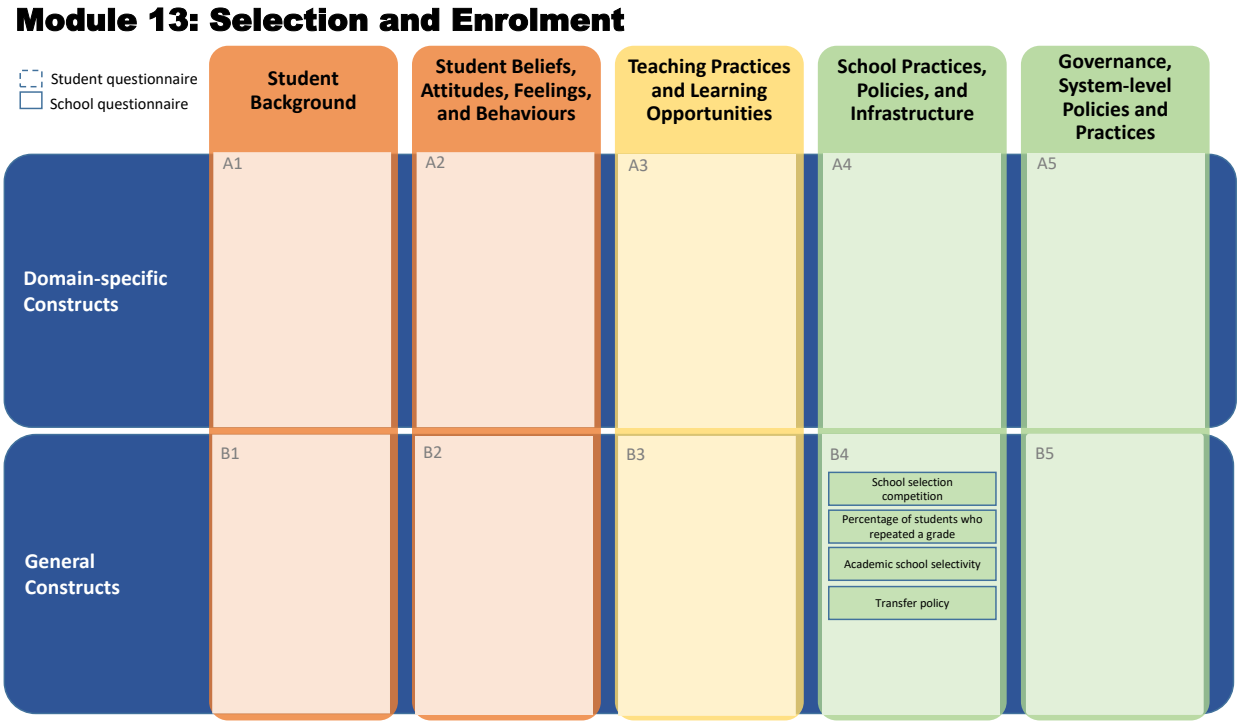


Figure 17. Constructs in Selection and Enrolment Module

4.14. School Autonomy

100. Education systems have been classified by the amount of control or local autonomy that is given to schools (i.e. the school board, staff, and school leaders) versus governing bodies at the local, regional, or national level when decisions on admission, curriculum, allocation of resources, and personnel have to be made. These indicators have been previously included in the PISA 2012 SCQ and are revisited in 2021. Domain-specific constructs recommended for measurement in the FT SCQ include administrators' reports of centralized versus local autonomy in setting the school's mathematics curriculum. General constructs recommended for measurement in the FT SCQ include administrators' reports of the primary responsibility for school decision making. As previously noted, the possibility of using OECD system-level data collection to collect complementary information will be explored.

101. Figure 18 below illustrates how all proposed constructs in this module map on the taxonomy.

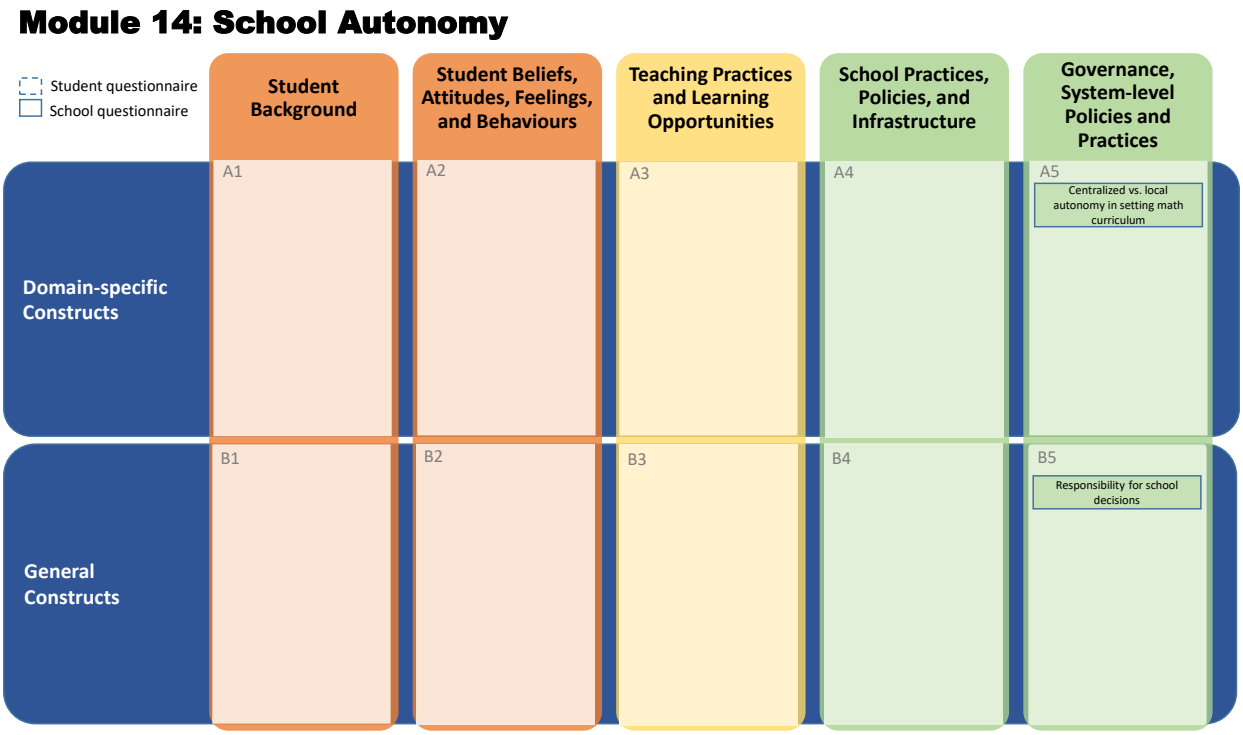


Figure 18. Constructs in School Autonomy Module

4.15. Organization of Student Learning at School

102. Large portions of students' educational experiences tend to occur at school in the classroom environment. During time spent in the classroom, students are exposed to subject content, curriculum materials, instructional strategies, skills, and a diversity of backgrounds and perspectives contributing to overall climate. Learning time and the intended curriculum content in school have been found to be closely related to student outcomes (e.g. Abedi et al., 2006; Cogan, Schmidt, & Guo, in press; Scherff & Piazza, 2008; Schmidt & Maier, 2009). Overall students' learning time and achievement are correlated as the time allowed for learning constrains students' opportunities to learn, though there are large differences within countries, across countries, and among different groups of students and schools (Ghuman & Lloyd, 2010; OECD, 2011). A generally positive relationship has been replicated in international comparative research (e.g. OECD, 2011; Martin, Mullis, & Foy, 2008; Schmidt et al., 2001; Schmidt & Burroughs, 2016; Schmidt, Burroughs, Zoido, & Houang, 2015).

103. Related to learning time is the way intended learning content is designed, structured, and communicated during that time in school. Understanding how a school curriculum functions requires a consideration of how it is organized and how students gain access to it. For example, a school's curriculum can be understood by examining what coursework is required and optional; whether students are tracked or grouped by achievement; and what standards are used to develop subject content. Curriculum may vary largely across tracks, grades, schools, and countries (Schmidt et al., 2001; Martin et al., 2008). Overall, there may be variations between the curriculum designed at the system level, the curriculum communicated by the teacher or in the textbook, and the curriculum as understood by students and their parents.

104. Domain-specific constructs recommended for the FT STQ include students' mathematics class periods per week and use of digital devices for mathematics. Domain-specific constructs recommended for the FT SCQ include administrators' reports of the average time in a class period, the average number of

students in these classes, percentages of students below/above the pass mark, student ability grouping in math, the school offering study help, emphasis on instruction, tracking policies, digital device policies, and selection of courses. This module complements Modules 16 (*Exposure to Mathematics Content*) and Module 17 (*Mathematics Teacher Behaviours*) in mapping out a broad view of students' OTL at school.

105. Figure 19 below illustrates how all proposed constructs in this module map on the taxonomy.

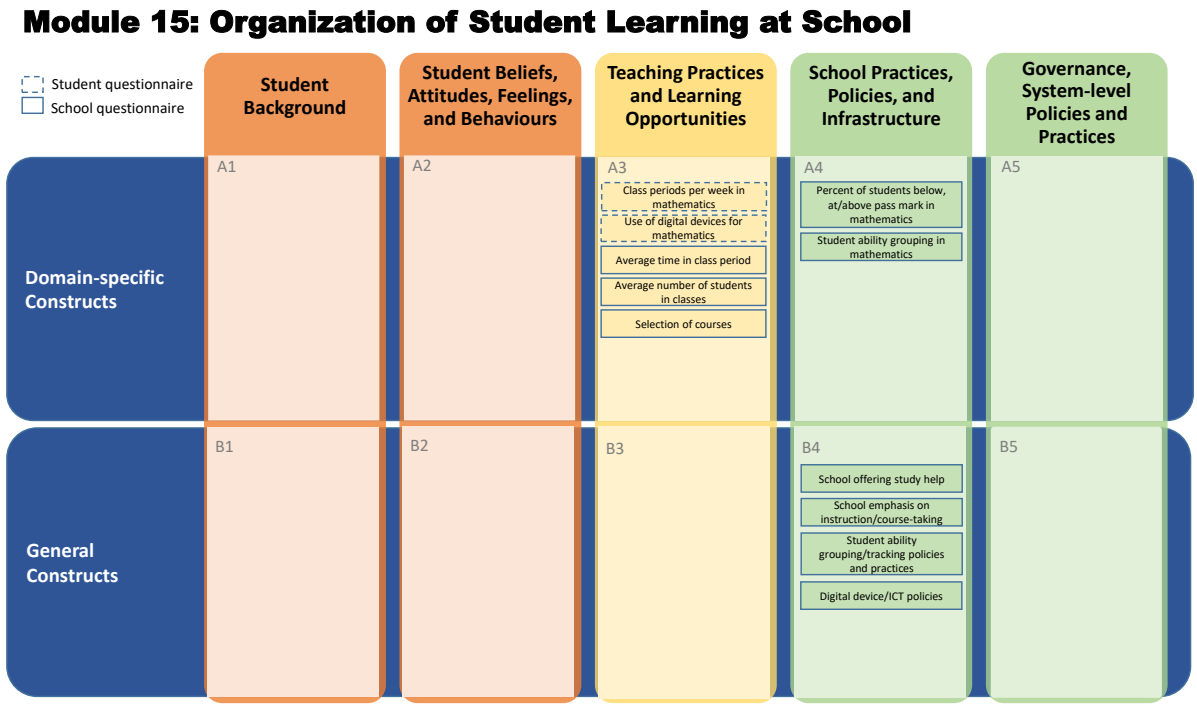


Figure 19. Constructs in Organization of Student Learning at School Module

4.16. Exposure to Mathematics Content

106. This module focuses on one key aspect of the broader OTL constructs, specifically students' exposure to relevant mathematics content. As such it focuses on the first three types of OTL-related variables described by Stevens (1993):

- *Content coverage variables* that measure whether or not students cover the curriculum for a particular grade level or subject matter;
- *Content exposure variables* that consider the time allowed for and devoted to instruction and the depth of teaching provided;
- *Content emphasis variables* that consider which topics within the curriculum are selected for emphasis and which students are selected to receive instruction emphasizing either lower-order skills (i.e. rote memorization) or higher-order skills (i.e. critical problem solving); and
- *Quality of instructional delivery variables* that measure how classroom teaching practices (i.e. presentation of lessons) affect students' academic performance.

107. PISA 2012 aimed to capture domain-specific (mathematics) OTL profiles in the STQ through the presentation of tasks reflecting mathematical abilities and content categories outlined in the PISA mathematics framework. Students were asked to judge whether and how often they had seen similar tasks in their mathematics lessons; thus, OTL measures in PISA 2012 (experience with pure and applied math

tasks, experience with problem types in mathematics, and familiarity with mathematics concepts) were mainly concerned with aspects of content coverage and exposure.

108. One specific area for new development in PISA 2021 is around students' OTL with regard to mathematics reasoning skills. The PISA 2021 FT questionnaires will hone in more precisely on students' exposure to mathematical problems of different complexity, such as with high versus low reasoning skill requirements. Mathematical problems with high reasoning skill requirements are those that have more than one possible solution and the student must provide a justification for the solution they have selected. Mathematical problems with low reasoning skill requirements are those with only one possible solution.

109. The goal of PISA 2021 is to measure in-school OTL (i.e. content coverage and exposure) at the school and country level in a way that allows for a clearer differentiation between types of mathematics problems and mathematics content—for instance, country-level differences in opportunities to learn formal mathematical modelling or applied mathematics problems. Domain-specific constructs recommended for the FT STQ include students' exposure to different types of mathematics content (formal and applied mathematics tasks), exposure to mathematics reasoning and 21st century skills related to mathematics, and exposure to different types of mathematics problems (problems requiring reasoning or *not* requiring reasoning). A domain-specific construct pertaining to the standardisation of the school's mathematics curriculum is also recommended for the FT SCQ.

110. Figure 20 below illustrates how all proposed constructs in this module map on the taxonomy.

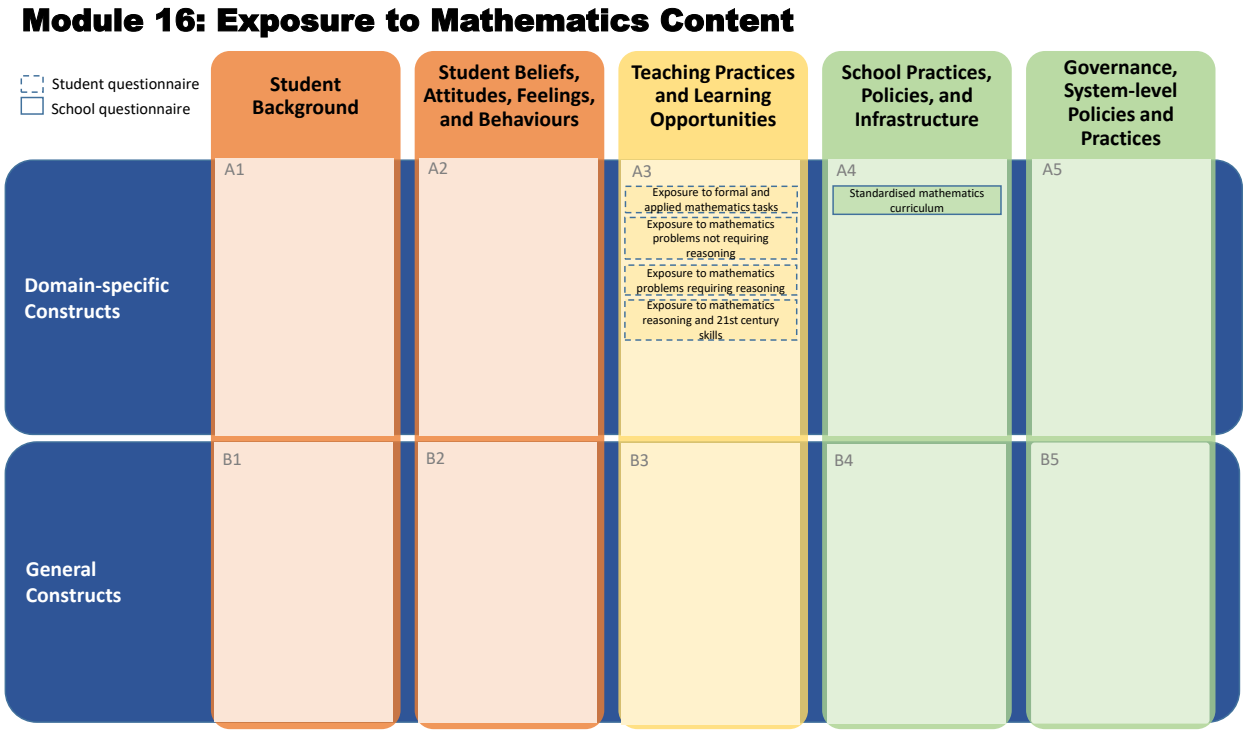


Figure 20. Constructs in Exposure to Mathematics Content Module

4.17. Mathematics Teacher Behaviours

111. How student learning is organized (Module 15) and what content is being taught (Module 16) are conceptually distinct from constructs that capture teaching practices and behaviours (instructional quality), in that teaching practices and behaviours can serve as vehicles through which different levels of content

coverage and exposure may occur. What teachers do has the strongest direct school-based influence on student learning outcomes (Hattie, 2009). Effective instruction is rooted in part in the repertoire of practices through which teachers facilitate students' thinking and understanding of subject content and concepts. Previous research has shown that proximal variables, such as classroom characteristics and teaching and learning practices, are more closely associated with student achievement than distal variables measured at the school- and system-level (e.g. Hattie, 2009; Slavin & Lake, 2008; Wang, Haertel, & Walberg, 1993).

112. Though understood differently across the field, there is general agreement that teachers' instructional practices, or instructional quality, is a multidimensional concept (e.g. Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Kane & Cantrell, 2010). The 2018 OECD Teaching and Learning International Survey (TALIS) framework identifies the following dimensions of teaching practices as having an influence on student achievement:

- *Classroom management,* or the actions taken by teachers to ensure order and effective use of time during lessons (van Tartwijk & Hammerness, 2011);
- *Teacher support,* such as providing extra help when needed, listening to and respecting students' ideas and questions, caring about and encouraging students, and providing emotional support to them (Klieme, Pauli, & Reusser, 2009);
- *Clarity of instruction,* that is, teachers' clear and comprehensive instruction and learning goals, connection of old and new topics, and summarization of lessons (Hospel & Galand, 2016; Kane & Cantrell, 2010; Seidel, Rimmele, & Prenzel, 2005);
- *Cognitive activation,* or the use of instructional activities involving evaluation, integration, and knowledge application in the context of problem solving, through which students engage in knowledge construction and higher order thinking (Lipowsky et al., 2009); and
- *Instructional assessment and feedback*, more specifically, the provision of constructive feedback through formative and summative assessment (Hattie & Timperley, 2007; Kyriakides & Creemers, 2008; Scheerens, 2016) or homework (Cooper, Robinson, & Patall, 2006).

113. Previous TALIS main study results from 2008 found that in 23 countries, participation in professional development and teaching high-ability classes raised the frequency of teachers implementing practices to improve clarity of instruction, teacher support, and cognitive activation (via enhanced activities). It is important to note that while effective pedagogical practices overlap across subjects and student populations, some practices may vary by particular subjects and populations. For instance, TALIS data indicate that mathematics and science teachers reported less student-oriented instructional support and less frequent use of enhanced activities compared to teachers who taught other subjects (OECD, 2009).

114. While TALIS has focused on measurement of general teaching practices, PISA 2021 complements these efforts by measuring closely aligned constructs that are domain-specific (i.e. mathematics focused), as has been done in previous cycles.

- *Disciplinary climate in mathematics* examines disciplinary issues that hinder mathematics learning in the classroom, complementing the TALIS dimension of *classroom management*;
- *Mathematics teacher feedback* is conceptually similar to the dimension of *teacher support,* and is also complemented by the construct of *mathematics teacher support* covered in Module 6;
- *Structure of mathematics instruction* is conceptually similar to the dimension of *clarity of instruction*, however, PISA focuses more specifically on how lessons are structured for learning mathematics, and whether new and old topics are connected, summarization of lessons occurs, and learning goals are communicated;
- *Cognitive activation in mathematics* is conceptually similar to the dimension of *cognitive activation*, however, PISA is focused specifically on the extent to which teachers encourage mathematical thinking and reasoning skills as highlighted in the PISA 2021 mathematics framework; and

- *Teachers' use of assessments* and *mathematics teacher feedback* are conceptually similar to the dimension of *instructional assessment and feedback*. In PISA, school administrator reports of the *use of mathematics assessments* also provide additional information about instructional assessment and feedback in mathematics.

115. Aspects of classroom disciplinary climate, teacher support, cognitive activation, and teacher behaviour (student-oriented) were measured in PISA 2012. Previous research indicates that several of the dimensions defined above correlate with students' mathematics outcomes. For instance, the international PISA 2003 report found that disciplinary climate in the mathematics classroom was strongly associated with mathematical literacy, while other variables (e.g. class size, mathematical activities offered at the school level, avoidance of ability grouping) had no substantial relationship once socioeconomic status was accounted for (OECD, 2004). Additionally, teacher support has been found to be positively linked to students' interest in mathematics after accounting for socioeconomic status (Vieluf, Kaplan, Klieme, & Bayer, 2009). Finally, cognitive activation in the form of providing learners opportunities to develop and practice mathematical competencies have been broadly discussed in mathematics education (e.g. Blum & Leiss, 2007).

116. Addressing teacher and teaching-related factors in PISA is a challenge, because sampling is by age rather than by grade or class. Nevertheless, aggregated student data and the optional teacher questionnaire can be utilized to describe several aspects of teacher background and practices, and the learning environment offered in classrooms.

117. Figure 21 below illustrates how all proposed constructs in this module map on the taxonomy.

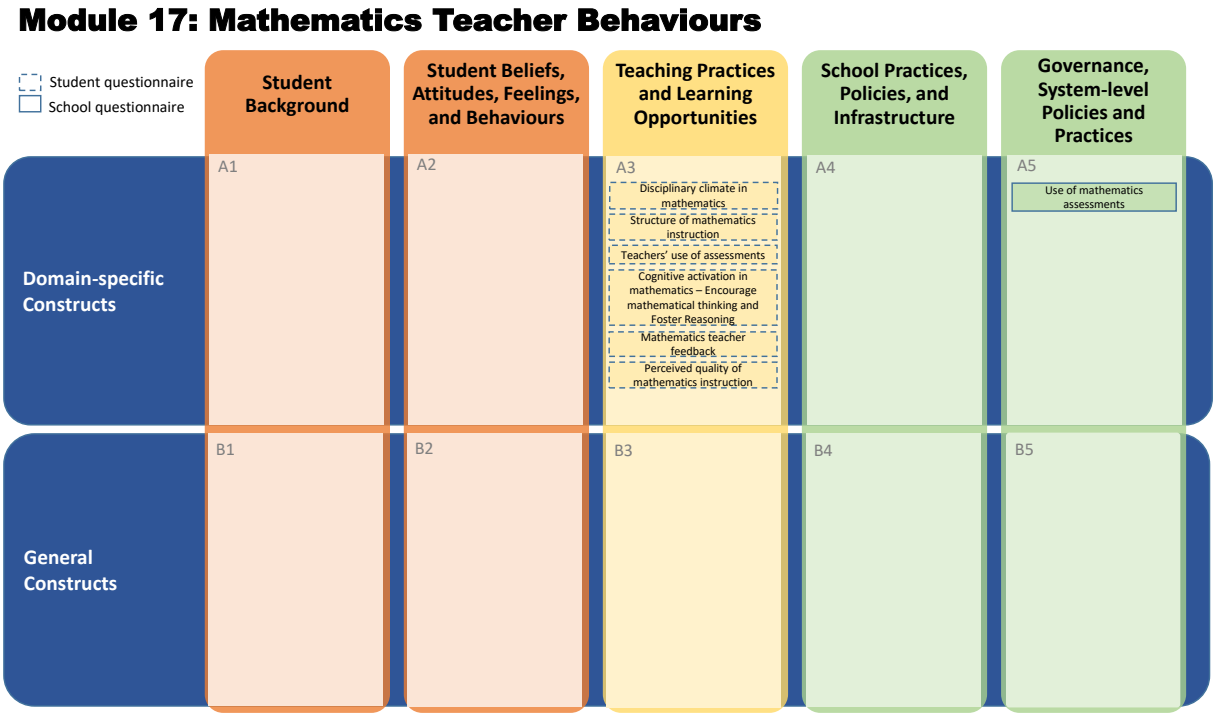


Figure 21. Constructs in Teacher Behaviours Module

4.18. Teacher Qualification, Training, and Professional Development

118. OECD's annual International Summit on the Teaching Profession (ISTP; Schleicher, 2014) has exemplified the continuously growing focus on teacher-related policies for improving the quality of

teachers, teaching, and learning. In addition to teacher's professional behaviour (e.g. interactions with students in the classroom and with their parents or guardians), the composition of the teaching force in terms of age and educational level, their initial education and qualifications, their individual beliefs and competencies, as well as professional practices on the school level (e.g. professional development, interactions with parents) have been topics of educational policy discussions.

119. A number of studies have demonstrated a clear influence of teacher-related factors on student learning and outcomes (e.g. Schmidt, Burroughs, Cogan, & Houang, 2016). Several studies and reviews show positive relationships between teachers' initial education and their teaching effectiveness (e.g. Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005). Research has shown that when teachers have opportunities to expand and develop their teaching practices and their knowledge of instructional approaches, they are more likely to provide a broader range of learning opportunities for students and be more effective in improving students' learning outcomes (Harris, 2002; Rankin-Erickson & Pressley, 2000).

120. General constructs recommended for measurement in the FT SCQ include administrators' reports of teacher qualifications, and in-house professional development opportunities. Recommended domain-specific constructs include mathematics teacher qualifications and mathematics in-house professional development opportunities.

121. Figure 22 below illustrates how all proposed constructs in this module map on the taxonomy.

Figure 22. Constructs in Teacher Qualification, Training, and Professional Development Module

4.19. Assessment, Evaluation and Accountability

122. Assessing students and evaluating schools are common practices in most countries (Ozga, 2012). Since the 1980s policy instruments, such as performance standards, standard-based assessment, annual

reports on student progress, and school inspectorates, have been promoted and implemented across continents. Reporting and sharing data from assessments and evaluations with different stakeholders provides multiple opportunities for monitoring, feedback, and improvement. In recent years, there has been a growing interest in the use of assessment and evaluation results through feedback to students, parents or guardians, teachers, and schools as one of the most powerful tools for quality management and improvement (OECD, 2010, p. 76). In addition, formative assessment, also known as assessment for learning, has been one of the dominant movements (Baird, Hopfenbeck, Newton, Stobart, & Steen-Utheim, 2014; Black, 2015; Hattie, 2009). Accountability systems based on these instruments are increasingly common in OECD countries (Rosenkvist, 2010; Scheerens, 2002, p.36).

123. Prior PISA cycles have covered aspects of assessment, evaluation, and accountability in the SCQ by identifying a variety of purposes for the assessment of students. School administrators have been asked whether they use test results to make comparisons with other schools at the district or national level, as well as to improve teacher instruction (e.g. by asking students for written feedback on lessons, teachers, or resources). However, extant research indicates that there are very few low-income countries that have a national assessment system in place that can track learning in a standardized manner to provide feedback into education policies and programs (Birdsall, Bruns, & Madan, 2016).

124. The evaluation of schools is used as a means of assuring transparency; making judgments and decisions about systems, programs, educational resources and processes; and guiding overall school development (Faubert, 2009), and evaluation criteria may be defined and applied from the viewpoints of different stakeholders (Sanders & Davidson, 2003). Evaluation can either be external (i.e. the process is controlled and headed by an external body and the school does not define the areas that are judged) or internal (i.e. the process is controlled by the school itself and the school defines the areas that are judged) (Berkenmeyer & Müller, 2010). The evaluation may be conducted by members of the school, or by persons/institutions commissioned by the school. Different evaluation practices generally coexist and benefit from each other (Ryan, Chandler, & Samuels, 2007). For instance, external evaluation can expand the scope of internal evaluation and also validate results and implement standard or goals. Additionally, internal evaluation can improve the interpretation and increase the utilization of external evaluation results. However, improvement of schools seems to be more likely when an internal evaluation is applied, compared to external evaluation. Thus, processes and outcomes of evaluation may differ between internal and external evaluation. Moreover, country and school-specific context factors may influence the implementation of evaluations as well as the conclusions and impact for schools. In many countries, individual evaluation of teachers and principals, separate from school-wide evaluation, is also common (Faubert, 2009; Santiago & Benavides, 2009). One study looked at 12 different school management programs in low- and middle-income countries and found that interventions from these management systems did not improve factors such as completion rates and did not have any significant effect on learning outcomes. However, in instances where the program included creating school improvement plans, decentralizing financial-decision making, and generating annual report cards on school performance, there was an improvement in learning outcomes (Snilstveit et al., 2016).

125. In the past several years, a number of countries have implemented national standards to assess students' learning outcomes. Together with formative assessment practices, summative assessment systems influence the way teachers teach and students learn. In particular, formative assessment practices can enhance students' achievement (Black & Wiliam, 1998). However, there is a large variation in the implementation of formative assessment practices, which has also been reported in recent studies in the United States, Canada, Sweden, Scotland, Singapore, and Norway, among others (DeLuca, Klinger, Pyper, & Woods, 2015; Hopfenbeck, Florez Petour, & Tolo, 2015; Jonsson, Lundahl, & Holmgren, 2015; Hayward, 2015; Ratnman-Lim & Tan, 2015; Wylie & Lyon, 2015).

126. Domain-specific constructs recommended for measurement in the FT SCQ include administrators' reports of the use of mathematics achievement data in accountability systems. General constructs recommended for measurement in the FT SCQ include administrators' reports of monitoring teacher practices, feedback to teachers, assessment use in the school overall, the use of social and emotional learning assessments, the use of social and emotional learning data in accountability systems, and school evaluation. As previously noted, the possibility of using OECD system-level data collection to collect complementary information will be explored.

127. Figure 23 below illustrates how all proposed constructs in this module map on the taxonomy.

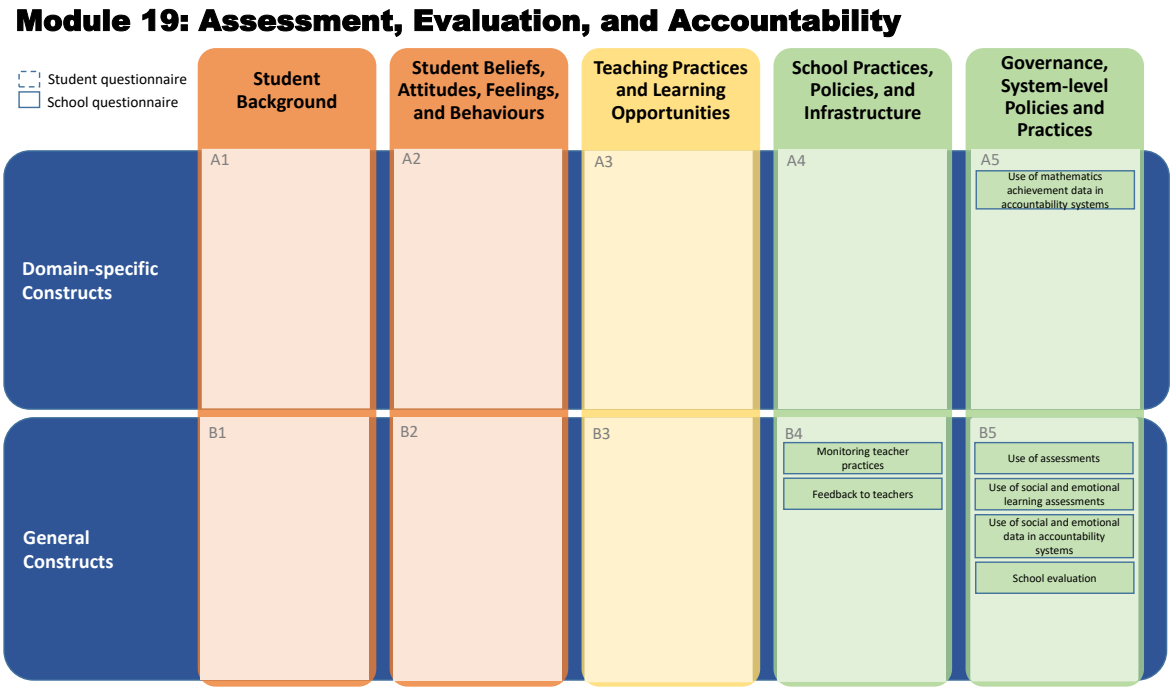


Figure 23. Constructs in Assessment, Evaluation, and Accountability Module

4.20. Parental/Guardian Involvement and Support

128. Parents and guardians are an important audience as well as powerful stakeholders in education, and open communication and collaboration between school leadership and students' parents or guardians are essential to student success. Parental/guardian involvement in education has been conceptualized as parents' or guardians' interactions with schools and their children to encourage academic success (Hill & Tyson, 2009). This involvement is multidimensional and includes school-based involvement (e.g. attending parent-teacher meetings, volunteering at school, or participating in school governance), home-based involvement (e.g. assisting with homework; participating in intellectual enrichment activities not directly related to school but that help develop children's cognitive and metacognitive processes), and academic socialization (i.e. parents' or guardians' educational goals and expectations for their children in general and in specific subjects, and the ways in which these goals and expectations are communicated) (Epstein, 2001; Hill & Tyson, 2009; Kim & Hill, 2015; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, 2016). Parental/guardian involvement may also vary by whether the participation is initiated by parents or guardians, students, teachers, or schools. For example, analyses of PISA 2012 data from seven countries have found that school principals' reports of parent-initiated involvement related positively to between-

school differences in student achievement, while within schools, parent reports of teacher-initiated involvement related negatively to student achievement (Sebastian, Moon, & Cunningham, 2017).

129. In addition to parents' or guardians' involvement in school activities, the support provided in the family plays an important role in fostering student learning and helping children and adolescents develop confidence, stress resistance, and other social and emotional characteristics important for academic and non-academic success. Several meta-analyses show a positive relationship between parental involvement in education and student achievement (Fan & Chen, 2001; Hill & Tyson, 2009; Jeynes, 2007; Kim & Hill, 2015), and parents' academic socialization of their children was found to have a strong positive relationship with achievement (Fan & Chen, 2001; Hill & Tyson, 2009; Kim & Hill, 2015). This correlation generally held across race and ethnicity and when accounting for socioeconomic differences within the United States (Jeynes, 2007; Kim & Hill, 2015).

130. Figure 24 below illustrates how all proposed constructs in this module map on the taxonomy.

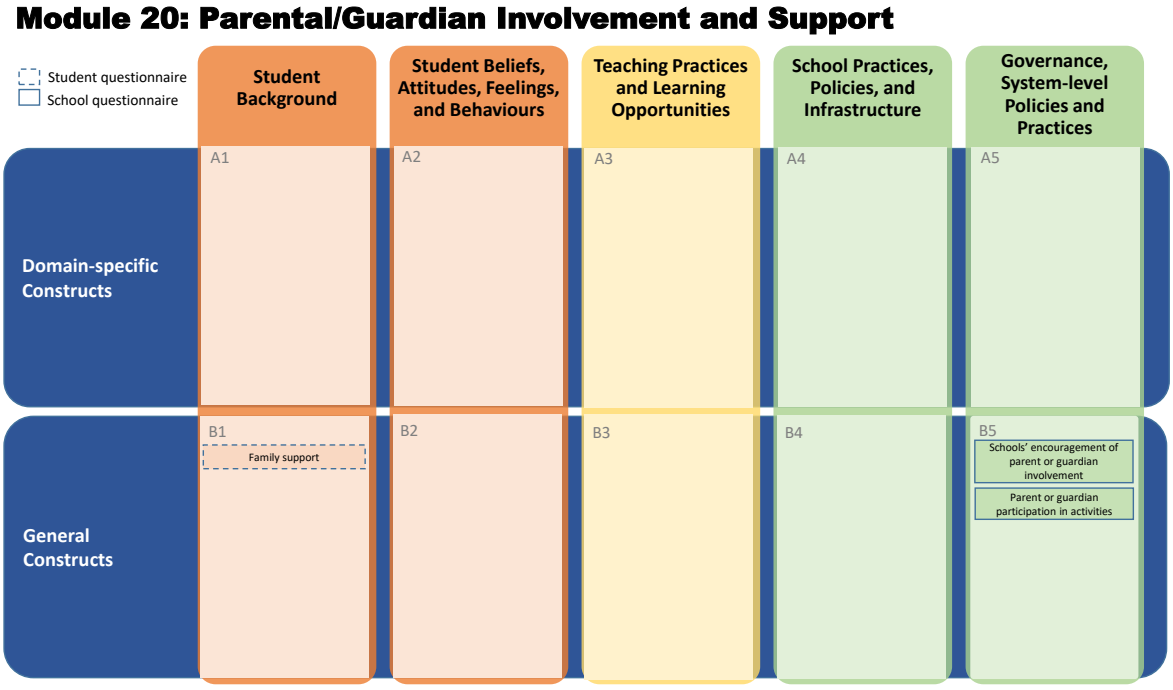


Figure 24. Constructs in Parental/Guardian Involvement and Support Module

5. PISA 2021 Survey Design Principles

131. PISA has made significant contributions to the enhancement and refinement of survey design principles. However, its previous frameworks have not systematically evaluated different methodological approaches or described a comprehensive list of best practices and survey design principles to guide item development. For example, across PISA cycles there have been frequent changes to the number of response options, response option labels, the number of items within scaled indices, or the use of reversed keyed items. Moreover, lack of cross-cultural comparability of questionnaire scales partly due to response styles in PISA is a well-known challenge. While potential strategies for alternative item types (e.g., Kyllonen & Bertling, 2013) as well as statistical approaches (e.g. He, Van de Vijver, Fetvadjiev, et al., 2017) have been explored, these have not always had the expected impact or have not led to noticeable shifts in how PISA data is reported and used. Lastly, measuring the above outlined constructs in PISA 2021 further faces the challenges of implementing robust measurement approaches while keeping student burden low. This framework section presents a clear set of survey design principles to further enhance construct validity of the questionnaire measures in the PISA 2021 FT and strengthen the basis for cross-national and cross-cycle comparisons.

132. Table 6 below gives an overview of the proposed principles, each of which will be described in more detail below.

Table 6. PISA 2021 Survey Design Principles

PISA 2021 Survey Design Principles	
Question types	1. Continue administering rating-scale type items with common response options grouped into matrix questions but harmonize the length of matrix questions to a range of 5-10 items to balance efficiency with controlling cognitive load and respondent fatigue. 2. Continue using alternative item formats introduced during the 2012-2018 cycles (e.g. anchoring vignettes, forced choice, situational judgments tests, slider bars) in cases where there is clear empirical evidence that these methods improve measurement. 3. Minimize the use of fill-in/free response type questions to reduce risk of coding inconsistencies and burden on countries for human coding.
Question Wording	4. For new item development, develop both positively and negatively framed items while avoiding double inversions and evaluate performance in the FT. 5. For new item development, place contextual cues (e.g. "outside of school", "during mathematics lessons"), if applicable, directly in the item rather than the question stem to improve clarity and reduce wordiness and complexity of question stems. 6. For new item development, avoid double- or multi-barrelled questions. 7. For new item development, harmonize the number of examples in question or balance reading load while avoiding potential student misinterpretations of examples as definitions. 8. For new matrix questions developed to capture reflective constructs, ensure sufficient distinctness of items by avoiding including items that are too similar (e.g. items that share a substantial number of words or repeat phrases used in other items).
Response Options	9. Use quantifiable frequency response options where possible and consider possible FT experiments to compare frequency- and agreement-type response options. 10. For new item development, consider increasing the number of response options from 4 to 5 where feasible to allow for more differentiation of responses across students but do not introduce a fifth (middle) category to the established PISA 4-point agreement scale given its longstanding use in PISA, unless there is a specific reason for the particular construct. 11. Display response options in ascending (lowest – highest) order for new questions but retain original order of intact scales retained from previous PISA cycles to facilitate cross-cycle comparisons.
Scaled Indices	12. Continue measuring reflective constructs with multi-item indices scaled based on Item Response Theory (IRT). 13. For new development, administer 6-10 items per construct for the FT to allow for item selection and/or within-construct matrix sampling during the MS; and target approximately five items per construct for every student for construct representation and reliability during the MS.
Routing	14. Use the affordances of the digital delivery platform to use deterministic routing for those questions where collection of more detailed information can be limited to a defined subset of students based on their responses to a previous question that defines a clear routing path.
Matrix Sampling	15. For questions reflective of latent constructs, use a within-construct matrix sampling design whereby individual students answer a subset of items from a larger set of items for each construct. 16. For questions representing manifest or formative constructs, collect data on each question from every student. 17. During the FT implement additionally a construct-level rotation design as used in previous cycles with multiple booklets to allow for data collection on additional constructs as well as implementation of select methodological experiments to guide MS item selection.
Use of log file data	18. Make questionnaire assembly decisions informed by timing data, to the extent that data from previous cycles or other testing programs is available. 19. Utilize log file data to detect response patterns that may impact the quality of collected survey data (e.g. straight lining, rapid responding). 20. Explore use of log file data to enhance survey-based measures of student test-taking motivation (e.g. timing data may be utilized to add to a measure of effort during the PISA test).

5.1. Question Types

5.1.1. Use of Matrix Questions

133. Table 7 below provides an overview of the number of items included in matrix questions across past PISA cycles. On average matrix questions have included between 3 and 6 items, with some exceptions of questions with just two items, as well as a notable number of questions with 7 or more items.

134. For the PISA 2021 FT, we aim to harmonize the number of items in a matrix across questions to optimize the costs and benefits of using matrix questions over discrete single items. Recent research in the context of the *National Assessment of Educational Progress* (NAEP) showed that data quality of matrix questions is comparable to quality of discrete items, with the main difference that matrix questions take much less time to answer (Almonte & Bertling, 2018). The response time benefit plays out especially the longer the matrix is, given that students have to read the stem initially and that time will be added to the first item response. At the same time, data quality suffers if matrices become too long. For example, findings from NAEP show that missing data rates increase if matrices become too long to fit on one screen without scrolling (i.e. higher missing rates are found particularly for those items at the end of a matrix that are not visible without scrolling). While reminders in the digital platform (e.g. prompts alerting respondents when an item on a page has not been answered) may help remedy these effects, it is not clear whether such reminders are equally well understood by test takers across the wide range of the PISA population.

135. For the PISA 2021 FT, we aim to limit the number of items in a matrix to approximately 5-10 items. Upon analysis of FT data, final decisions for the length in the MS should be made.

Table 7. Number of Items in Matrix Questions across PISA Cycles

Number of Subitems	YR2000	YR2003	YR2006	YR2009	YR2012	YR2015	YR2018	Year PISA-D	Total	Average
2 sub-items	0	0	0	0	4	0	0	3	7	0.9
3 sub-items	5	2	3	2	4	6	14	5	41	5.1
4 sub-items	1	1	3	4	8	5	14	2	38	4.8
5 sub-items	2	3	4	5	6	9	10	2	41	5.1
6 sub-items	4	3	6	0	5	4	9	4	35	4.4
7 sub-items	4	0	1	3	0	2	3	1	14	1.8
8 sub-items	2	2	4	0	3	4	2	2	19	2.4
9 sub-items	1	0	0	2	5	2	2	1	13	1.6
10 sub-items	0	2	1	0	1	0	1	3	8	1.0
11 sub-items	1	0	0	1	0	2	0	2	6	0.8
12 sub-items	0	1	1	0	0	0	0	0	2	0.3
13 sub-items	0	0	0	1	1	0	0	0	2	0.3
14 sub-items	0	1	0	0	0	0	0	0	1	0.1
15 sub-items	0	0	0	0	0	0	0	1	1	0.1
16 sub-items	0	1	0	0	1	1	1	0	4	0.5
17 sub-items	0	0	2	1	1	0	0	0	4	0.5
18 sub-items	1	0	0	0	0	0	0	0	1	0.1
24 sub-items	1	0	0	0	0	0	0	0	1	0.1
28 sub-items	1	0	0	0	0	0	0	0	1	0.1
Total	23	16	25	19	39	35	56	26	239	

Note. Green bars denote frequency distributions for individual PISA years, orange bars denote frequencies of total counts across all years, and blue bars denote average frequencies across all PISA years.

5.1.2. Use of Alternative Item Formats

136. Innovative item formats have been explored extensively across the PISA 2012 and 2015 PISA cycles. For instance, PISA 2012 explored the use of anchoring vignettes, situational judgment test items, overclaiming items, and forced choice (Kyllonen & Bertling, 2013). PISA 2015 continued using anchoring vignettes and introduced slider bars to take full advantage of the digital delivery platform.

137. Since the introduction of alternative items formats to PISA in 2012, their use in other LSA context questionnaires has so far found rather limited applications and validity studies have resulted in mixed results (e.g., Bertling & Kyllonen, 2014; Primi, Santos, John, DeFruyt, & Hauck-Filho, 2018; Stankov, Lee, & von Davier, 2017). Anchoring vignettes and situational judgment tests come with the added

complexity that they pose greater demands on respondent time than more traditional rating-scale multiple-choice questions in order to fully exercise the benefits of these techniques. For instance, research with PISA 2012 anchoring vignettes showed that the technique could improve cross-cultural comparability of resulting scales when vignettes were applied to self-report items designed to measure the same construct (Bertling & Kyllonen, 2014), which corresponds to the originally proposed application of the technique (e.g. King & Wand, 2007), but the application of one or few sets of vignettes to multiple distinct scales capturing entirely different constructs may be problematic from a validity perspective (e.g. Stankov et al., 2017; von Davier, Shin, Khorramdel, & Stankov, 2017). Including customized vignettes for every construct in the questionnaire, on the other hand, is not feasible within the time constraints of the PISA STQ administration. The most promising use of vignettes in the context of PISA may not be to recode original student responses but rather consider student responses to vignettes as additional complementary information on students' interpretations of the response options across countries and their use of the entire range of the offered scales (Bertling, 2018).

138. Situational judgment tests are known for their relatively lower internal consistencies (a finding confirmed by PISA 2012 data; Bertling, 2012) calling for longer scales in order to meet reliability standards for LSAs. Forced choice items have a similar problem. While promising psychometric models are available that allow for the derivation of normative scales through ipsative data (e.g. Brown & Maydeu-Olivares, 2013; Stark, Chernyshenko, & Drasgow, 2005), these methods require large numbers of items and pairing of many constructs in order to yield robust results. These conditions are typically not met in LSAs where most constructs are operationalized only through a few items and limited time is available. Mixed results have also been reported regarding test-taker perceptions of forced choice items, with sometimes negative impressions of forced choice items.

139. The most promising technique so far among the innovative item formats explored in PISA 2012 is the use of overclaiming items to adjust subjective topic familiarity ratings for students' tendencies to overclaim what they know and can do. The technique has been widely used in psychological and educational research (e.g. Bensch, Paulhus, Stankov, & Ziegler, 2017; Ziegler, Kemper, & Rammstedt, 2013), and recent applications in the context of the NAEP program in the United States, for instance, confirmed promising findings found in the context of PISA 2012. Another benefit of the overclaiming technique is that it comes at a relatively low cost – only few items need to be added to existing scales. Despite these benefits, an important caveat is that the overclaiming technique lends itself only to a very limited number of constructs (i.e. subjective ratings of familiarity with a topic), which makes it less promising as a technique to address cross-cultural equivalence concerns more broadly across a larger range of constructs (e.g. attitudinal or behavioural constructs).

140. In light of these considerations, it is recommended to keep the number of innovative item formats in the PISA 2021 FT instruments small and limit it to those formats for which gains in validity are expected and/or additional relevant information about students' response behaviours can be collected.

5.1.3. Minimize Use of Open-ended Fill-in-the-blank Questions

141. Open-ended questions that ask the respondent to fill-in a response using constrained or unconstrained free text entry may be problematic for several reasons. In addition to concerns about potentially larger response time burden for the respondent, one of the main challenges in the context of PISA is that analysis of resulting data requires an initial step of coding student responses into quantifiable categories, as well as the necessary quality control steps to ensure coding accuracy. Accuracy of open-ended student responses is a well-known issue with regard to the coding of open-ended responses specifically for parental occupation questions (e.g. Kaplan & Kuger, 2016; Tang et al., 2017). For the PISA 2021 FT it is recommended to minimize the use of fill-in/free response type questions except for

cases where text entry is limited to a small number of digits (e.g. questions about the number of days per week) to reduce risk of coding inconsistencies and burden on countries for human coding.

5.2. Question Wording

5.2.1. Use of Positive and Negative Statements

142. Balancing positively with negatively framed statements in questionnaire items designed to measure bipolar latent constructs is an established tradition in psychological measurement. For bipolar constructs, including both positively and negatively framed statements helps ensure that the entire range of a given construct from both poles of the theoretically defined construct is well represented. For unipolar constructs, which are defined theoretically only with regard to one pole, balancing statements might be less necessary. Balancing statements, however, may be still useful in these cases to minimize the risk of inviting undesired survey responding behaviours, such as "straightlining" (i.e. a response pattern where respondents chose options regardless of their content by creating a straight line across options chosen for several items in a matrix question), and it bears the chance to explore whether additional data cleaning steps may improve the validity and reliability of scales based on such items.

143. On the flipside, researchers have reported that respondents with poor reading proficiency may have difficulty responding accurately to scales that combine both positively and negatively worded items, specifically when negations are used, potentially leading to double-negatives (e.g., "I strongly disagree that mathematics is not one of my favourite subjects."). This problem may be minimized by refraining from using simple negations of positive statements when writing negatively framed statements (but see Cacioppo & Berntson, 1994). Table 13 below illustrates how negatively framed items can be written without the need to include negations.

Table 8. Examples of Positively and Negatively Framed Statements

Positively framed statement (examples)	Reversed keyed with negations (examples)	Negatively framed without negations (examples)
I am full of energy.	I **am not** full of energy.	I tire out quickly.
I finish things I start.	I **don't** finish things I start.	I leave things unfinished.

144. Another alternative approach that has been proposed is to present respondents with questions that intersperse items from scales of more or less socially desirable traits, rather than using reverse-scored items (e.g. Gehlbach & Barge, 2012). Interspersing items from different constructs in one matrix has been implemented in PISA only in a few select cases (e.g. assessment of mathematics anxiety and mathematics self-concept in a combined matrix question in PISA 2012) with the overarching number of items designed to represent a scale being grouped into one single matrix. The idea of interspersing items from different constructs in a common matrix has been recently explored in NAEP with findings pointing to only little differences in the factor structure and reliability of resulting indices. Potential benefits of creating construct heterogenous matrices should be carefully weighed against potential risks, including potentially increased cognitive load due to content variation across items in a matrix.

145. New item development for the PISA 2021 FT will explore these principles of balanced scales, and final decision for the use of balanced versus unbalanced scales and construct homogeneous versus heterogeneous matrices for the MS will be made based on FT data in consultation with the PISA Questionnaire Expert Group (QEG) and Technical Advisory Group (TAG).

5.2.2. Contextual Cue Placement

146. Questionnaire items often ask students to report a behavioural frequency or indicate agreement with a statement when considering a specific contextual cue that may be provided in the question stem or in each individual item (see Table 9 below for an example). While placement of contextual cues in the question stem may seem somewhat more efficient from a reading load perspective, it may be less advisable considering research findings that respondents often place only little attention on reading information in the question stem. Placing an important contextual cue in the question stem bears the risk of students missing this piece of information and, consequentially, providing general rather than specific responses to each item. Recent findings from a large-scale pilot in the context of the United States' NAEP assessment are in line with this assumption (Qureshi, Alegre, & Bertling, 2018).

147. New item development for the PISA 2021 FT will, therefore, place contextual cues preferably in the actual statement rather than the item stem.

Table 9. Examples of Contextual Cue Placement in Question Stem vs. Item

Contextual cue placement in stem only (example)	Contextual cue placement in each item (example)
Thinking about **your mathematics class**, how much do you agree or disagree with each of the following statements? (a) I come to class prepared. (b) I finish my homework right away. (c) I enjoy participating in group activitites.	How much do you agree or disagree with each of the following statements? (a) I come to my **mathematics** class prepared. (b) I finish my **mathematics** homework right away. (c) I enjoy participating in group activities in my **mathematics** class.
Word count: 38	Word count: 40

5.2.3. Avoid Multi-barrelled Statements

148. An established key principle in survey methodology is not to combine multiple ideas or statements into a single item because of the resulting multi-barreledness and statistical confounding of student responses (e.g. Dillman, Smyth, & Christian, 2014; Gehlbach & Artino, 2018). Table 10 below shows examples of double- or multi-barreled items, alongside alternative wording as single statement items. New item development for the PISA 2021 FT will avoid use of multi-barreled items.

Table 10. Examples of Single- vs. Multi-barrelled Statements

Double- or Multi-barrelled statement (examples)	Alternative wording as multiple single statement items.
I am relaxed and handle stress well.	Statement 1: I am relaxed. Statement 2: I handle stress well.
I am helpful and unselfish with others.	Statement 1: I am helpful to others. Statement 2: I am unselfish with others.

5.2.4. Choose a Meaningful Number of Examples

149. A notable number of questions used in previous PISA STQ and SCQ include examples. These examples are necessary to convey what information the respondent is asked to provide and to clarify potential ambiguities of broad terms, such as "classical literature" or "digital devices". Table 11 illustrates that items may differ with regard to the number of examples used and outlines potential validity concerns related to the use of too few or too many examples in an item. In order to maximize the utility of examples in PISA 2021 FT instruments, it is recommended to harmonize the number of examples to a range of 2-5, if feasible. In addition, country-specific examples should be allowed for inclusion, if feasible.

Table 11. Illustration of Questions with Different Numbers of Examples

Number of examples provided	Example item	Potential validity concern(s)
Single example	Which of the following are in your home? Classical literature (**e.g. <Shakespeare>**) *(from PISA 2018)*	• Students may misinterpret the parenthetical as a definition rather than an example if only a single example is provided.
More than 5 examples	What kind of job does your father have? Machine Operator (**e.g., dry-cleaner, worker in clothing or shoe factory, sewing machine operator, paper products machine operator, crane operator, bus driver, truck driver**) *(from PISA-D)*	• The long list of examples increases the reading load of the question, and consequentially the cognitive load, which may affect understanding particularly for respondents with lower proficiency levels. • Some respondents may also misinterpret the long list of terms in the parenthetical as a full list of possible exemplars of a larger category rather than as examples.

5.2.5. Minimize Surface-level Similarities in Wording across Matrix Question Items

150. While most matrix questions used in the PISA STQ and SCQ are designed to measure latent constructs by asking respondents a range of similar, yet related questions, it is important that statements are sufficiently distinct to avoid issues of co-linearity between data collected on each item, which may complicate IRT-scaling and inflated internal consistencies. Moreover, including statements that are too similar in the questionnaire may limit the value of the questions for reporting, unless there is strong reason to keep item wording consistent with previously used items' wording or for comparability with other studies. Table 12 below provides an example of statements deemed potentially too similar alongside an illustration how surface-level similarities between the items in the same matrix may be reduced.

Table 12. Example of Questions with Surface-level Similarities

Scale with potentially too similar items	Rewording of items to reduce surface-level similarities
To what extent do you agree or disagree with the following statements? (a) **I finish** what I start. (b) **I finish** tasks despite difficulties in the way.	To what extent do you agree or disagree with the following statements? (a) **I finish** what I start. (b) **I complete** tasks despite difficulties in the way.

5.3. Response Options

5.3.1. Number of options

151. Across the past seven PISA cycles, the STQ and SCQ have used a broad range of rating scale response option sets, most of which included four response options (see Table 13 for an overview).

152. Based on current knowledge in survey method research, five response options have been proposed as an optimal number for any survey question to collect data of sufficient variability (Revilla, Saris, & Krosnick, 2014) and researchers have cautioned against using response options with too many categories as well as neutral middle categories (Alwin, Baumgartner, & Beatty, 2018). PISA 2021 questionnaires will balance the need to have sufficiently many data points along which student responses can be distinguished with the respondents' inability to distinguish too many response options and the desire to keep response options as simple as possible to facilitate translations and adaptations. For new item development, it is recommended to increase the number of response options from four to five where feasible to allow for

more differentiation of responses across students and more advanced statistical modelling. At the same time, it is *not* recommended to introduce a fifth (middle) category to the established PISA 4-point agreement scale given its longstanding use in PISA, unless there is a specific reason for the particular construct why a middle category would improve validity or cross-cultural comparability.

Table 13. Previously Used Rating-scale Response Options in PISA 2000-2018

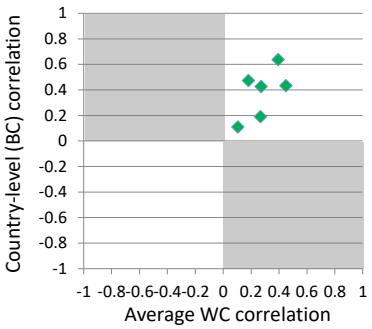
Type of response option	Order	Number of Options	Response Label	2000	2003	2006	2009	2012	2015	2018	Total
Agreement	Decrease	4	Strongly agree / Agree / Disagree / Strongly disagree	0	7	5	0	13	2	2	29
Agreement	Increase	4	Disagree/Disagree somewhat/Agree somewhat/Agree	1	0	0	0	0	0	0	1
Agreement	Increase	4	Strongly disagree/Disagree/Agree/ Strongly agree	3	0	0	3	0	7	16	29
Amount	Decrease	4	Very confident/Confident/Not very confident/Not at all confident	0	1	0	0	1	0	0	2
Amount	Decrease	4	Very important/Important/Of little importance/Not important at all	0	0	1	0	0	0	0	1
Amount	Decrease	4	Very likely/Likely/Slightly likely/Not at all likely	0	0	0	0	1	0	0	1
Frequency	Decrease	4	Always or almost always/Often/Sometimes/Never or rarely	0	0	0	0	2	0	0	2
Frequency	Decrease	4	Frequently / Sometimes / Rarely / Never	0	0	0	0	5	0	0	5
Frequency	Decrease	4	In all lessons/In most lessons/In some lessons/Never or hardly ever	0	0	1	0	0	1	0	2
Frequency	Decrease	4	Very often/Regularly/Sometimes/Never or hardly ever	0	0	1	0	0	1	0	2
Frequency	Increase	4	Almost never/Sometimes/Often/Almost always	1	0	0	1	0	0	0	2
Frequency	Increase	4	Never or almost never/A few times a year/A few times a month/Once a week or more	0	0	0	0	0	2	1	3
Frequency	Increase	4	Never or almost never/About once a week/2 to 3 times a week/Almost every day	0	0	0	0	0	0	0	0
Frequency	Increase	4	Never or almost never/Some lessons/Many lessons/Every lesson or almost every lesson	0	0	0	0	0	3	2	5
Frequency	Increase	4	Never or hardly ever/A few times per year/About once a month/Several times a month	1	0	0	0	0	0	0	1
Frequency	Increase	4	Never or hardly ever/In some lessons/In most lessons/In all lessons	0	0	0	3	0	0	1	4
Frequency	Increase	4	Never or hardly ever/Once or twice a year/About 3 or 4 times a year/More than 4 times a year	1	0	0	0	0	0	0	1
Frequency	Increase	4	Never/One or two times/Three or four times/Five or more times	0	0	0	0	0	1	1	2
Frequency	Increase	4	Never/Rarely/Sometimes/Always	0	0	0	0	0	0	1	1
Frequency	Increase	4	Never/Some lessons/Most lessons/Every lesson	2	0	0	0	0	0	0	2
Frequency	Increase	4	Never/Sometimes/Most of the time/Always	1	0	0	0	0	0	0	1
Amount	Decrease	5	Very much like me/Mostly like me/Somewhat like me/Not much like me/Not at all like me	0	0	0	0	2	0	4	6
Amount	Increase	5	Not at all true of me/Slightly true of me/Moderately true of me/Very true of me/Extremely true of me	0	0	0	0	0	0	1	1
Amount	Increase	5	Not at all true/Slightly true//Very true /Extremely true	0	0	0	0	0	0	2	2
Frequency	Increase	5	Never or almost never/A few times a year/About once a month/Several times a month/Several times a week	0	0	0	1	0	0	1	2
Frequency	Increase	5	Never or hardly ever/A few times a year/About once a month/Several times a month/Several times a week	4	0	0	0	0	0	0	4
Frequency	Increase	5	Never/A few times a year/About once a month/Several times a month/Several times a year	0	0	0	1	0	0	0	1

Note. Green bars denote frequency distributions for individual PISA years; orange bars denote frequencies of total counts across all years.

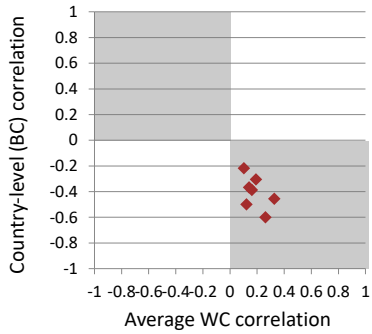
5.3.2. Use of Agreement and Frequency Scales

153. While the overwhelming majority of questions in past PISA cycles have used agreement-type response options (see Table 13), decades of survey methodological research have demonstrated a range of issues with this type of verbal framing of questions, including their proneness to acquiescence response bias and high cognitive burden (e.g. Revila et al., 2014). Bertling & Kyllonen (2014) have shown that scales in the PISA 2012 STQ were especially prone to the so-called "Attitude-Achievement Paradox" (i.e. a phenomenon whereby scales correlate positively with achievement within a group [e.g. country] but correlations flip to the negative when aggregated group-level data [e.g. country-level data] is considered) when positively framed agreement response options were used. In contrast, scales using negatively framed agreement response options or behavioural frequency response options were not affected by the phenomenon (see Figure 25 below). These findings seem to indicate that frequency-based response options may be preferable over agreement-type options.

Scale	Framing	Construct	Correlations with Achievement		
			TOT	BC	WC
Behavioral	+	Min in<class period> - <Math>	.04	.11	.11
Likert	-	Disciplinary Climate	.19	.47	.18
Behavioral	+	Perceived Control of Mathematics Performance	.24	.19	.27
Behavioral	+	Experience with Pure Math Tasks at School	.27	.43	.27
Behavioral	+	Familiarity with Math Concepts	.44	.64	.39
Behavioral	+	Math Self-Efficacy	.43	.43	.45



Scale	Framing	Construct	Correlations with Achievement		
			TOT	BC	WC
Likert	+	Attitude towards School: Learning Outcomes	.08	-.22	.10
Likert	+	Math Work Ethic	.04	-.50	.12
Likert	+	Instrumental Motivation for Mathematics	.07	-.37	.14
Likert	+	Math Interest/ Intrinsic Motivation for Mathematics	.06	-.39	.16
Likert	+	Perseverance	.13	-.30	.19
Likert	+	Openness for Problem Solving	.18	-.60	.26
Likert	+	Math Self-Concept	.26	-.46	.33



(Analyses based on PISA 2012 FT data)

Note. TOT= correlation with achievement for total (pooled) sample across all countries; BC= between-country correlation based on aggregated country-level data; WC= average within-country correlation across all countries

Figure 25. Scales Affected vs. Not Affected by Attitude Achievement Paradox in PISA 2012 (from Bertling & Kyllonen, 2014).

154. It should be noted that response option type and construct were confounded in the aforementioned analyses in PISA 2012, which is why additional research in the specific context of PISA is recommended prior to considering *replacement* of agreement-type questions with frequency-type questions across all constructs. Please note, many of the constructs described in the previous sections, especially the outlined social and emotional characteristics, by definition entail a subjective (and possibly culturally dependent) component and metric or scalar invariance across different cultural groups may therefore be unwarranted. While most of these subjective constructs have traditionally been assessed with agreement type scales, different possible response option sets for PISA 2021 have been explored in cognitive interviews in several countries. Based on the cognitive interview findings, the PISA 2021 Field Trial may implement a few methodological experiments to discern whether cross-cultural comparability can be enhanced by replacing agreement scales focused on the intensity dimension of a construct with frequency scales targeted primarily at the frequency dimension of a construct.

Table 14. Examples of Possible Response Option Sets

Scale type	Response options	Recommendations for PISA 2021 FT
Agreement	Strongly disagree – Disagree – Agree – Strongly agree OR Strongly disagree – Disagree – Neither disagree nor agree – Agree – Strongly Agree	• Retain for re-administration of previously used PISA questions • Limit use in new questions • Explore use in new questions where desirable for comparability purpose with other surveys • Implement methodological comparisons with other options as feasible
Similarity-to-self	Not at all like me – A little bit like me – Somewhat like me – Quite a bit like me – Exactly like me OR Not at all like me – Not much like me – Somewhat like me – Mostly like me – Very much like me	• Limit use based on findings from cognitive interviews and translation challenges.
Abstract Frequency	Never– Rarely – Sometimes – Often – Always	• Limit use in new questions based on findings from cognitive interviews and translation challenges • Implement methodological comparisons with other options as feasible
Absolute Approximate Frequency	Never – About once or twice a year – About once or twice a month – About once or twice a week – Every day or almost every day	• Consider for new questions • Implement methodological comparisons with other options as feasible
Relative Approximate Frequency	Never or almost never – Less than half of the time – About half of the time – More than half of the time – All or almost all the time OR Never or almost never – Less than half of the lessons – About half of the lessons – More than half of the lessons – Every lessons or almost every lesson	• Consider for new questions • Implement methodological comparisons with other options as feasible
Absolute count	Never– Once – Two or three times – Four or five times – More than five times	• Consider for new questions • Implement methodological comparisons with other options as feasible

155. When considering different response options, another important perspective to take into account is the reporting perspective. It is recommended to prioritize response options that allow for more informative and less ambiguous reporting to technical and non-technical audiences. Table 15 below illustrates the range of possible item-level reporting messages based on hypothetical comparison of two countries based on student responses to the statement “My teacher gave me feedback on math assignments”.

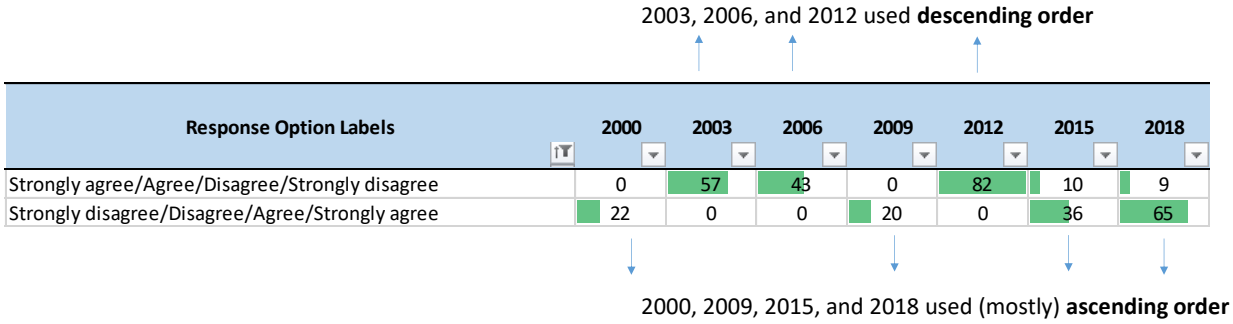
Table 15. Possible Reporting Messages Based on Different Response Option Sets

	Direct Reporting of Relative Frequency Response Category 1	Direct Reporting of Relative Frequency Response Category 2	Direct Reporting of Relative Frequency Response Category 3	Direct Reporting of Relative Frequency Response Category 4	Direct Reporting of Relative Frequency Response Category 5
Relative Frequency	"X% of students in <country A> and Y% in <country B> said their teachers **never** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **less than half of the time.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **about half of the time.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **more than half of the time.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **every time or almost every time / all or almost all of the time.**"
Abstract Frequency	"X% of students in <country A> and Y% in <country B> said their teachers **never** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers **rarely** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers **Sometimes** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers **often** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **always or almost always.**"
Absolute Approximate Frequency	"X% of students in <country A> and Y% in <country B> said their teachers **never** gave them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **about once or twice a year.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **about once or twice a month.**"	"X% of students in <country A> and Y% in <country B> > said their teachers gave them feedback on a math assignment **about once or twice a week.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **every day or almost every day.**"
Absolute Count	"X% of students in <country A> and Y% in <country B> said their teachers **never** gave them feedback on a math assignment this school year."	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **once this school year.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **two or three times this school year.**"	"X% of students in <country A> and Y% in <country B> > said their teachers gave them feedback on a math assignment **four or five times this school year.**"	"X% of students in <country A> and Y% in <country B> said their teachers gave them feedback on a math assignment **more than 5 times this school year.**"
Agreement 5-point	"X% of students in <country A> and Y% in <country B> **strongly disagreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **disagreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **neither disagreed nor agreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **agreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **strongly agreed** with the statement that their teachers give them feedback on a math assignment."
Agreement 4-point	"X% of students in <country A> and Y% in <country B> **strongly disagreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **disagreed** with the statement that their teachers give them feedback on a math assignment."	n/a	"X% of students in <country A> and Y% in <country B> **agreed** with the statement that their teachers give them feedback on a math assignment."	"X% of students in <country A> and Y% in <country B> **strongly agreed** with the statement that their teachers give them feedback on a math assignment."

5.3.3. Harmonizing Directionality of Response Options

156. Figure 26 below shows how the directionality of response options for the most commonly used PISA questionnaire response options changed since the first PISA cycle in 2000. While response options were administered strictly in ascending order in 2000 and 2009, response options were administered strictly

in descending order in 2003, 2006, and 2012. The 2015 and 2018 cycles used a hybrid approach where most questions used ascending order but some questions introduced in earlier cycles were kept in descending order. While harmonizing the directionality of response options in PISA 2021 would likely improve the student experience by making it more consistent across the questionnaire, statistical concerns about backwards comparability of data need to be taken into account. Past FT experiments for PISA 2015 had shown notable effects on item parameters of the direction of response options, and therefore the direction of response options for scales retained from previous PISA cycles will remain unchanged, which may lead to the PISA 2021 FT continuing to include a few questions of opposite directionality to maintain comparability on select constructs.



Note. Green bars denote frequency distributions for individual PISA years.

Figure 26. Variation in Directionality of Response Options from PISA 2000-2018

5.4. Scaled indices

5.4.1. Distinguishing Manifest, Reflective, and Formative Constructs

157. The constructs outlined in this module can be distinguished into constructs that are manifest in nature (i.e. are directly observable and reportable based on respondent answers to a single question) and constructs that are not directly observed and cannot be reported on based on respondent answers to a single question but require the creation of indices for reporting. The latter category can be further differentiated into reflective constructs and formative constructs (for an overview see e.g. Bollen & Lennox, 1991; Edwards & Bagozzi, 2000). Table 16 below lists some examples from the PISA context for manifest, reflective, and formative constructs.

Table 16. Examples of Manifest, Reflective, and Formative constructs in PISA

Manifest constructs in PISA (examples)	Reflective constructs in PISA (examples)	Formative constructs in PISA (examples)
Parental Education	Sense of Belonging	ESCS
Parental Occupation	Mathematics Self-efficacy	Exposure to Mathematics Content
Mathematics Class Periods per Week	Perseverance	Quality of Educational Resources

158. Reflective constructs can be formalized into latent variable models, which often make a unidimensionality assumption of a single statistical cause that determines responses on the items reflective of the construct (MacCallum & Browne, 1993). Social science usually assumes constructs are reflective (Bollen, 2002), and most of the student attitudes, values, and beliefs constructs described in this framework fall into this category: the underlying trait determines how students think, feel, and behave in certain situations.

159. In contrast, formative constructs are theoretically inconsistent with latent variable models. Socioeconomic status is often considered the archetypical example of a formative construct (e.g. Bollen & Lennox, 1991). Another example from PISA that would classify as formative are students' OTL in Mathematics. Unlike the case with reflective constructs, indicators such as parental education or students' exposure to certain type of mathematics problems are not assumed to be caused by ESCS or OTL respectively. Instead, different levels of ESCS or OTL are assumed to emerge when a set of theoretically defined components are combined together. As a result, changes to the item composition necessarily changes the construct. Formative constructs therefore are less suitable for the use of IRT modelling (Howell, Breivik, & Wilcox, 2007).

160. Previous PISA cycles have used IRT to create scaled indices for all constructs classified as reflective and a combination of IRT and other methods (e.g. principal components analysis, or PCA) were used for formative constructs, most notably ESCS. For PISA 2021, different scaling approaches for reflective and formative constructs should be considered during development and after the FT based on guidance from the QEG and TAG. One possible consideration is that measurement equivalence may be relevant for reflective constructs only and less applicable to formative constructs, particularly if multiple group confirmatory factor analysis (MG-CFA) models are utilized to evaluate measurement invariance because the latter method rests on the assumption of one latent trait underlying the scale.

5.4.2. Number of Items per Scaled Index

161. Despite the consistency in scaling indices based on IRT, there is considerable variation with regard to the number of items used in scaled indices across the questionnaires from PISA 2000 – PISA 2018. Some reflective constructs have been targeted with a single item (e.g. Growth mindset) or very few items (e.g. Sense of purpose) whereas 10 or more items were used to scale other constructs (e.g. Familiarity with Mathematical Concepts). Table 17 lists additional examples for short and long questionnaires scales across the last three PISA cycles.

Table 17. Examples of Short and Long Questionnaire Scales in PISA

Examples of short student questionnaire scales (5 items or less)		Examples of long student questionnaire scales (8 or more items)	
Less than 4 items	• PISA 2018: Growth Mindset (1 item) • PISA 2015 and 2018: Life Satisfaction (1 item) • PISA 2018: Attitude Towards School; Sense of Purpose (3 items)	8 items	• PISA 2012: Mathematics Self-Efficacy • PISA 2015: Science Self-Efficacy; Collaboration
4 items	• PISA 2015: Interest and Valuing of Science • PISA 2018: Motivation, Performance Anxiety	9 items	• PISA 2012: Sense of Belonging; Cognitive Activation; Mathematics Work Ethic • PISA 2018: Positive and Negative Affect; Test Language Reading Activities
5 items	• PISA 2012: Perseverance; Enjoyment of Problem Solving • PISA 2015: Test Anxiety • PISA 2018: Enjoyment of Reading; Perspective Taking	10 or more items	• PISA 2012: Familiarity with Mathematical Concepts (13 items); Home Possessions (17 items) • PISA 2015: Science Learning Activities (10 items) • PISA 2018: Engagement in Global Issues (10 items)

162. While including single items or very short scales may be appealing from the administration perspective, it might be problematic from the measurement perspective. In order to provide valid and

reliable measurement of most contextual variables and additional constructs across participating education systems, it is crucial to rely on multiple indicators for the construct at hand.

163. PISA 2021 will continue measuring reflective constructs with multi-item indices scaled based on Item Response Theory. For new development, it is recommended to administer around 6-10 items per construct for the FT to allow for evaluation of the intended scales, item selection, and/or within-construct matrix sampling during the MS. In the MS, any reflective construct should be measured with as sufficient number of items to reach reasonable levels of internal consistency and restrict representation. A desired number of at least five items for every student will be considered as a starting point during FT analyses. The actual number will depend on both statistical criteria and complexity of the construct in terms of content, which may lead to more or less than five items for certain constructs.

5.5. Routing

164. When PISA questionnaires were delivered on paper, the possibilities to customize individual student experiences through routing were extremely limited. The transition to the digital delivery platform in 2015 opened new possibilities for a routing approach, whereby respondents receive different questions based on their responses to previous survey questions. The approach has been used for specific questions, such as to administer follow-up questions, but it has so far not been widely used to increase the efficiency of collecting data for key PISA constructs, such as ESCS. At the same time, there may be opportunities in using routing in the context with ESCS because many of the current questions are relevant only for a subset of student respondents. For instance, if a student indicates that all of their parents or guardians completed at least ISCED 3, questions about the parents' or guardians' ability to read and write would be not appropriate.

165. The challenge of administering ESCS questions relevant to all student respondents becomes more difficult in PISA 2021 given the large increase in the number of participating countries, specifically participation of many lower- and middle-income countries that previously participated in PISA-D. The number of questions needed to cover the entire ESCS range (including PISA and PISA-D questions) may exceed the number of questions that an individual student can answer in time available for measuring ESCS, and not all questions are applicable to all countries and all students within countries. Given the manifest nature of parental occupation, parental education, and certain home possession constructs (e.g. number of books), routing approaches such as the one illustrated in Figure 5 of this document should be explored to yield a greater depth of data with increased efficiency.

166. Beyond the measurement of ESCS, deterministic routing and skip patterns will be explored in the FT for manifest constructs where a clear path can be specified a priori. When introducing routing, an additional important consideration is how to provide countries that will administer questionnaires on paper with an as seamless as possible respondent experience.

5.6. Matrix Sampling

167. Constraints of overall testing time and the large sample sizes in large-scale assessments make matrix-sampling approaches, whereby different respondents receive different sets of items, a viable option to reduce burden while maintaining content coverage across relevant areas. Matrix-sampling approaches are the standard practice for the subject-area tests in educational large-scale assessments (Comber & Keeves, 1973; OECD, 2014) and have more recently been used as an alternative to single-form questionnaire designs.

168. PISA 2012 utilized a three-booklet questionnaire matrix sampling design whereby individual students received one of three possible booklets containing only a subset of all survey questions

administered. This approach, which is illustrated in Figure 27, allowed for testing a total of 41 minutes of questionnaire material in the main survey with each individual student's time limited to 30 minutes, i.e., the design allowed for collection of data on 33 percent more questions than in previous cycles without increasing individual student burden (Adams, Lietz, & Berezner, 2013).

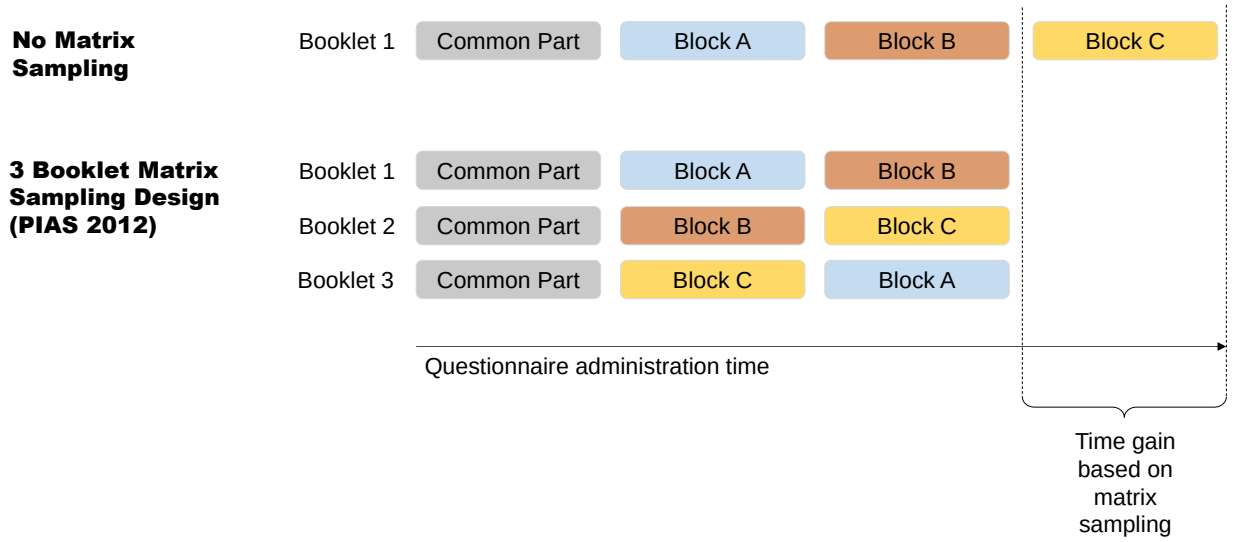


Figure 27. Schematic Illustration of the PISA 2012 3-booklet Matrix Sampling Design

169. A disadvantage of the 2012 three-form design was that entire constructs were rotated rather than rotating individual items within constructs. Thus, one student might answer questions on certain constructs while another student might answer questions on entirely different constructs, but no student answered questions on all constructs. While many researchers reported very small to negligible impact on the overall measurement model, including conditioning and estimation of plausible values (Adams, Lietz, & Berezner, 2013; Almonte, McCullough, Lei, & Bertling, 2014; Kaplan & Wu, 2014; Monseur & Bertling, 2014), methodological concerns about possible attrition in sample size when conducting multivariate regression models and biases in the estimation of plausible values under the construct-level 2012 rotation design have also been raised (von Davier, 2014).

170. PISA 2015 and 2018 reverted back to a single questionnaire form and extended the questionnaire time from 30 to 35 minutes to find a compromise between providing a non-matrix sampled data set and including more variables than feasible to include in a 30-minute booklet.

171. Over the past five years, research has advanced and brought forward new insights about risks and benefits of using matrix sampling for questionnaires, including the exploration of alternative approaches that may prevent the challenges encountered with the 2012 design (e.g. Bertling & Weeks, 2018a, 2018b; Kaplan & Su, 2016).

172. PISA 2021 will utilize an alternative matrix sampling design to the one used in PISA 2012, which would rotate questions within constructs instead of across constructs.

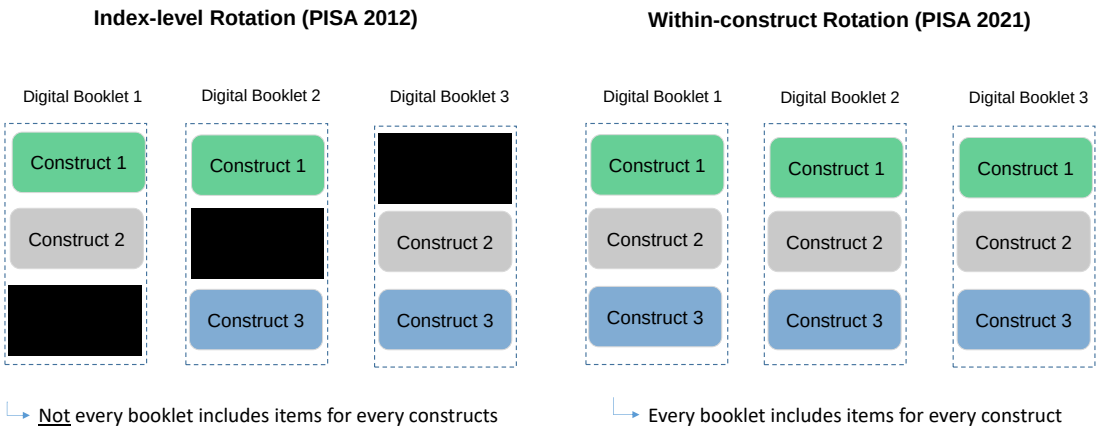


Figure 28. Comparison of Construct-level Missing Data Structure Resulting from Alternative Matrix Sampling Approaches

173. Figure 28 illustrates the differences between index-level and within-construct matrix sampling designs in terms of construct-level missing data. Unlike the PISA 2012 design, in the PISA 2021 FT design, every student will receive questions on all constructs but only answer a subset of all questions for each construct.

174. Bertling and Weeks (2018a, 2018b) presented findings from a series of simulation studies using PISA 2012 data to the PISA TAG and QEG and concluded that there is no statistical reason to rule out within-construct matrix sampling as a potential operational design for the PISA 2021 MS. Differences found in a first study between fixed vs. random selection of anchor items and rotated items were practically negligible, suggesting that both designs would be feasible in PISA (Bertling & Weeks, 2018a). Results from a second study (Bertling & Weeks, 2018b) clearly indicated that within-construct matrix sampling with a random choice of rotated items offers the best results among different matrix sampling approaches. Moreover, findings are in strong support that a design where five items are randomly selected from each item matrix will offer superior data for backwards trend analyses than a single form shortened five item scale or designs with anchor items.

175. Based on discussing a range of possible alternative designs with the PISA TAG, the PISA 2021 FT will utilize a design where a random set of five items per construct (drawn from a set of 8-10 items total for each construct) is administered to each student given that this design led to the most promising findings in the simulation study.

176. In addition to this within-construct matrix sampling design, PISA 2021 will continue utilizing a construct-level rotation approach during the FT only to gather data on a larger number of new and revised questions and allow item reduction and methodological split ballot FT experiments that can guide MS questionnaire assembly. This is illustrated in Table 18.

Table 18. Proposed Matrix Sampling Approaches for FT and MS

	PISA 2021 FT	PISA 2021 MS
Within-construct matrix sampling	Yes	Considered
Construct-level rotation	Yes	Not considered

5.7. Log File Data

177. Since 2015, the PISA assessment has made the transition to computer-based formats. Besides the answers to cognitive and context questionnaire material, the electronic assessment platform captures basic test takers' behavioural data, also known as log-file data (OECD, 2017a). These log-file data can be used for various purposes. For instance, in PISA 2015 and 2018, the answering time was used to guide content selection after the FT.

178. Survey response behaviours captured by log-file data may also be used to relate to cognitive processes (Almond, Deane, Quinlan, Wagner, & Sydorenko, 2012; Couper & Kreuter, 2013; Naumann, 2015; Yan & Tourangeau, 2008). In recent studies, log-file analysis has been used to measure motivation (Hershkovitz & Nachmias, 2009), or to link answering behaviour to aspects of personality (Papamitsiou & Economides, 2017) or students' learning styles (Agudo-Peregrina, Iglesias-Pradas, Conde-González, & Hernández-García, 2014; Efrati, Limongelli, & Sciarrone, 2014).

179. Accordingly, research interest in this area is growing rapidly. While the Programme for the International Assessment of Adult Competencies (PIAAC) study has published an online LogData analyzer tool that allows for easy access to these data for secondary analyses, open access to PISA log data is still missing. The PISA questionnaires in 2021 will once again be assessed via a CBA platform, thus the captured log-file data could be used to explore relationships between answering behaviour and outcomes, in addition to informing content selection post-FT.

180. As fundamental research is missing on the relationship between context indicators as assessed by tests and questionnaires and corresponding data from log-files, making the PISA data accessible for further research seems to be a promising starting point. Although Jude and Kuger (2018) point out that currently "no theoretical frameworks exist specifying which kind of log-file data would be the most promising to contribute additional information in ILSAs," making the data accessible could help researchers explore theories and compare relationships in different countries.

6. References

Abedi, J., Courtney, M., Leon, S., Kao, J., and Azzam, T. (2006), *English language learners and math achievement: A study of opportunity to learn and language accommodation* (CSE Report No. 702, National Center for Research on Evaluation, Standards, and Student Testing. Los Angeles, CA: University of California.

Adams, R. J., Lietz, P., and Berezner, A. (2013), On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-Scale Assessments in Education, 1*, 1-27.

Agudo-Peregrina, Á. F., Iglesias-Pradas, S. Conde-González, M. A., and Hernández-García, A. (2014), Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542-550.

Almlund, M., Duckworth, A. L., Heckman, J. J., and Kautz, T. (2011), Personality psychology and economics. *Handbook of the economics of education, 4*, 1-181.

Almond, R., Deane, P., Quinlan, T., Wagner, M., and Sydorenko, T. (2012), A preliminary analysis of keystroke log data from a timed writing task. *ETS Research Report RR-12-23.* ETS Research Report Series.

Almonte, D. E., McCullough, J., Lei, M., and Bertling, J. P. (2014, April), Spiraling of contextual questionnaires in the NAEP TEL pilot assessment. In Bertling, J. P. (Chair), *Spiraling contextual questionnaires in educational large-scale assessments.* Coordinated Session at NCME Conference, Philadelphia, PA.

Almonte, D. E., & Bertling, J.P. (2018). Effects of item format (discrete vs. matrix) on grade four student responses. In *New insights on survey questionnaire context effects from multiple large-scale assessments.* Symposium at the annual meeting of the National Council on Measurement in Education, New York, NY.

Alwin, D.F., Baumgartner, E. M., and Beatty, B. A. (2018), "Number of Response Categories and Reliability in Attitude Measurement," *Journal of Survey Statistics and Methodology*, 6, 212-239.

Astor, R. A., Benbenisty, R., and Estrada, J. N. (2009). School violence and theoretically atypical schools: The principal's centrality in orchestrating safe schools. American Educational Research Journal, 46, 423–461.

Astor, R. A., Guerra, N., and Van Acker, R. (2010). How can we improve school safety research?. *Educational researcher*, *39*(1), 69-78.

Babcock, P., and Bedard, K. (2011), The wages of failure: New evidence on school retention and long-run outcomes. *Education Finance and Policy, 6*, 293-322.

Baird, J., Hopfenbeck, T. N., Newton, P., Stobart, G., and Steen-Utheim, A. T. (2014), *State of the Field Review: Assessment and Learning*. Report for the Norwegian Knowledge Centre for Education.

Bansak, K., Hainmueller, J., and Hangartner, D. (2016). How economic, humanitarian, and religious concerns shape European attitudes toward asylum seekers. *Science*, aag2147.

Beal, S. J., and Crockett, L. J. (2010), Adolescents' occupational and educational aspirations and expectations: Links to high school activities and adult educational attainment. *Developmental Psychology, 46,* 258-265.

Bensch, D., Paulhus, D. L., Stankov, L., and Ziegler, M. (2017). Teasing apart overclaiming, overconfidence, and socially desirable responding. *Assessment, 1073191117700268*.

Berkenmeyer, N., and Müller, S. (2010), Schulinterne Evaluation - nur ein Instrument zur Selbststeuerung von Schulen? / School-internal evaluation-only an instrument for self-control of schools? In H. Altrichter & K. Maag Merki (Eds.), *Handbuch neue Steuerung im Schulsystem / Handbook New Control in the School System* (pp. 195-218). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bertling, J. P. (2012). Students' approaches to problem solving-comparison of MTMM models for SJT data from PISA 2012. *Technical report presented at the PISA 2012 Questionnaire Expert Group and Problem Solving Expert Group Meeting*, Heidelberg, Germany

Bertling, J. P. (2018). Anchoring Vignettes and Situational Judgment Tests in PISA 2012. *Presentation at OECD Methodological Conference: Cross-country Comparability of Questionnaire Scales in Large-Scale Surveys.* November 9-10, 2018, Paris.

Bertling, J. P., and Kyllonen, P. C. (2014), Improved measurement of noncognitive constructs with anchoring vignettes. Presentation at 79th Annual Meeting of the Psychometric Society, Madison, Wisconsin.

Bertling, J. P., Marksteiner, T., and Kyllonen, P. C. (2016), General noncognitive outcomes. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 255-281). New York: Springer International Publishing.

Bertling, J. P., and Weeks, J. (2018a). *Plans for Within-construct Questionnaire Matrix Sampling in PISA 2021.* Paper presented to PISA Technical Advisory Group (TAG(1808)10). August 2018, Princeton.

Bertling, J. P., and Weeks, J. (2018b). *Within-construct Questionnaire Matrix Sampling: Comparison of Different Approaches for PISA 2021.* Paper presented to PISA Questionnaire Expert Group. October 2018, Oxford.

Binder, M. (2009), Why are some low-income countries better at providing secondary education? *Comparative Education Review, 53,* 513-534.

Birdsall, N., Bruns, B., and Madan, J. (2016), Learning data for better policy: A global agenda. CGD policy paper, 2. Washington, DC: Center for Global Development

Black, P. (2015), Formative assessment: An optimistic but incomplete vision. *Assessment in Education: Principles, Policy and Practice, 22,* 161-177.

Black, P., and Wiliam, D. (1998), Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5,* 7-74.

Blau, D., and Curie, J. (2006), Pre-school, day care, and after-school care: Who's minding the kids? In *Handbook of the economics of education* (pp. 1164-1278). Amsterdam: Elsevier.

Blum, W., and Leiss, D. (2007), Investigating quality mathematics teaching: The DISUM Project. In C. Bergsten & B. Grevholm (Eds.), *Developing and researching quality in mathematics teaching and learning* (pp. 3-16). *Proceedings of MADIF 5*, SMDF, Linköping.

Bowman, N. A. (2010), College diversity experiences and cognitive development: A meta-analysis. *Review of Educational Research, 80*, 4-33.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*, 605-634.

Bollen, K. A., and Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*, 305-314.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009), Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*, 416-440.

Broh, B. A. (2002), Linking extracurricular programming to academic achievement: Who benefits and why? *Sociology of Education, 75*, 69-95.

Brown, A., and Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods, 18*(1), 36.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., and Easton, J. Q. (2010), *Organizing schools for improvement: Lessons from Chicago*. Chicago: University of Chicago Press.

Buchmann, C., and Dalton, B. (2002), Interpersonal influences and educational aspirations in 12 countries: The importance of institutional context. *Sociology of Education, 75*, 99-122.

Butler, J. and R. Adams (2007). The Impact of Differential Investment of Student Effort on the Outcomes of International Studies, *Journal of Applied Measurement 8*, 279-304.

Callahan, R. M. (2005). Tracking and high school English learners: Limiting opportunity to learn. *American Educational Research Journal, 42*, 305-328.

Cacioppo, J. T., and Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401-423. doi:10.1037/0033-2909.115.3.401

Cantril, H. (1965). *The pattern of human concerns.* New Brunswick, NJ: Rutgers University Press.

Carroll, J. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Citro, C. F., and Michael, R. T. (1995). Measuring poverty. A new approach, panel on poverty and family assistance: Concepts, information needs, and measurement methods. *National Research Council. Available online from http://www.nap.edu/catalog/4759.html*.

Cogan, L. S., and Schmidt, W. H. (2015). The concept of opportunity to learn (OTL) in international comparisons of education. In K. Stacey & R. Turner (Eds.), *Assessing Mathematical Literacy* (pp. 207-216). Switzerland: Springer International.

Cogan, L. S., Schmidt, W. H., and Guo, S. (in press). The role that mathematics plays in college- and career-readiness: Evidence from PISA. *Journal of Curriculum Studies*.

Comber, L. C., and Keeves, J. P. (1973), *Science education in nineteen countries: International studies in evaluation.* New York: John Wiley and Sons.

Cooper, H., Robinson, J. C., and Patall, E. A., (2006), Does homework improve academic achievement? A synthesis of research, 1987 – 2003. *Review of Educational Research, 76,* 1-62.

Couper, M. P., and F. Kreuter (2013), Using Paradata to Explore Item Level Response Times in Surveys, *Journal of the Royal Statistical Society, Series A (Statistics in Society), 176,* 271-286.

Cowan, C. D., Hauser, R., Kominski, R., Levin, H., Lucas, S., Morgan, S., and Chapman, C. (2012), *Improving the measurement of socioeconomic status for the National Assessment of Educational Progress: A theoretical foundation.* National Center for Education Statistics. Retrieved from http://files.eric.ed.gov/fulltext/ED542101.

Creemers, B. P. M., and Kyriakides, L. (2008), *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London: Routledge.

Crothers, L. M., Schreiber, J. B., Schmitt, A. J., Bell, G. R., Blasik, J., Comstock, L. A., and Lipinski, J. (2010), A preliminary study of bully and victim behavior in old-for-grade students: Another potential hidden cost of grade retention or delayed school entry. *Journal of Applied School Psychology, 26*(4), 327-338. http://dx.doi.org/10.1080/15377903.2010.518843

Cunha, F., Heckman, J. J., Lochner, L. J., and Masterov, D. V. (2006), Interpreting the evidence on life cycle skill formation. In E. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (pp. 697-812). Amsterdam: Elsevier.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005), Does teacher preparation matter?: Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42). Retrieved December 31, 2018 from https://epaa.asu.edu/ojs/article/download/147/273.

Debeer, D., Buchholz, J., Hartig, J., and Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39(6),* 502-523.

DeLuca, C. Klinger, D., Pyper, J., and Woods, J. (2015), Instructional rounds as a professional learning model for systemic implementation of assessment for learning. *Assessment in Education: Principles, Policy and Practice, 22,* 122-139.

Demakakos, P., Nazroo, J., Breeze, E., and Marmot, M. (2008). Socioeconomic status and health: the role of subjective social status. *Social science & medicine, 67*(2), 330-340.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014), *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Hoboken, NJ: John Wiley.

Duncan, G. J., and Murnane, R. J. (Eds.) (2011), *Whither opportunity? Rising inequality, schools, and children's life chances.* New York: Russell Sage Foundation.

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., and Mac Iver, D. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American psychologist, 48*(2), 90.

Edwards, J. R., and Bagozzi, R. P. (2000). On the nature and direction of the relationship between constructs and measures. *Psychological Methods, 5,* 155-174.

Efrati V., Limongelli C., and Sciarrone, F. (2014), A data mining approach to the analysis of students' learning styles in an e-learning community: A case study. In C. Stephanidis & M. Antona (Eds.), *Universal access in human-computer interaction.* UAHCI 2014. Lecture Notes in Computer Science, (Vol. 8514). Cham: Springer.

Eklöf, H., Pavešič, B. J., and Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education, 27(1),* 31-45.

Epstein, J. L. (2001), *School, family, and community partnerships: Preparing educators, and improving schools.* Boulder, CO: Westview Press.

Fan, X., and Chen, M. (2001), Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review, 13,* 1-22.

Faubert, V. (2009), *School evaluation: Current practices in OECD countries and a literature review.* Paris: OECD Education Working Papers.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., and Büttner, G. (2014), Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29,* 1-9.

Fuligni, A. J., and Stevenson, H. W. (1995), Time use and mathematics achievement among American, Chinese, and Japanese high school students. *Child Development, 66,* 830-842.

Gehlbach, H., and Artino, A. R. (2018), The survey checklist (manifesto). *Academic Medicine: Journal of the Association of American Medical Colleges, 93*(3), 360-366. doi:10.1097/ACM.0000000000002083

Gehlbach, H., and Barge, S. (2012), Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology, 34,* 417-433. doi:10.1080/01973533.2012.711691

Ghuman, S., and Lloyd, C. (2010), Teacher absence as a factor in gender inequalities in access to primary schooling in rural Pakistan. *Comparative Education Review, 54,* 539-554.

Goodman, E., Adler, N. E., Kawachi, I., Frazier, A. L., Huang, B., & Colditz, G. A. (2001). Adolescents' Perceptions of Social Status: Development and Evaluation of a New Indicator. *Pediatrics, 108*(2), e31-e31.

Greene, J. P., and Winters, M. A. (2009), The effects of exemptions to Florida's test-based promotion policy. Who is retained? Who benefits academically? *Economics of Education Review, 28,* 135-142.

Griffith, C. A., Lloyd, J. W., Lane, K. L., and Tankersley, M. (2010), Grade retention of students during grades K-8 predicts reading achievement and progress during secondary schooling. *Reading & Writing Quarterly, 26,* 51-66.

Gurin, P., Dey, E. L., Gurin, G., and Hurtado, S. (2004), The educational value of diversity. In P. Gurin, J. S. Lehman, & E. Lewis (Eds.), *Defending diversity: Affirmative action at the University of Michigan* (pp. 97-188). Ann Arbor: University of Michigan Press.

Gurin, P., Dey, E. L., Hurtado, S., and Gurin, G. (2002), Diversity and higher education: Theory and impact of educational outcomes. *Harvard Educational Review, 72,* 330-366.

Hanushek, E. A., and Wöβmann, L. (2011), The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Wöβmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89-200). Amsterdam: Elsevier.

Harris, A. (2002), *School Improvement: What's In It For Schools?* London: Routledge.

Hattie, J. (2009), *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London & New York: Routledge.

Hattie, J., and Timperley, H. (2007), The power of feedback. *Review of Educational Research, 77,* 81-112.

Hayward, L. (2015), Assessment is learning: The preposition vanishes. *Assessment in Education: Principles, Policy and Practice, 22,* 27-43.

He, J., Van de Vijver, F. J. R., Fetvadjiev, V. H., Dominguez-Espinosa, A., Adams, B. G., Alonso-Arbiol, I., Aydinli-Karakulak, A., Buzea, C., Dimitrova, R., Fortin Morales, A., Hapunda, G., Ma, S., Sargautyte, R., Schachner, R. K., Sim, S., Suryani, A., Zeinoun, P., and Zhang, R. (2017), On enhancing the cross-cultural comparability of Likert-scale personality and value measures: A comparison of common procedures. *European Journal of Personality, 31,* 642-657. doi:10.1002/per.2132

Heckman J. J., Stixrud, J., and Urzua, S. (2006), The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24,* 411-482.

Hershkovitz, A., and Nachmias, R. (2009), Learning about online learning processes and students' motivation through Web usage mining. *Interdisciplinary Journal of E-Learning and Learning Objects, 5.*

Hill, N. E., and Tyson, D. F. (2009), Parental involvement in middle school: A meta-analytic assessment of the strategies that promote achievement. *Developmental Psychology, 45,* 740-763.

Hopfenbeck, T., Florez Petour, M. T., and Tolo, A. (2015), Balancing tensions in educational policy reforms: Large-scale implementation of assessment for learning in Norway. *Assessment in Education: Principles, Policy and Practice, 22,* 44-60.

Hopfenbeck, T. N., and Kjaernsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal, 27(3),* 406-422.

Hospel, V., and Galand, B. (2016), Are both classroom autonomy support and structure equally important for students' engagement? A multilevel analysis. *Learning and Instruction, 41,* 1-10.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12,* 205-218.

Hoy, W. K., Hannum, J., and Tschannen-Moran, M. (1998). Organizational climate and student achievement: A parsimonious and longitudinal view. *Journal of School Leadership, 8,* 336-359.

Jerrim, J. (2015). Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent. *Oxford Review of Education, 41(3),* 310-333.

Jeynes, W. H. (2007), The relationship between parental involvement and urban secondary school student academic achievement: A meta-analysis. *Urban Education, 42,* 82-110.

Jia, Y., Way, N., Ling, G., Yoshikawa, H., Chen, X., Hughes, D., ... and Lu, Z. (2009). The influence of student perceptions of school climate on socioemotional and academic adjustment: A comparison of Chinese and American adolescents. *Child development*, *80*(5), 1514-1530.

Jonsson, A., Lundahl, C., and Holmgren, A. (2015), Evaluating a large-scale implementation of Assessment for Learning in Sweden. *Assessment in Education: Principles, Policy and Practice, 22,* 104-121.

Jude, N. (2016), The assessment of learning contexts in PISA. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 39-51). New York: Springer International Publishing.

Jude, N. and 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). Washington: National Academy of Education. http://naeducation.org/wp-content/uploads/2018/02/2018-Questionnaire-Design-for-ILSA_v02-1.pdf

Kane, T., and Cantrell, S. (2010), *Learning about teaching: Initial findings from the measures of effective teaching project.* Retrieved from: https://docs.gatesfoundation.org/documents/preliminary-findings-research-paper.pdf.

Kaplan, D., and Kuger, S. (2016), The methodology of PISA: Past, present, and future. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning* (pp. 53-73). New York: Springer International Publishing.

Kaplan, D. and Su, D. (2016), On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics, 41,* 57-80.

Kaplan, D., and Wu, D. (2014, April), Imputation issues relevant to context questionnaire rotation. In Bertling, J. P. (Chair), *Spiraling contextual questionnaires in educational large-scale assessments.* Coordinated Session at NCME Conference, Philadelphia, PA.

Kim, S. W., and Hill, N. E. (2015), Including fathers in the picture: A meta-analysis of parental involvement and students' academic achievement. *Journal of Educational Psychology, 107,* 919-934.

King, G., and Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis, 15(1),* 46-66.

Klieme, E., Pauli, C., and Reusser, K. (2009), The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137-160). Münster: Waxmann.

Klieme, E., Backhoff, E., Blum, W., Buckley, J., Hong, Y., Kaplan, D., Levin, H., Scheerens, J., Schmidt, W., Van de Vijver, F. J. R., and Vieluf, S. (2013), PISA 2012 context questionnaires framework. In OECD (Eds.), *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy* (pp. 167-258). Paris: OECD Publishing.

Klieme, E., and Kuger, S. (Eds.) (2014), *PISA 2015 draft questionnaire framework*. Paris: OECD Publishing. Retrieved from: http://www.oecd.org/pisa/pisaproducts/PISA-2015-draft-questionnaire-framework.pdf.

Kloosterman, R., and De Graaf, P. M. (2010), Non-promotion or enrollment in a lower track? The influence of social background on choices in secondary education for three cohorts of Dutch pupils. *Oxford Review of Education, 36,* 363-384.

Kutsyuruba, B., Klinger, D. A., and Hussain, A. (2015), Relationships among school climate, school safety, and student achievement and well-being: A review of the literature. *Review of Education, 3,* 103-135.

Kyllonen, P. C., and Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis,* 277-285.

Kyriakides, L, and Creemers, B. P. (2008), Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: A study testing the validity of the dynamic model. *School Effectiveness and School Improvement, 19,* 183-205.

Lareau, A., and Weininger, E. B. (2003), Cultural capital in educational research: A critical assessment. *Theory and Society, 32,* 567-606.

Lee, J., and Stankov, L. (2018). Non-cognitive predictors of academic achievement: Evidence from TIMSS and PISA. *Learning and Individual Differences, 65,* 50-64.

Lemeshow, A. R., Fisher, L., Goodman, E., Kawachi, I., Berkey, C. S., and Colditz, G. A. (2008). Subjective social status in the school and change in adiposity in female adolescents: findings from a prospective cohort study. Archives of pediatrics & adolescent medicine, 162(1), 23-28.

Levin, K. A., and Currie, C. (2014). Reliability and validity of an adapted version of the Cantril Ladder for use with adolescent samples. *Social Indicators Research, 119(2),* 1047-1063.

Lipowsky, F., Rakoczy, K., Pauli, C., Drollinger-Vetter, B., Klieme, E., and Reusser, K. (2009), Quality of geometry instruction and its short-term impact on students' understanding of the Pythagorean Theorem. *Learning and Instruction, 19,* 527-537.

Little, I., Goe, L., and Bell, C. (2009), *A practical guide to evaluating teacher effectiveness.* Washington, DC: National Comprehensive Centre for Teacher Quality.

Loukas, A. (2007), What is school climate? High-quality school climate is advantageous for all students and may be particularly beneficial for at-risk students, *Leadership Compass, 5,* 1-3.

MacCallum, R. C., and Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin, 114,* 533-541.

Mahoney, J. L., and Cairns, R. B. (1997), Do extracurricular activities protect against early school dropout? *Developmental Psychology, 33,* 241-253.

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., and Arens, A. K. (in press). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology.*

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., and Lichtenfeld, S. (2017). Long-term positive effects of repeating a year in school: Six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores. *Journal of Educational Psychology, 109,* 425-438.

Martin, M.O., Mullis, I. V. S., and Foy, P. (2008), TIMSS 2007 international science report: Findings from IEA's Trends in International Mathematics and Science Study at the eighth and fourth grades. Chestnut Hill, MA: Boston College.

McDonnell, L. M. (1995), Opportunity to learn as a research concept and a policy instrument. *Educational Evaluation and Policy Analysis, 17,* 305-322.

Milem, J. F., Chang, M., and Antonio, A. (2005), *Making diversity work on campus: A research-based perspective.* Washington, DC: Association of American Colleges and Universities.

Minor, E. C., Desimone, L. M., Spencer, K., and Phillips, K. J. R. (2015), A new look at the opportunity-to-learn gap across race and income. *American Journal of Education, 121,* 241-269.

Monseur, C., and Bertling, J. P. (2014, April), Questionnaire rotation in international surveys: Findings from PISA. In Bertling, J. P. (Chair), *Spiraling contextual questionnaires in educational large-scale assessments*. Coordinated Session at NCME Conference, Philadelphia, PA.

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H. W., and Lichtenfeld, S. (2016). Don't aim too high for your kids: Parental over-aspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology, 111,* 166-179.

Naumann, J. (2015), A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior, 53,* 263-277.

OECD. (2004), *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD Publishing.

OECD. (2009), *Creating effective teaching and learning environments: First results from TALIS.* Paris: OECD Publishing.

OECD. (2010), *PISA 2009 results: What students can do? Student performance in reading, mathematics and science, Vol. 1*. Paris: OECD Publishing.

OECD (2011), *Quality time for students: Learning in and out of school, PISA*. Paris: OECD Publishing.

OECD. (2013), *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. Paris: OECD Publishing.

OECD (2014), *PISA 2012 technical report.* Paris: OECD Publishing.

OECD. (2017a), *PISA 2015 technical report: Chapter 17 questionnaire design and computer-based questionnaire platform.* Paris: OECD Publishing.

OECD. (2017b), *Social and emotional skills: Well-being, connectedness and success.* Paris: OECD Publishing.

OECD. (2018), *PISA for Development assessment and analytical framework: Reading, mathematics and science.* Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264305274-en

OECD, European Union, UNESCO Institute for Statistics. (2015), *ISCED 2011 Operational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications*, OECD Publishing. http://dx.doi.org/10.1787/9789264228368-en. Creative Commons Attribution CC BY-NC-ND 3.0 IGO.

Opdenakker, M. C., and Van Damme, J. (2000), Effects of schools, teaching staff and classes on achievement and well-being in secondary education: Similarities and differences between school outcomes. *School Effectiveness and School Improvement*, *11*, 165-196.

Ou, S. R., and Reynolds, A. J. (2010), Grade retention, postsecondary education, and public aid receipt. *Educational Evaluation and Policy Analysis*, *32*, 118-139.

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

Papamitsiou, Z., and Economides, A. A. (2017), Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us? *Computers in Human Behavior*, *75*, 423-438.

Penk, C. (2015). *Effekte von Testteilnahmemotivation auf Testleistung im Kontext von Large-Scale-Assessments.* Berlin: Humboldt-University

Pekrun, R. (2017). Emotion and achievement during adolescence. *Child Development Perspectives. 11*, 215-221.

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., and Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. Child Development, 88, 1653-1670.

Pettigrew, T. F., and Tropp, L. R. (2006), A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology*, *90*, 751-783.

PISA Governing Board (2017), *Report from PISA 2021 background questionnaire strategic advisory group.* Paris: OECD.

Primi, R., Santos, D., John, O.P., De Fruyt, F., and Hauck-Filho, N. (2018), Dealing with Person Differential Item Functioning in Social-Emotional Skill Assessment Using Anchoring Vignettes. In: Wiberg, M., Culpepper, S., Janssen, R., González, J., & Molenaar, D. (Eds) *Quantitative Psychology. IMPS 2017. Springer Proceedings in Mathematics & Statistics, 233*, 275-286.

Purves, A.C. (1987), The evolution of the IEA: A memoir. *Comparative Education Review*, *31*, 10-28.

Quon, E. C., and McGrath, J. J. (2014). Subjective socioeconomic status and adolescent health: a meta-analysis. *Health Psychology*, *33*(5), 433.

Qureshi, F., Alegre, J.M., & Bertling, J.P. (2018). Effect of contextual cue placement on achievement goals item responses, response time, and scalability. In *New insights on survey questionnaire context effects from multiple large-scale assessments.* Symposium at the annual meeting of the National Council on Measurement in Education, New York, NY.

Rankin-Erickson, J. L., & Pressley, M. (2000), A survey of instructional practices of special education teachers nominated as effective teachers of literacy. *Learning Disabilities Research & Practice, 15*, 206-225.

Ratnman-Lim, C. T. L. and Tan, L. H. Kiat (2015), Large-scale implementation of formative assessment practices in an examination oriented culture. *Assessment in Education: Principles, Policy and Practice*, 22, 61-78.

Revilla, M., W. E. Saris, and J. A. Krosnick (2014), Choosing the Number of Categories in Agree/Disagree Scales, *Sociological Methods & Research, 43*, 73-97.

Rosenkvist, M. A. (2010), *Using student test results for accountability and improvement: A literature review. OECD Education Working Papers, No. 54*. Paris: OECD Publishing.

Rumberger, R. W., and Palardy, G. J. (2005), Test scores, dropout rates, and transfer rates as alternative indicators of high school performance. *American Educational Research Journal, 42*, 3-42.

Rutkowski, D., and Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education. 8(3),* 259-278.

Ryan, K. E., Chandler, M., and Samuels, M. (2007), What should school-based evaluation look like? *Studies in Educational Evaluation*, *33*, 197-212.

Rychen, D. S., and Salganik, L. (2003), *Key competencies for a successful life and a well-functioning society*. Goettingen: Hogrefe & Huber.

Sanders, J. R., and Davidson, E. J. (2003), A model for school evaluation. In T. Kellaghan, D. L. Stufflebeam, & L. A. Wingate (Eds.), *International handbook of educational evaluation* (*Kluwer international handbooks of education,* Vol. 9, pp. 806-826). Dordrecht: Kluwer Academic Publishers.

Santiago, P., and Benavides, F. (2009), *Teacher evaluation: A conceptual framework and examples of country practices.* Paris: OECD Publishing. Retrieved from http://www.oecd.org/education/school/44568106.pdf.

Scheerens, J. (2002), School self-evaluation: Origins, definition, approaches, methods and implementation. In D. Nevo (Ed.), *School-based evaluation: An international perspective.* (*Advances in program evaluation,* Vol. 8, pp. 35-69). Amsterdam, the Netherlands: JAI.

Scheerens, J. (2016), *Educational effectiveness and ineffectiveness: A critical review of the knowledge base.* Dordrecht: Springer.

Scheerens, J., and Bosker, R. J. (1997), *The foundations of educational effectiveness.* Oxford: Pergamon Press.

Scherff, L., and Piazza, C. L. (2008), Why now more than ever, we need to talk about opportunity to learn. *Journal of Adolescent and Adult Literacy*, *52*, 343-352.

Schleicher, A. (2014), *Equity, excellence and inclusiveness in education: Policy lessons from around the world. Background report for the 2014 International Summit on the Teaching Profession.* Paris: OECD Publishing.

Schmidt, W. H., and Burroughs, N. A. (2016). The trade-off between excellence and equality: What international assessments tell us. *Georgetown Journal of International Affairs*, *17*, 103-109.

Schmidt, W. H., Burroughs, N. A., Cogan, L. S., & Houang, R. T. (2016), The role of subject-matter content in teacher preparation: An international perspective for mathematics. Journal of *Curriculum Studies, 49*(2), 111–131.

Schmidt, W. H., Burroughs, N. A., Zoido, P., and Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educational Researcher, 44*, 371–386.

Schmidt, W. H., and Maier, A. (2009), Chapter 44: Opportunity to learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 541-559). New York: Routledge.

Schmidt, W. H., McKnight, C. C., Houang, R. T., Wang, H., Wiley, D., Cogan, L. S., and Wolfe, R. G. (2001), *Why schools matter: A cross-national comparison of curriculum and learning*. San Francisco, CA: Jossey-Bass.

Sebastian, J., Moon, J. M., and Cunningham, M. (2017), The relationship of school-based parental involvement with student achievement: A comparison of principal and parent survey reports from PISA 2012. *Educational Studies, 43*, 123-146.

Seidel, T., Rimmele, R., and Prenzel, M. (2005). Clarity and coherence of lesson goals as a scaffold for student learning. *Learning and instruction*, *15*(6), 539-556.

Slavin, R. E., and Lake, C. (2008). Effective programs in elementary mathematics: A best-evidence synthesis. *Review of Educational Research*, *78*(3), 427-515.

Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., and Jimenez, E., 2016. The impact of education programmes on learning and school participation in low- and middle-income countries: a systematic review summary report, 3ie Systematic Review Summary 7. London: International Initiative for Impact Evaluation (3ie).

Stankov, L., Lee, J., and von Davier, M. (2017). A note on construct validity of the anchoring method in PISA 2012. *Journal of Psychoeducational Assessment, 0734282917702270.*

Stark, S., Chernyshenko, O. S., and Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise-preference model. *Applied Psychological Measurement, 29(3),* 184-203.

Stevens, F. (1993), Applying an opportunity-to-learn conceptual framework to the investigation of the effects of teaching practices via secondary analyses of multiple-case-study summary data. *Journal of Negro Education, 62,* 232-248.

Tang, J. H., Whorton, R., Bertling, J., Broer, M., Xie, Q., Rui, N., and Ward, W. (2017, April), *The development and applications of alternative student socioeconomic status measures.* Symposium at AERA Conference, San Antonio, TX.

United Nations. (2015), Transforming our world: The 2030 agenda for sustainable development. Resolution adopted by the seventieth session of the United Nations General Assembly.

van de Vijver, F., and He, J. (2014), Report on social desirability, midpoint and extreme responding in TALIS 2013. OECD Education Working Papers (No. 107).

van Targwijk, J., and Hammerness, K. (2011), The neglected role of classroom management in teacher education. *Teaching Education, 22,* 109-112.

Vieluf, S., Kaplan, D., Klieme, E., and Bayer, S. (2012), *Teaching practices and pedagogical innovations: Evidence from TALIS.* Paris: OECD Publishing.

von Davier, M. (2014), Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis.* Boca Raton, FL: CRC Press.

von Davier, M., Shin, H. J., Khorramdel, L., and Stankov, L. (2017), The effects of vignette scoring on reliability and validity of self-reports. *Applied psychological measurement, 0146621617730389.*

Wang, M. C., Haertel, G. D., and Walberg, H. D. (1993), Toward a knowledge base for school learning. *Review of Educational Research, 63,* 249-294.

Wike, R., Stokes, B., and Simmons, K. (2016). Europeans fear wave of refugees will mean more terrorism, fewer jobs. *Pew Research Center, 11,* 2016.

Willms, J. D. (2006*), Learning divides: Ten policy questions about the performance and equity of schools and schooling systems.* Montreal: UNESCO Institute for Statistics.

Wylie, C., and Lyon, C. (2015), The fidelity of formative assessment implementation: Issues of breadth and quality. *Assessment in Education: Principles, Policy and Practice, 22,* 140-160.

Yan, T., and R. Tourangeau (2008), "Fast Times and Easy Questions: The Effects of Age, Experience, and Question Complexity on Web Survey Response Times," *Applied Cognitive Psycholog*y, 22, 51-68.

Ziegler, M., Kemper, C., and Rammstedt, B. (2013). The vocabulary and overclaiming test (VOC-T). *Journal of Individual Differences.*