Organisation for Economic Co-operation and Development

EDU/PISA/GB(2024)9

For Official Use **English text only**

26 March 2024

DIRECTORATE FOR EDUCATION AND SKILLS
PROGRAMME FOR INTERNATIONAL STUDENT ASSESSMENT

Governing Board

QUALITY STANDARDS FOR PISA

57th meeting of the PISA Governing Board

17-19 April 2024
Valletta, Malta

The PGB is invited to **DECLASSIFY** this document to facilitate its future publication.

Andreas Schleicher, Director for Education and Skills and Special Advisor on Education Policy to OECD's Secretary-General (andreas.schleicher@oecd.org)

JT03540303

Table of contents

Acknowledgements

The project was initiated and steered by Mario Piacentini (OECD) and coordinated by Ava Guez (OECD), with the support of Francesco Avvisati (OECD).

The following authors contributed to this document by authoring the different sections of the PISA Quality Standards:

PISA Quality Standards sections	Authors
Part I: Principles	
Chapter 2: Validity	Stephen Sireci and Javier Suarez-Alvarez (University of Massachussets Amherst)
Chapter 3: Reliability	Stephen Sireci and Javier Suarez-Alvarez (University of Massachussets Amherst)
Chapter 4: Comparability	Stephen Sireci and Javier Suarez-Alvarez (University of Massachussets Amherst)
Chapter 5: Fairness	Stephen Sireci and Javier Suarez-Alvarez (University of Massachussets Amherst)
Part II: Standards	
Chapter 6: Design and coordination	
6.1 Standards for defining new test domains and questionnaire content	Ketan (University of Massachussets Amherst)
6.2 Standards on continuous research and development	Ava Guez (OECD)
6.3 Standards for target population, sampling design and sampling operations	Sabine Meinck (IEA)
6.4 Standards on functionalities of the PISA digital tools	Ketan (University of Massachussets Amherst)
6.5 Standards on the use of alternative forms	Lucia Tramonte (University of New Brunswick)
6.6 Standards on test and questionnaire design	Peter van Rijn (ETS)
Chapter 7: Development of data collection instruments	
7.1 Standards for developing assessment frameworks and specifications for tests and questionnaires	Nina Jude (University of Heidelberg)
7.2 Standards for developing, reviewing, evaluating and revising test items and scoring rubrics	Ava Guez (OECD)
7.3 Standards for developing questionnaire items	Nina Jude (University of Heidelberg)
7.4 Standards for translations and adaptations	Elica Krajceva (Capstan)
Chapter 8: Scoring and analysis	
8.1 Standards on analysis of field trial data	Peter van Rijn (ETS)
8.2 Standards on computing survey weights and preparing datasets for estimating sampling variance	Sabine Meinck (IEA)
8.3 Standards on constructing and validating scales from the cognitive tests	Peter van Rijn (ETS)

8.4 Standards on constructing and validating scales from the questionnaire	Jonas Bertling (ETS)
8.5 Standards on constructing proficiency-level descriptors	Peter van Rijn (ETS)
Chapter 9: Reporting of PISA data	
9.1 Standards on producing and reviewing initial reports of findings	Lucia Tramonte (University of New Brunswick)
9.2 Standards on releasing items and data products	Ava Guez (OECD)
9.3 Standards on technical reports	Ava Guez (OECD)

We are grateful for the comments received on drafts of the different sections from the following reviewers, including international experts and national experts from PISA participating countries and economies: Samantha Burg, Enis Dogan, Belinda Gaskin, Emma Holmberg, Fredrik Jensen, Yu Kameoka, Steve May, Claudia Matus, Emma Medina, Hjalte Meilvang, Christian Monseur, Leslie Rutkowski, and Megan Welsh. Within the OECD Secretariat, we thank Janine Buchholz, Tiago Caliço, Catalina Covacevich, Tiago Fragoso, Tue Halgreen, Miyako Ikeda and Claudia Tamassia, for their valuable comments and input on the drafts, and Josephine Murasiranwa who provided administrative support. We are also grateful for the support and advice from Andreas Schleicher, Director for Education and Skills, and Yuri Belfali, Head of the Early Childhood and Skills Division. Megan Welsh (University of California Davis) supported the initial work on conceptualisation and design. We thank members of the PISA Governing Board and experts from the PISA Technical Advisory Group (TAG) and of the Research, Development and Innovation Group (RIG) for their advice and guidance from the early stage of the project.

Quality Standards for PISA

Chapter 1. Introduction to the PISA Quality Standards

1. Over the last two decades, the Organisation for Economic Co-operation and Development (OECD)'s Programme for International Student Assessment (PISA) has become a global yardstick for evaluating and comparing quality, equity, and efficiency in learning outcomes in education systems around the worlds. It has also played a significant role in driving education reforms, allowing policy makers to make more informed decisions. PISA evaluates every three (now four) years what students know and can do through a set of standardised assessments in science, mathematics, reading, and an innovative domain. It also collects data from students, teachers, schools and education systems on attitudes, well-being, and learning environments. PISA is a collaborative enterprise between the participating countries and economies, the experts and organisations that are part of the PISA Consortium, and the OECD Secretariat.

2. PISA has been at the forefront of educational measurement and has consistently aimed for the highest methodological standards, from assessment design to the reporting of data. However, there has not been, until now, any effort to explicitly articulate these standards for the whole assessment cycle. Indeed, the current PISA technical standards cover the procedures that ensure the consistent implementation of PISA in participating countries and economies, but do not address other phases of the project cycle (e.g. international coordination, instrument development, analysis and reporting) or aspects of assessment quality like relevance, validity, fairness, and comparability.

3. This document was produced to fill this gap. It was conducted under the PISA Research, Development and Innovation (RDI) programme, following a decision of the PISA Governing Board (PGB). These comprehensive, PISA-specific standards are intended to take into account the unique characteristics of PISA, such as the need to balance regular updates with the preservation of trend measures. They aim to make more visible to the public what is inside the complex PISA machine, and to illustrate how the quality and value of PISA builds on the coordinated actions of different actors. They aim to consolidate good practices within PISA, providing a comprehensive reference on processes and methodologies that have proven to be effective in previous cycles, and on quality targets that PISA should strive to achieve through continuous improvements in these processes and methodologies. The standards also inform the appropriate use of PISA data in analytical work conducted by the OECD or outside the OECD.

4. This introductory chapter provides a short description of PISA and explains the objectives, scope and expected use of the Quality Standards.

1.1. Characteristics of PISA

1.1.1. The objectives of PISA

PISA is a triennial survey of 15-year-old students around the world that assesses the extent to which they have acquired key knowledge and skills essential for full participation in social and economic life. PISA assessments do not just ascertain whether students near the end of their compulsory education can reproduce what they have learned; they also examine how well students can extrapolate from what they have learned and apply their knowledge in unfamiliar settings, both in and outside of school. (OECD, 2023)

5. The OECD conducted PISA for the first time in 2000 as a response to the need for cross-national comparisons on students' performance.

6. The PISA long-term strategy describes PISA's mission as follow:

> *PISA supports participating Members and Partners in achieving high quality lifelong learning by improving the quality of learning outcomes, increasing equity in learning opportunities, and enhancing the effectiveness and efficiency of education systems. (OECD, 2023)*

7. The PISA long-term strategy further defines principles that guide PISA:

i. *PISA is policy-oriented and meets the needs of educational policy making for enhancing teaching and learning.*

ii. *PISA is a regularly administered system-level assessment of learning outcomes, well-being and learning environments of students.*

iii. *PISA assesses knowledge, skills, attitudes and values of students and their capacity to apply these competences in real life for lifelong learning, employment and citizenship.*

iv. *PISA provides valid, comparable and reliable data respecting the principles of fairness and scientific integrity.*

v. *PISA enables international comparison across a wide range of countries and over time.*

vi. *PISA is a collaborative and innovative effort to develop forward-looking assessments to inform emerging policy* issue.

8. In line with this mission and principles, the key uses of PISA have been (a) comparing country results on a set of indicators for identifying high-performing educational systems in terms of educational quality and equity; (b) monitoring trends in country performance over time; and (c) generating insights about effective policy and practices that are associated with better outcomes. While these key uses remain at the core of what PISA does, the objectives of PISA evolve across cycles as the study reaches more countries, extends its coverage of learning domains and becomes of interest to other stakeholders in addition to education policymakers. These objectives are defined by the PISA Governing Board for periods of ten years in the PISA long-term strategy.

1.1.2. The stakeholders and actors of PISA

9. The quality of PISA depends on the collaborative, coordinated work of different actors:

- The *PISA Governing Board (PGB)*, composed of delegates of each participating countries and economies, determines the policy priorities for PISA and oversees its implementation.
- The *PISA National Centres* are responsible of implementing national adaptations of test and questionnaire items, administering the data collection and reviewing the appropriateness of the test material for their student population. National Project Managers ensure that the implementation of the survey is of high quality, and verify and evaluate the survey results, analyses, reports and publications.
- The *Expert Groups* are constituted of subject-matter experts and provide expert guidance on the development of the cognitive assessments, including framework and item development. National experts from participating countries

and economies also participate in working groups to ensure that the instruments are *internationally* valid and take into account their various cultural and educational contexts, and that the assessment materials enable the production of valid information.

- The *Technical Advisory Group (TAG)* is composed of educational assessment *experts* and advises on technical issues to ensure the quality of PISA data.
- The *PISA Consortium* includes the organisations responsible for developing the frameworks and instruments for tests and questionnaires and the sampling framework, supporting national teams with translations, preparation and administration of the data collection, processing and scaling the data, typically selected *as* contractors to the OECD through a competitive call for tender.
- The *sampled students and schools* are involved in administering and taking the assessment.
- The *OECD Secretariat* coordinates the overall programme, monitors its implementation, facilitates decision making on new content and cross-cycles research, selects and monitors the Contractors. It also produces indicators and *analysis* and drafts the international reports and publication on the results, in cooperation with the PISA Consortium and in close consultation with the PGB and National Centres.

1.2. Description of the PISA Quality Standards

1.2.1. The purpose of the PISA standards

10. The Quality Standards are primarily intended to guide all the actors involved in the development and use of PISA data. They make explicit the quality principles that should undergird all aspects of the development and use of the assessment. They reference and indicate best practices and quality criteria for design choices, instrument development, analysis and reporting. They outline the methodological and content innovation goals for PISA, related for example to the use of technology in item design, scoring and validation, or to making the test accessible to a larger population of students. As such, the standards provide guidelines, constitute a yardstick against which PISA can be assessed and indicate directions to guide future innovation in PISA. They are not intended to be prescriptive in order to balance the need for ensuring the quality of the assessment with the need to innovate and evolve as technologies and methodologies improve. In this respect, it is important to note that the Quality Standards are a living document that will be regularly updated as research on large-scale assessments advances, the content of PISA test and questionnaires gets refined, and the technology used in PISA changes.

11. The Standards can serve as a valuable reference for countries that are building their own assessment system.

12. A final goal of the Standards is to inform the public on the process and coordinated actions underlying the development of a complex international large-scale assessment like PISA.

1.2.2. Relation with the PISA Technical standards

13. The PISA Quality Standards differ from the PISA Technical Standards in its coverage, objectives and prescriptiveness. The PISA Technical Standards focus on the procedures that ensure the consistent implementation of PISA in different countries. They serve as a set of criteria for post hoc data adjudication (decisions on whether the data

for a specific country are of sufficient quality for inclusion in the international reports), but do not address other phases of the project cycle (e.g. international coordination, instrument development, analysis and reporting) or aspects of assessment quality like validity or fairness. The PISA Quality Standards complement the Technical Standards: their scope extends beyond defining obligations for countries to follow during data collection for data adjudication purposes by supporting the development of high-quality instruments and the valid interpretation and use of PISA results.

14. The current Technical Standards continue to exist as a separate document that covers the procedures that countries have to respect in order to meet the requirements of the data adjudication process. Hence, the Quality Standards described in this document do not cover procedures related to the organisation and implementation of the data collection by national teams.

1.2.3. Development process of the PISA Quality standards

15. During the first year of the project, the OECD Secretariat has sought inputs from several experts in the PISA Technical Advisory Group (TAG) and the Research, Development and Innovation Group (RIG), and from experts in national teams, on the scope and expected use of the new standards, on the organisation of the standards document and on the formulation of individual standards. The OECD Secretariat worked with external experts to produce the various chapters and specific standards. First drafts of the different sections were shared with the PISA Governing Board throughout the development process in order to receive input and feedback from national experts in participating countries and economies and reflect their views.

1.2.4. Coverage and use of the PISA Quality standards

16. The Standards are structured in two main parts to promote usefulness and accessibility to different audiences.

17. *Part I – Principles* outlines the general principles that guide decisions on what to assess, instrument design, data collection and reporting. These core principles are validity, reliability, comparability, and fairness. This quality target echoes the overall OECD quality framework for data quality ([STD/QFS(2011)1]). This section is organised in four chapters that define what validity, reliability, comparability and fairness mean in the context of PISA. All the chapters include references to the standards in Part II of the document, so that readers can easily understand how standards respond to more general quality concerns and guidelines. Policy makers and the general public can use this section to seek guidance on how the data should be interpreted and used, and to get a better sense of the scientific rigor and values that drive the development of the measures and their interpretation.

18. *Part II – Quality Standards* addresses the need for clear guidance for PISA stakeholders on what they are expected to accomplish in specific operations This section presents the specific Standards, following the life-cycle of PISA, from the definition of reporting goals to the publication of the reports, indicating expected actions from each actor and providing objective quality-assurance criteria based on evidence about the process and about the product. These operational guidelines help ensure PISA's technical quality. Standards have been developed in the following areas:

Design and coordination
Defining new test domains and questionnaire content
Continuous research and development
Target population, sampling design and sampling operations
Functionalities of the PISA digital tools
Use of alternative forms
Test and questionnaire design
Development of data collection instruments
Developing assessment frameworks and specifications for tests and questionnaires
Developing, reviewing, evaluating and revising test items and scoring rubrics
Developing questionnaire items
Translations and adaptations
Scoring and analysis
Analysis of field trial data
Computing survey weights and preparing datasets for estimating sampling variance
Constructing and validating scales from the cognitive tests
Constructing and validating scales from the questionnaire
Constructing proficiency-level descriptors
Reporting of PISA data
Producing and reviewing initial reports of findings
Releasing items and data products
Technical reports

19. Each section of the standards presents a rationale for establishing standards in the given activity, and specifies a set of standards for that activity. Importantly, the standards distinguish requirements or what has been regular practice in PISA, from recommendations or aspirational practices. The distinction is made with the following keywords:

- *"shall"* is used to denote a requirement which is important to meet to ensure the quality of PISA data and its uses; and
- *"should"* is used to denote a recommendation or guideline which is expected to enhance the overall quality of PISA (in terms of validity, reliability, comparability and/or fairness) but is not required, for instance because it may imply additional logistical constraints for ongoing or upcoming cycles.

20. Finally, each section enumerates the related quality assurance measures and criteria for the standards. The quality assurance section refers to both evidence on the *processes* that were followed for the activity, and to evidence on the *product* whenever the latter is available.

Part I. Principles

21. This section of the Standards defines the quality target of PISA, which builds on the four following core principle: validity, reliability, comparability and fairness. The section is organised in four chapters that define what validity, reliability, comparability and fairness mean in the context of PISA. All the chapters include references to the standards in Part II of the document, so that readers can easily understand how specific standards respond to more general quality concerns.

Chapter 2. Chapter 2: Validity

2.1. Introduction and definition of validity

22. Validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA and NCME, 2014, p. 11[1]). This definition emphasises that validity is not an inherent property of a test, but rather a concept that refers to the defensibility of the use of a test in practice, and the soundness of test score interpretations. It also implies that the reasoning and the empirical evidence that support the proposed interpretation of PISA results, and which constitute the "validity argument," (AERA, APA and NCME, 2014[1]; Kane, 2013[2]) should be made accessible. Thus, concerns for the intended (and final) use of the PISA results should permeate all aspects of the design and development of PISA, from initial conceptualisation of the domain of knowledge and skills to be tested, to test development and administration, through the analysis and reporting of results. Assumptions made in the context of test development or data analysis should be explicitly stated, grounded in theory and prior evidence, and, where feasible, put to the test using statistically sound procedures as part of a validation program.

23. All aspects of a testing program that relate to its technical quality and utility are relevant to validity. Thus, validity is an overarching concept that essentially serves as a "North Star" for an assessment program. Some aspects of validity, such as *Reliability* and *Comparability*, have specific importance and meaning and are therefore discussed more fully in separate chapters (see Chapters 2 and 3). Essentially, *Reliability* refers to the consistency of an assessment process in producing results, which is sometimes referred to as measurement *precision*. *Comparability* refers to the degree to which test results can be meaningfully compared across various testing contexts, which is particularly important for PISA, given the multinational and diverse contexts in which it is administered. A concept that overlaps largely with validity is *Fairness* (Chapter 4). However, although *Fairness* requires valid interpretation and use of test scores, it also involves consideration of the welfare and treatment of students and other stakeholders during and after the testing process, and so it also merits a separate chapter to fully discuss these considerations.

24. Noting that reliability, comparability, and fairness shall all hold for an interpretation to be valid, this chapter focuses on how validity can promote assessments of high technical quality that support valid interpretations and help testing programs realise their intended purposes. The discussion of validity is written with the expectation that the *Reliability*, *Comparability,* and *Fairness* chapters will also be consulted so readers will get a comprehensive understanding of how these critical components of a testing program overlap but are also uniquely important.

25. Given that validity refers to evidence and theory that supports the use of a test for its intended purposes, validation research for PISA shall be conducted in relation to intended interpretations and uses of PISA scores. The *purposes* for PISA assessments are described at the beginning of the main reports produced by the OECD on behalf of the PISA Governing Board: for example, the report *PISA 2018 Results (Volume I): What Students Know and Can Do* (OECD, 2019[3]) provides the following description of what PISA measures:

> *PISA is a triennial survey of 15-year-old students around the world that assesses the extent to which they have acquired key knowledge and skills essential for full participation in social and economic life. PISA assessments do not just ascertain whether students near the end of their compulsory education can reproduce what they have learned; they also examine how well students can extrapolate from what*

they have learned and apply their knowledge in unfamiliar settings, both in and outside of school.

26. The PISA long-term strategy [EDU/PISA/GB(2013)14] further describes the *intended uses* of PISA:

PISA responds to the need for cross-national comparisons on student performance. It aims to provide reliable information on how well education systems prepare students for further study, careers and life. PISA also provides a basis for international collaboration in order to define and implement effective educational policies.

27. Given these purposes and intended uses, validity evidence for PISA should include not only evidence confirming that the test measures the intended construct (note: historically, the knowledge and skill domain to be measured is referred to as a "construct" (Cronbach and Meehl, 1955[4]; Sireci, 2020[5])), as every assessment should, but also that the construct measured is consistent across countries administering the assessment (i.e. measurement equivalent; see Chapter 3 on *Comparability*), and that the assessment captures this construct with the sufficient precision to facilitate comparisons across countries and subgroups within countries. Unlike a traditional test that focuses on individual scores, validity evidence shall be provided to confirm PISA results are valid for comparing national and subgroup estimates of students' performance across jurisdictions and over time. It is also important to bear in mind that PISA assessments include measurement of both cognitive and non-cognitive constructs (e.g. socio-economic status, attitudes). Gathering validity evidence is crucial for drawing inferences on these multiple constructs across multiple countries, cultures, and other subgroups.

2.2. Threats to validity

28. An important aspect of validating the use and interpretation of educational test scores is ensuring that these tests measure the knowledge and skills they intend to measure (Messick, 1989[6]; Kane, 2013[2]). Messick (1989) described the major threats to valid assessment as "construct under-representation" and "construct-irrelevant variance."

29. The first threat occurs when an assessment does not fully measure the entirety of the intended domain. For example, the PISA 2018 reading test intends to measure three processes of reading (locating information, understanding, and evaluating and reflecting). Construct under-representation in PISA could happen if the test only contains items measuring locating information and understanding, but not evaluating and reflecting. Construct under-representation could occur in PISA test score interpretation if results were interpreted based on only the subset of items administered to a single student instead of all sampled students. A more subtle manifestation of construct under-representation might occur if items do not sufficiently differentiate themselves with respect to the full range of cognitive complexity intended to be measured on an assessment.

30. The second threat, construct-irrelevant variance, occurs when a test measures characteristics that are irrelevant to the intended domain. One example of this problem in PISA could occur if students who take the PISA math test on a computer need prior experience in successfully interacting with technology to correctly answer test items. If students who are good at mathematics get questions wrong because they are inexperienced in using a computer, their mathematics test scores will be underestimated because of the irrelevant construct (computer proficiency) that interferes with their ability to demonstrate their mathematics skills. For tests measuring mathematics and science, another potential source of construct-irrelevant variance may be linguistic complexity, for example, when students who possess strong math or science proficiency are unable to

correctly answer the test questions because of their difficulty in understanding the text associated with the items. Many scenario-based items attempt to make test questions more realistic, but can add unnecessarily to the reading load.

31. Protecting against these types of errors and promoting valid test score interpretations contribute to scientific integrity in educational testing. Thus, validity is an important component of scientific integrity because it represents a body of evidence built on theory and rigorous empirical research that supports the use of test scores for specified purposes.

32. The complex and sophisticated procedures used to develop and support PISA assessments are designed to provide information for the valid interpretation and use of PISA results. However, validity evidence is needed to confirm and justify each interpretation and use. Acquiring and analysing validity data is a comprehensive endeavor and so a description of the types of validity evidence needed to support the interpretation and use of PISA results is provided next.

2.3. Sources of validity evidence for building a validity argument

33. Gathering and summarising validity evidence to support the use of a test for a particular purpose is an arduous, but important endeavor. Since the validity of the use of a test for a particular purpose can never be absolutely "proven", the typical validation approach is to develop a strong "validity argument" (Kane, 2013[2]; Kane, 2006[7]) that justifies the intended use of the test by summarising a compelling body of evidence. The AERA et al. (2014[1]) *Standards* specify five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). The sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing. These five sources of validity evidence involve both quantitative (empirical) and qualitative (procedural) approaches and provide a framework for organising a validity research agenda and facilitating a sound validity argument.

2.3.1. Validity evidence based on test content

34. Validity evidence based on test content refers to the degree to which the content of an assessment is congruent with the assessment purpose (Pedrosa, Suárez-Álvarez and García-Cueto, 2014[8]; Sireci and Faulkner-Bond, 2014[9]). For educational tests such as PISA, this type of evidence addresses how well the knowledge and skill domain is defined, how well the items on a test represent this domain, how well the items measure the cognitive processes they are intended to measure, and the degree to which the test development process supports construct representation. Formerly known as "content validity," validity evidence based on test content addresses four important aspects: (a) domain definition, (b) domain relevance, (c) domain representativeness, and (d) appropriateness of test development procedures (Sireci and Faulkner-Bond, 2014[9]; Sireci, 1998[10]). The process used to develop the PISA frameworks is one example of validity evidence based on test content that demonstrates the domain was carefully designed through a consensus process (OECD, 2019[11]). Validity evidence addressing domain relevance and representativeness is usually gathered by having subject matter experts review test items and provide ratings regarding how well the items represent the objectives, content domains, cognitive levels, or real-world applications they are intended to measure. These types of studies include for instance analyses of content domain representation (Martone and Sireci, 2009[12]). These studies require recruiting a diverse panel of qualified subject matter experts who are knowledgeable of the populations tested,

training of these experts to review items, gathering data regarding their perceptions of the congruence of the test items to the measurement objectives, and analysing these data to evaluate this congruence (Sireci and Faulkner-Bond, 2014[9]). These experts can be teachers and other accomplished experts from business and society who are qualified to evaluate the knowledge and skills measured. Validity evidence based on the appropriateness of test construction procedures includes all of the quality control procedures in place to evaluate items statistically (e.g. item analyses) and qualitatively (e.g. conformance to item writing guidelines, sensitivity review) and to ensure item quality (e.g. training item writers, editing items, etc.).

2.3.2. Validity evidence based on response processes

35. Validity evidence based on response processes refers to results from studies that evaluate the degree to which test takers employ the cognitive processes a test intends to measure when they respond to test items. For example, if some test items are designed to measure how well students synthesise disparate types of information, evidence that students employed the higher cognitive skill of synthesis would be needed to confirm the items are eliciting the intended process of synthesising information. Confirming that the intended cognitive skills are being measured is difficult because cognitive processes are essentially invisible. However, test validators have used methods such as magnetic resonance imaging (Iancheva et al., 2018[13]; Testo et al., 2020[14]), think-aloud protocols, cognitive interviews (Padilla and Benítez, 2014[15]; Pepper et al., 2018[16]), and analysis of students' item response time and other log data from computer-based exams to understand the cognitive processes test takers use when responding to test items (Araneda et al., 2022[17]; Teig, Scherer and Kjærnsli, 2020[18]). For example, students' navigation patterns in the PISA 2018 reading assessment were associated with their reading proficiency levels (He, Borgonovi and Suárez-Álvarez, 2023[19]), such that students who visited all the pages and spent more time reading without rushing through the transited pages obtained the highest reading scores. The OECD has promoted validity research in this area by providing access to students' log data from their exams administered on computer as well as software for facilitating analyses of these data (e.g. Costa (2021[20])).

2.3.3. Validity evidence based on internal structure

36. Validity evidence based on internal structure refers to results from a variety of studies that evaluate the dimensionality and measurement precision of an assessment. A central aspect of validity evidence based on internal structure is confirming that the hypothesised dimensionality is reproduced in empirical analysis of the dimensionality of the data derived from students' responses to test and questionnaire items. For example, in each PISA main domain, the assessment is conceptualised as a unidimensional construct, which means a single dimension can be used to characterise students' performance in the subject area – although it should be noted that in practice, subscores are also estimated and reported to encourage more diagnostic discussion of results. Thus, the "hypothesised dimensionality" is characterised by a unidimensional scaling model (i.e. reading items only contribute to scores in reading), called item response theory (IRT)[1]. All of the documentation regarding the fit of these IRT models to the PISA data (see Chapter 9 in OECD (2020[21])) is an example of validity evidence based on internal structure. In addition, the stability of the PISA score scale over time is another important validity issue (OECD, 2020, p. Chapter 9[21]). That is, studies should be conducted to ensure the scores from assessments in different years are on the same scale, so that changes in student performance over time can be attributed to differences in student achievement, rather than to changes in the score scale itself.

37. Dimensionality assessment can provide important validity evidence for complex assessments such as PISA. For example, PISA assessments may have been developed based on specific hypotheses about sub-domains, such as the four domains of Creative Thinking application in the PISA 2022 innovative domain (visual, written, scientific problem solving, social problem solving). Whether these domains are observed in dimensionality analyses of PISA data is a critical validity question. Statistical techniques that can provide such validity evidence based on internal structure include exploratory and confirmatory factor analysis, multidimensional scaling, and other specialised methods. Evaluation of measurement precision (Chapter 3) and *measurement invariance* i.e. the consistency of the measurement structure across sub-groups of test takers or different versions of a test (Chapter 3) are examples of validity evidence based on internal structure (van de Vijver, Jude and Kuger, 2019[22]; Wells, 2021[23]). When invariance is evaluated at the item level, it is referred to as *differential item functioning*. PISA validity research includes comprehensive analyses of IRT invariance at the item and total scale levels (Glas and Jehangir, 2014[24]; Joo et al., 2021[25]; van de Vijver, Jude and Kuger, 2019[22]; OECD, 2020[21]).

2.3.4. Validity evidence based on relations to other variables

38. Validity evidence based on relations to other variables refers to studies that include test scores as one variable in a multivariate analysis of relationships or predictions. Given that test scores are intended to be manifestations of the constructs they intend to measure, hypothetical relationships between test scores and other variables can be proposed and tested by gathering data on variables external to the test itself. For example, scores on a PISA mathematics test would be hypothesised to be positively correlated with students' grades in a mathematics class. They would also be hypothesised to be less correlated with variables unrelated to or less related to mathematics, such as writing proficiency, dexterity, and verbal fluency. Campbell and Fiske (1959) distinguished between *convergent validity*, which refers to positive relationships among test scores and other variables that measure the same construct, and *discriminant validity*, which refers to relationships among test scores and variables that are presumed to be unrelated, or at least less related, to the construct measured. This distinction supported a "strong program" of validation (Cronbach, 1971[26]) where validators would propose thoughtful hypotheses about the expected relations between test scores and other variables and test them for expected convergent and discriminant relationships. The degree to which relationships among test scores and external variables hold up over subgroups of test takers is also a relevant validity issue, with consistent relationships across subgroups suggesting consistency in the validity of interpretations across subgroups.

39. Extensive examples of validity evidence based on relations of PISA test scores to other variables can be found in PISA initial and thematic reports (e.g. (OECD, 2019[3]; OECD, 2016[27]): these include performance gaps related to socio-economic status (OECD, 2019[3]), differences in performance across grade levels (Avvisati and Givord, 2021[28]), or the relationship between mathematics performance and students' (self-reported) familiarity with basic mathematics concepts (OECD, 2016[27]).

40. Lastly, when the purpose of a test involves predicting test takers' performance on a future outcome or domain (such as admissions test scores predicting subsequent performance in school, or PISA test scores reflecting readiness for the future), *predictive validity* studies are needed. Such studies require criterion data gathered at some point after the test is administered. The PISA Young Adult Follow-Up Study (Mamedova et al., 2021[29]) is one example of validity evidence based on relations to other variables within a predictive context. Predictive validity studies are contrasted with *concurrent validity*

studies that involve relationships among variables where the data are collected at similar points in time. In relation to this, it is particularly relevant to participating countries/economies in PISA to analyse the relationship between PISA scores and scores in other international large-scale assessments such as PIRLS (e.g. variations in PIRLS results can account for about 72% of the variation in PISA reading results across countries and economies (OECD, 2019, p. 68[3])). One should note however that this inference only holds at the international level and might not hold at the within-country level.

2.3.5. Validity evidence based on testing consequences

41. Validity evidence based on testing consequences refers to evaluating the degree to which the intended consequences of a testing program are realised, as well as the degree to which negative, unintended consequences do not occur. On one hand, validity evidence based on testing consequences can be thought of as validity evidence in general, because testing purposes specify intended consequences (e.g. test scores indicate how well students have mastered specific knowledge and skills and the curriculum can be adjusted accordingly). On the other hand, it is hard to predict all consequences that could emerge from a testing program and so how to approach gathering evidence of negative consequences can be unclear, and often evolves over the lifetime of a testing program. That said, there are many examples of studies that could be done to evaluate testing consequences (see Lane (2014[30]) for a review), as well as some good examples of applied studies in this area (e.g. (Breakspear, 2012[31]; Chung, 2016[32]; Jude et al., 2021[33]; Klieme, 2020[34])).Such studies can include surveys or interviews of students, teachers, policymakers and other stakeholders affected by a testing system, and analysis of curricular changes. Although not relevant to PISA, analysis of adverse impact can also be used as validity evidence based on testing consequences. One way to develop a research agenda for gathering validity evidence based on testing consequences is to seek out and document public criticisms of a testing program (e.g. the argument that PISA rankings lead to a narrowing of the curriculum) and treat them as hypotheses to be studied ((Sireci, 2013[35]; Sireci, 2020[5]). Chapter 5 (*Fairness*) also addresses issues related to testing consequences.

42. With respect to PISA, by 2012, several participating countries/economies had already initiated reforms in schools and education systems in response to PISA 2009 results (Breakspear, 2012[31]). These reforms ranged from revising and updating parts of the study plans and curriculum to deeper changes in the country's educational policies and practices, such as the use of PISA results to justify new educational laws. However, transferring policies and practices from one educational context to another does not always have the desired results (Chung, 2016[32]). On some occasions, translating findings from international assessments to educational practices at the school level can have unintended consequences (Jude et al., 2021[33]). An increasing number of publications have started evaluating the consequences of testing in PISA (Klieme, 2020[34]).

2.3.6. Synthesising validity evidence as a validity argument

43. The aforementioned five sources of validity evidence provide a helpful way to gather and organise validity evidence to support the use of a test for a particular purpose. Given that these sources of evidence include both procedural aspects (e.g. quality control procedures in test development, administration and scoring) and empirical evidence based on research (e.g. validity evidence based on relations to other variables) it is important to summarise and clearly report the evidence. This summarising and documenting step can be considered a synthesis of all validity evidence in the form of an "argument" that the use of test scores for their intended purposes is justifiable. The term "argument" as used in the context of the validity of educational tests refers to a compelling and logical proposal

that an analysis of the validity evidence supports the use of the test scores for their intended purposes. The synthesis of validity evidence represented in a validity argument can be presented in a validity chapter of a document such as a technical manual, or as a stand-alone document.

2.4. Guidelines for ensuring validity in PISA

44. The technical and other PISA documentation represents important sources of validity evidence for PISA. For example, OECD (2020[21]) describes PISA test design and development, administration, and scoring processes. Evidence can also be found in PISA initial and thematic reports (e.g. (OECD, 2019[3]; OECD, 2021[36])), and in publications from the research community (e.g. (Klieme, 2020[34]; Rutkowski et al., 2010[37]; Rutkowski and Rutkowski, 2016[38]).However, a strong validity argument is coherent, organised, comprehensive, and compelling (Kane, 2006[7]; Kane, 2013[2]) (OECD, 2019[3]). With these objectives in mind, the following guidelines should be pursued in documenting the validity evidence built into PISA assessments, and for conducting validity studies develop a sound and comprehensive validity argument to support the interpretation and use of PISA results.

2.4.1. Clearly articulate the purposes and intended uses of PISA results

45. Given that validity refers to the degree to which evidence and theory support the use of a test for its intended purposes, it is important that validation research is guided by these purposes. As mentioned earlier, PISA has public statements of the intended uses and purposes of PISA results (OECD, 2019[3]; Schleicher, 2019[39]; OECD, 2019[11]). It is important these statements are clearly articulated, repeated and emphasised in documentation related to PISA's validity argument.

In practice: This guideline is reflected in quality standards for all phases of the PISA cycle, from design to instrument development and reporting. Thus, the objectives of PISA should be clearly stated in various documentation that pertain to PISA's validity argument, including the proposal for the innovative domain (see Standards 6.1), the test and questionnaire design plans (see Standards 6.6), the tests and questionnaire frameworks (see Standards 7.1), the initial reports of results (see Standards 9.1) and the technical reports (see Standards 9.3).

2.4.2. Clearly describe the constructs intended to be measured by each assessment component

46. PISA assessment frameworks (e.g. (OECD, 2019[11])) define the constructs and describe the types of processes and scenarios exhibited in the tasks that PISA uses to assess them. They also discuss how student performance is measured and reported. These constructs include Reading, Mathematics, and Science as well as innovative domains such as Global Competence. Contextual Questionnaire frameworks describe the constructs measured in students, school, teacher and parent questionnaire. Descriptions of the constructs should support appropriate interpretations of PISA results. For example, possible threats to this principle arise if domain items (e.g. Mathematics, Reading, Science, Problem Solving) and its subscales items (e.g. single, and multiple source reading subscales) are not sufficiently differentiated between them. Although group means differences can still be observed among highly correlated constructs, construct representativeness should be clearly described to justify separate score reporting. Constructs that exhibit high conceptual overlap, such as social and emotional skills

included in the PISA questionnaires, should also provide evidence of the incremental value of each construct (Kankaraš and Suarez-Alvarez, 2019[40]; OECD, 2019[3]).

In practice: The standards for developing assessment frameworks and specifications for tests and questionnaires (Standards 7.1) require PISA frameworks to clearly define the constructs that will be assessed in each PISA cycle, to present the theoretical foundation underpinning the target domains and subdomains, and to clearly articulate what claims will be made about students' knowledge, skills, attitudes and contexts of learning. Furthermore, the standards on constructing and validating scales from cognitive scales and on constructing proficiency levels and descriptors specify that subscales must be defensible with respect to their measurement properties, evaluating the value of subscores beyond overall proficiency scores (Standards 8.3). For the questionnaires, scaled indices shall only be created for constructs with an appropriate number of items ensuring construct representation consistent with published theories about the respective construct, and correlations between constructs should be considered for validity evidence (Standards 8.4). These results are reported in the technical reports (Standards 9.3). Finally, the initial reports of results should clearly present the constructs and accurately reflect what the assessment measures as presented in the frameworks and supplemented by validation studies, avoiding extrapolations (Standards 9.1).

2.4.3. Identify the studies most needed to validate the intended purposes and uses of PISA assessments and evaluate negative consequences

47. Theoretically, there is an unlimited number of validity studies that could be done; however, an infinite validity agenda is not realistic. Therefore, identification and prioritisation of the studies most likely to provide important evidence for supporting the interpretation and use of PISA results is needed (Kane, 2013[2]). The five sources of validity evidence are helpful for identifying and prioritising these studies, but it is important to note not all five sources may be needed to justify a particular use or interpretation (Sireci, 2013[35]; Sireci, 2020[5]). Identification of the studies to be conducted can benefit from the voiced criticisms of a testing program.

48. Once the purposes and relevant criticisms are identified, studies should be proposed to confirm that the intended uses are being realised and unintended negative consequences are not occurring. For example, studies confirming that the content and other intended attributes of PISA assessments are on-target are certainly needed, and studies confirming that the intended response processes are being measured are also important. Steps taken in test construction such as item analyses, item calibration and model fit analyses, and measurement precision evaluations provide important internal structure validity evidence. Some testing purposes may require evidence involving relations to other variables (e.g. do students who score proficient on PISA demonstrate academic success on other measures), and some "invalidity" hypothesis may require demonstrating a lack of relationship between PISA scores and construct-irrelevant variables. Evaluating the consequences of the use of PISA results may require a longitudinal program of research that involves both shorter-term and longer-term studies that monitor the effects of PISA over time.

In practice: Conducting studies to validate the uses of PISA data is required in the standards from the stage of item development itself. Qualitative pilot studies, including cognitive laboratories, and quantitative pilot studies administered in multiple

countries and languages should be part of item development to confirm that the tasks effectively capture the targeted constructs, minimising construct irrelevant variance. The validation of items should notably include analysis of response process data, such as reports of thinking aloud processes in cognitive labs and log-files. This analysis is particularly important to detect differences in familiarity with the content and engagement across countries and groups of students. (Standards 7.2and 7.3). Quantitative studies that should be carried are further detailed in the chapter on scoring and analysis, such as conducting experiments to support changes in the design/methodology and inspecting psychometric features of tests in the Field Trial (Standards 8.1), conducting nonresponse bias analysis (Standards 8.2), evaluating the invariance of item parameters across countries and economies in all domains (Standards 8.3 and 8.4), etc. These results shall be reported in the Technical reports (Standards 9.3).

On the other hand, evaluating the consequences of the use of PISA results is still aspirational at this stage, and may be proposed to the PISA Governing Board as the focus of a future research project, e.g. as part of the PISA Research, Development and Innovation (RDI) Programme.

2.4.4. Synthesise the validity evidence to develop a validity argument and evaluate and improve the argument over time

49. The preceding steps involve identifying the studies to be conducted, designing them, and completing them. Rather than have these studies and documentation of test development processes exist in individual reports that remain unlinked, they can be synthesised into a validity argument that clearly and succinctly summarises their importance for justifying the use of PISA scores for their intended purposes. As studies are conducted, synthesised, and documented, that documentation should also be evaluated for what is needed in the future. That evaluation should be formative in that it informs future validation research, and results in continuous improvement in the valid interpretation and use of PISA results. As mentioned earlier, the validity argument could be presented as a summary to a validity chapter in a technical manual, or as a separate document. Variations of the validity argument documentation may be appropriate for different audiences that differ with respect to psychometric and technical expertise.

In practice: The PISA Technical reports gather evidence from studies that constitute validity evidence, but have historically not synthetised this evidence clearly into a validity argument, nor included evidence from validations studies that happen during item development. The Quality standards now recommend to present these validation studies (e.g. cognitive laboratories, quantitative pilots) in the technical reports, and to add a section dedicated to synthetising validity evidence, pointing to other specific sections of the technical reports for further technical details (Standards 9.3).

Chapter 3. Reliability and accuracy

3.1. Introduction and definitions of reliability and accuracy

50. The value of PISA to inform evidence-based policy making relies on the degree of precision with which population-level statistics are estimated and reported. For PISA to provide value for governments, PISA mean scores (and other indicators derived from PISA) need to be reliable, which means they are consistent (stable) from a measurement perspective. Consistency of measurement is straightforward in physical measures. For example, weighing oneself repeatedly on the same scale typically produces the same (consistent) weight. In education, reliable test scores reflect scores users can trust as consistent if the test were administered multiple times.

51. Reliability and accuracy are critical to valid educational and psychological testing. Essentially, reliability refers to the extent to which a measure is free of random measurement error, which would result from a lack of consistency in scores across (hypothetical) repetitions of the testing procedure. Here, it is important to distinguish between systematic and random errors. *Systematic errors* affect validity, but not reliability, and refer to situations in which factors (e.g. gender, ethnicity, fatigue, language, etc.) that are not conceptually relevant to the trait or construct (e.g. reading, science, or mathematics knowledge and skills) that the test is intended to measure affect the scores of test-takers. A *random error*, in contrast, has no relationship to other observed and unobserved variables and affects test-takers indiscriminately. For example, a systematic error in PISA could result from errors during the translation of assessment instruments (see Chapter 4 on Comparability for further details). In contrast, a random error could be due to fluctuations in the degree to which a student engages with a test on a particular testing occasion.

52. Tests that fail to reduce random measurement errors produce unreliable/imprecise[1] results. Generally, results that are not reliable threaten the possibility of drawing valid interpretations. Yet, a reliable result does not guarantee a valid interpretation either. In other words, providing reliable and precise test scores is a necessary, but not sufficient condition to make valid interpretations of the data (see Chapter 2 on Validity for further details).

53. For survey-based assessments such as PISA, the assessment is not designed to report individual scores, but population-level statistics. Thus, reliability and accuracy in measurement in PISA are not only affected by inconsistency in measurement at the test-taker level *(measurement error*) - which remains important because it affects these derived group statistics –, but also by difficulties in obtaining representative samples of 15-year-olds (*sampling error*). In addition, PISA cycles are meant to be compared across cycles, which adds another source of uncertainty associated with equating of scales (*linking error*).

54. These three sources of error – measurement error, sampling error, and linking error – together contribute to *total survey error* (Groves and Lyberg, 2010[41]). The measurement error is the smallest, accounting for around 0.5 to 0.8 of a country's mean performance estimate, depending on the domain. By contrast, the sampling error accounts for around

[1] This document uses reliability and precision as equivalent terms. In survey-based assessments, such as PISA, measurement precision is conceptualized as a lack of measurement error, which in addition to sampling error and linking error compose the total survey error (Groves and Lyberg, 2010[41]). Other specialized literature uses measurement precision to denote the more general notion of reliability and consistency regardless of the method used to calculate it (AERA, APA and NCME, 2014[1]).

2 to 3 PISA score points for most countries. Finally, the linking error varies between 1.5 to 3.5 score points for comparisons across cycles, depending on the cycles compared and the domain (OECD, 2019[3]).

55. Sources of error affect both cognitive assessment data and context questionnaire data. Since PISA's main interest has traditionally been in comparing mean test scores across countries, developing methods for improving measurement precision in the cognitive assessment has received most of the attention. Yet, most of the information provided in the PISA reports is based on context questionnaire data - e.g. almost every result accounts for students' and schools' socio-economic status. Although reliability for each scaled index is reported (see OECD (2020[21]), Chapter 16), there is room for researching and implementing new methods to improve contextual questionnaire scores' reliability estimation, such as using alternative reliability coefficients to Coefficient Alpha (Sijtsma, 2009[42]) based on frameworks that account for the covariance between different dimensions (e.g. Multidimensional Item Response Theory (MIRT (Ferrando and Navarro-González, 2021[43])).

56. This chapter presents factors that influence these three sources of error and thus affect reliability/precision and provides suggestions for documenting the reliability and accuracy of PISA scores.

3.2. Threats to reliability and accuracy

57. Minimising measurement error, sampling error, and linking error is crucial for reporting accurate test results and making valid comparisons between cycles. This section presents threats to reliability and accuracy in PISA associated to these three types of error.

3.2.1. Measurement error

58. The measurement error (also called imputation error) is the error associated with using a particular set of items to measure a certain skill. Unlike sampling and linking error, which are specific to PISA design features, the measurement error is common to all educational and psychological testing.

59. Score reliability is a function, among other things, of the number of items per student: measurement error typically decreases with the number of items used in the test, which in PISA context translates to the individual assessment time. This means that major domains in PISA (e.g. Reading in PISA 2018), which typically contain more items, have lower measurement errors than minor domains (e.g. Mathematics and Science in PISA 2018).

60. Measurement error is also related to the reliability of human coding, where this is required. Indeed, PISA assessments include constructed-response items, which, with a few exceptions, require human scoring (e.g. writing a short summary after reading a passage). Improving reliability and precision for this type of item requires ensuring a sufficient level of consistency in the ratings given by the raters (inter-rater reliability), and among ratings of a single rater (intra-rater reliability). Imagine two teachers are correcting the exams of the same group of students. A high inter-rater reliability means that both teachers would assign the same score to a given student. In large-scale educational testing such as PISA, consistency in scoring is often achieved by ensuring raters apply the same scoring rubric to assign their scores - so a particular test-taker's score is unaffected by the specific person who graded their response. High intra-rater reliability means that each rater is consistently applying the same rubric/scale to assign scores to students, so factors such as the order in which the scorer corrects the exam have no significant impact on test takers' scores. Increasing the number of raters (for the same response) typically improves reliability.

The use of machine scoring also allows to avoid human errors and improve reliability (Yamamoto et al., 2017[44]).

61. Score reliability is also influenced by the variability in responses: tests in which the variability in the responses is low tend to exhibit low estimates of score reliability. In other words, reliability estimates can only be as high as the variability in the responses. A test in which most examinees answer all items correctly (or incorrectly) or agree (or disagree) with all the statements presented will not have high estimates of reliability because there is not enough variability in the responses. Perhaps more illustrative is an example taken from the OECD Survey on Social and Emotional Skills in which the scores of 10-year-olds typically showed comparatively lower reliability coefficients than 15-year-olds partly due to lower variability in the responses of younger cohorts (Kankaraš and Suarez-Alvarez, 2019[40]; OECD, 2021[45]). The same principle explains why dichotomous response formats (right/wrong, yes/no) usually require more questions to obtain the same precision as a 4- or 5-point, multiple-choice format.

62. Score reliability is also a function of how well the test difficulty matches the ability level of students. Generally, nationally representative samples will result in wide variation in the responses, which positively affects reliability. Yet, this is only true if the questions included in the test are as varied as the population, allowing the population variability to manifest itself in different patterns of answer. Other factors, such as the good quality of the items (e.g. if a question is good at showing the difference between high- and low-proficiency students), also drive low measurement errors (see, for example, results of the hybrid model in PISA 2015 Technical Report (OECD, 2017[46])).

63. Lastly, measurement error in the context of PISA also relates to the quality of the conditioning (imputation[2]) model. The quality of the conditioning model, that is, how well the contextual questionnaire data can predict the cognitive items, directly impacts the amount of measurement error. If the relationships are strong, the test results are likely to be more reliable and accurate. If not, there might be discrepancies, leading to potential inaccuracies in assessing the student's true capabilities.

64. Importantly, low measurement error should not be achieved by sacrificing content validity. For example, replacing long scenario-based items with multiple related items may increase the reliability of the scores. Still, it would jeopardise construct representation and claims that relate the scores to real-life situations (see Chapter 2 on Validity for further information). The same would apply to context questionnaire data; introducing content redundancy in the items may help increase the reliability of the scores at the cost of missing the opportunity to represent better the construct measured.

65. In comparison to the sampling and linking error, the measurement error in PISA is the smallest. The introduction of multi-stage adaptive testing (MSAT) in reading in PISA 2018 has contributed to reducing measurement errors ((Yamamoto, Shin and Khorramdel, 2019[47]). The extension of MSAT to other domains (mathematics in PISA 2022, and science in PISA 2025) is likely to reduce measurement error for these domains as well.

[2] See OECD (2020a, Chapter 9) for a description of the imputation methodology for obtaining plausible values for proficiency scales and for using these to estimate descriptive statistics for populations and subpopulations.

3.2.2. Sampling error

66. The sampling error is the error associated with selecting a representative sample of 15-year-old students instead of the full population of 15-year-olds for each participating country/economy in PISA. Except in small countries, which conduct a census of all 15-year-old students, the sampling error is by far the most important source of error in PISA country-level results in terms of magnitude and its consequences for data interpretation – which enable generalising the results at the system-level (OECD, 2019, p. 45[3]). PISA uses two-level sampling in order to reduce the costs of data collection: schools are sampled first and students from within the selected schools are sampled in a second stage. This means that the sampling error decreases the greater the number of schools and (to a lesser extent) of students included in the assessment. Sampling error is also reduced through the use of adequate stratification variables: when predictors of test performance at the school level are included among the stratification variables, the sampling error is smaller for an equivalent number of sampled schools.

3.2.3. Linking error

67. The linking error is the error associated with comparing PISA results across cycles (e.g. PISA 2018 with PISA 2022). To make PISA results directly comparable across cycles, scales need to be equated. This means that results are transformed so that they can be expressed on the same metric. The linking error only affects trend comparisons, but in those comparisons, it can be the largest source of error – even larger than the sampling error. Since PISA 2015, the linking error was minimised due to improvements on the test design (e.g. greater number of trend items), scaling procedures (e.g. concurrent calibration), and the absence of framework revisions for assessing mathematics and science (OECD, 2020[21]).

68. The sampling, linking, and measurement error described in this chapter are sources of error that can be easily quantified in the context of PISA. However, this does not mean they are the only sources of error affecting the reliability of PISA results. Minimising both quantifiable and non-quantifiable sources of errors is important to maximise reliability/precision of PISA results. For instance, test administration conditions can also impact reliability. Test administration shall be consistent, not only in the questions and scoring procedures but also in the conditions in which students answer the test. This includes the physical space (the quieter and more comfortable, the better), time limits (absence of time pressure), and instructions (the clearer, the better). For example, school disruption during the Covid-19 pandemic threatened the reliability of assessments during that period because of the extra difficulty of ensuring consistent test administration conditions among students (Sireci and Suarez-Alvarez, 2022[48]).

3.3. Guidelines for ensuring reliability and accuracy in PISA

69. Test developers and distributors have the primary responsibility for obtaining and reporting reliability evidence. In PISA, this responsibility primarily relies on the OECD Secretariat and the international contractors (PISA Consortium). However, final users of PISA (governments represented on the PISA Governing Board, practitioners, and researchers) are also responsible for documenting the precision of the indicators derived from PISA data when making decisions that impact educational policy and practice. Since the reliability of measures varies substantially across countries, it is all the more important for data users to report reliability for their particular context.

70. Various independent professional bodies have developed standards to address common issues when documenting reliability to inform evidence-based interpretations –

see for example Chapter 2 of the Standards for Educational and Psychological Testing (AERA, APA and NCME, 2014[1]), and Guideline 2 of the International Test Commission Guidelines for the large-scale assessment of linguistically and culturally diverse populations (ITC, 2019[49]). The literature on large-scale international assessment has also significantly contributed to addressing common and specific issues when documenting reliability in large-scale international assessments. See, for example, PISA 2022 Technical Standards (OECD, 2020[21]; OECD, 2020[50]) and the Technical Standards for IEA Studies (Martin, Rust and Adams, 1999[51]).

71. Despite the general agreement among experts on what is required to ensure and document scores' reliability, PISA test scores have too often been used inappropriately (Klieme, 2020[34]). Having in mind PISA goals and assessment design features, the following guidelines concerning reliability are proposed to inform evidence-based interpretations of PISA test scores. These guidelines apply to both cognitive assessment data and contextual questionnaire data. Each of the following guidelines should be ensured for the interpretation of each intended score use and adequately documented for each intended user:

3.3.1. Systematically communicate the uncertainty associated with statistical estimates to the intended users of these statistics

This guideline advocates for identifying the potential sources of error – quantifiable and non-quantifiable - for each population-level estimate and intended use. In PISA these include, but are not limited to, within-country group-mean comparisons, between-country group-mean comparisons, percentage of students below minimum proficiency level, and trend comparisons. It is important to realise that individuals with very different statistical knowledge use PISA scores. Overexplaining the results to specialised audiences might be a burden while underexplaining the results to non-specialised audiences might be misleading. In all cases, PISA results should always be accompanied by standard errors or confidence intervals that reflect the uncertainty associated with the sample estimates to make appropriate inferences about the population parameters (e.g. means and proportions). The fact that statistics are good enough to be included in a PISA report doesn't mean that all have the same quality (see Chapter 4 on Comparability for further details). In the case of reports targeting specialised audiences, such as the PISA Technical Report, sufficient information shall be provided to fully reproduce the results. This may come with the need for PISA National Centres and the PISA consortium to agree on releasing intermediate steps in the process, which otherwise would be impossible for secondary analysts to guess.

In practice: These guidelines are reflected in the standards related to reporting (Chapter 9). The international reports of results shall contain all the necessary information readers need to understand the results and draw appropriate inference, including how to interpret differences in PISA scores, discussions of statistical and practical significance and a description of the sources of uncertainty in PISA. Furthermore, the uncertainty associated with aggregate statistics (standard errors or confidence interval) shall always be reported (Standards 9.1). For specialised audiences, the technical reports shall precisely describe how analyses were carried (Standards 9.3), and code scripts enabling external researchers to reproduce results shall be made available, along with extensive metadata to facilitate their use (Standards 9.2).

3.3.2. Provide appropriate evidence of reliability/precision at the test and proficiency level[3] for each participating jurisdiction and adjudicated region

72. This guideline advocates for documenting evidence of reliability/precision at the overall test score (scale score), and proficiency level for each participating jurisdiction and adjudicated region in each PISA cycle. Likewise, knowing the reliability/precision of test scores is crucial to ensure the scores meet acceptable reliability/precision criteria. This information is not only relevant for test developers and researchers, but also for any users intending to use the scores to make inferences and interpretations.

73. When it is not possible to provide widely accepted reliability/precision indicators (e.g. coefficient alpha, test information functions, conditional standard errors, etc.), alternative indicators shall be sufficiently documented. For example, the introduction of MSAT in reading and the rotated and incomplete assessment design in other domains of recent PISA cycles forbid calculating classical reliability coefficients for each cognitive domain. Instead, score reliability in PISA 2015 and PISA 2018 was estimated using the formula: 1 – (expected error variance/total variance) (OECD, 2020[21]).

74. When inferences and interpretations are made for subgroups of students, and if reliability estimates or standard errors for one or more of these groups differ significantly from most others, information on the reliability/precision of scores by those subgroups should also be documented.

In practice: Reliability indicators are presented for each country/economy and adjudicated region in the Technical reports (Standards 9.3).

3.3.3. Provide appropriate evidence of scoring agreement for constructed-response items within and across each participating jurisdiction and adjudicated region.

75. This guideline advocates for documenting evidence of scoring agreement at the item- and domain level, as well as by coding category (e.g. no credit, partial credit, full credit). In PISA 2018, at least one fourth of new reading items in the computer-based assessment were human-scored constructed-response items (Annex A, OECD, 2020a). Given this important proportion of human scored items in the assessment, it is important to ensure their reliability/precision in order not to jeopardise the reliability/precision of the scores for the entire assessment and the possibility of using these scores for valid interpretations. Since one of the goals of PISA is comparing results across subgroups of students and across countries, the scoring agreement evidence should be provided within and across each participating jurisdiction. Whenever possible, this process should be optimised and streamlined using technology to avoid human errors and reduce coding burden (Yamamoto et al., 2017[52]).

[3] This is related to the concept of decision consistency in the specialized literature. Decision consistency refers to the extent to which the observed classifications of examinees would be the same across replications of the testing procedure (AERA, APA and NCME, 2014[1]). However, PISA does not classify students but items into different proficiency levels (see (OECD, 2020[21]), Chapter 15).

In practice: Evidence of scoring agreement in open-ended items is presented in the Technical reports (Standards 9.3).

3.3.4. Clearly describe reliability information in the Technical Report and in the initial reports

76. For the goals of PISA, it is crucial to ensure that the scores meet acceptable reliability/precision criteria for every participating jurisdiction and adjudicated region (ITC, 2019[49]). It is equally important to provide sufficient details regarding the methods used to estimate the indices reported, the nature of the group from which the data were derived, and the conditions under which the data were obtained (AERA, APA and NCME, 2014[1]). Technical information of these procedures should be exhaustively described in the PISA Technical Report. Yet, since PISA Technical Reports are usually not available in full at the time of the initial report's release, sufficient information should be included in the initial reports to adequately interpret the data according to these guidelines. This information should be documented for each intended user (e.g. policy makers, practitioners, researchers) in a language the user is familiar with and for each intended score use (e.g. within-country group-mean comparisons, between-country group-mean comparisons, percentage of students below minimum proficiency level, trend comparisons, etc.).

In practice: The Technical reports present the methods used for estimating reliability indices, along with detailed descriptions of the sample and administration process (Standards 9.3). The initial reports are further required to present all the necessary information for readers to draw appropriate inferences, including descriptions of the population, procedures for data collection, and user-friendly explanations of reliability information. In addition, data quality indicators are required to be reported along with the results (Standards 9.1).

Chapter 4. Comparability

4.1. Introduction and definitions of comparability

77. Comparability refers to the degree to which examinees' scores on a test can be meaningfully compared, as well as the degree to which aggregate statistics (e.g. country-level mean scores) derived from test scores can be meaningfully compared (AERA, APA and NCME, 2014[1]; Berman, Haertel and Pellegrino, 2020[53]; ITC, 2019[49]; ITC, 2017[54]) . PISA score users should be confident that countries and educational jurisdictions with the same score have a population of students who are on average equally proficient with respect to the knowledge and skills the test was intended to measure. However, comparability is never an all-or-nothing decision: confidence in the validity of comparisons follows a continuum and varies as a function of several factors. For instance, it decreases as the content of the test, the administration conditions, and students' samples diverge.

78. Given that the intended use of PISA scores is to construct aggregate indicators, and to compare these internationally, comparability is a critical cornerstone of validity in PISA. PISA is the largest international large-scale assessment of educational outcomes available. Since its inception in 2000, PISA has grown from 28 countries and jurisdictions to more than 80 in 2022. PISA provides policymakers with unique indicators about their education system that enable them to benchmark their results to other education systems and track them across time (trend comparisons). For PISA to provide value for its stakeholders, PISA cognitive assessment and context questionnaire scores obtained from students need to be comparable, both within countries and across different countries, languages, and cycles. This chapter defines basic requirements for a comparability argument and provides suggestions for documenting the evidence supporting comparability in the context of PISA.

79. Like all large-scale assessments, PISA is carried out through standardised tests in which all people are evaluated with the same, or equivalent, test forms, conditions, and scoring protocols. If the test and testing conditions are not standardised, the scores are not likely to be comparable (e.g. if some students have more time than others to answer the questions or have received help from their teacher during the test administration). However, standardisation does not guarantee valid comparisons. In other words, standardisation is necessary, but insufficient, to ensure score comparability.

80. In some instances, greater comparability of scores may be attained if standardised procedures have flexibility to adapt the test to the situation and characteristics of the examinees and the interpretations of the results are made based on these conditions (Randall, 2021[55]; Randall et al., 2022[56]; Sireci, 2020[57]; Sireci and Suarez-Alvarez, 2022[48]). For example, sampling rules in PISA allow the limited[4] exclusion of special education students and students with insufficient assessment language experience. However, students with special education needs should not be automatically excluded, if they are able to sit the test. In order to enable their participation, countries may resort to a number of adaptations for these students, including the use of a special *Une-Heure* booklet - a shorter version of the test designed for students with special education needs. Adaptations that compensate for a physical or intellectual disability are an example of how comparability may be enhanced by deviating from the standardised procedures.

[4] After all exclusions (school and within-school), the resultant population is required to cover at least 95% of the desired target population (OECD, 2020[50]).

4.2. Threats to comparability

81. Comparability of country-level indicators requires both the comparability of individual measurements (scores), the consistent translation of assessment instruments and the consistent quality of samples from which country-level inferences are drawn. The basic principle for ensuring individual comparability is to avoid changing factors – across countries and cycles - that may generate differences in scoring that are not related to examinees' abilities (Berman, Haertel and Pellegrino, 2020[53]). In the context of trend comparisons, (Beaton and Zwick, 1990[58]) phrased this principle as "when measuring change, do not change the measure" (p. 165). However, a rigid application of this principle is not always possible, nor desirable, in assessments like PISA. For example, in order to reflect changes in society and technology, and to continue measuring what students can do in real-life situations, periodic changes in the tests are unavoidable. Changes in construct definitions in the assessment frameworks may affect trend comparisons (which would look different if such changes were not introduced), but, if appropriately justified and documented, do not threaten comparability, because they reflect adaptations to changing conditions. For example, one of the changes in the reading framework between 2009 and 2018 was a greater emphasis on multiple-source texts which are common when reading in digital environments (OECD, 2021[36]). Other times, changes may be driven by unexpected events like school closure during the Covid-19 pandemic. This type of change would likely impact within- and between-country comparisons during that year(s) as well as future trend comparisons with the affected cycle (OECD, 2020[50]; Sireci and Suarez-Alvarez, 2022[48]).

82. The main comparability threats in PISA are driven by variations in (a) sample coverage (b) translation and adaptation, (c) construct coverage, (d) assessment conditions, (e) test administration, and (f) scoring criteria.

4.2.1. Variations in sample coverage

83. PISA's primary challenge to minimise unjustified variations which would threaten comparability lies in enforcing 80+ PISA participating jurisdictions as well their contractors to meet PISA Technical Standards (AERA, APA and NCME, 2014[1]). The purpose of the PISA Technical Standards is to specify the data collection procedures to create an international dataset of a quality that allows for valid cross-national inferences. For example, Standard 1.1 specifies that the PISA target population is composed of "students between 15 years and 3 (completed) months and 16 years and 2 (completed) months at the beginning of the testing period", which ensures that the tested students have approximately the same age across countries."

84. These and other standards were designed with the objective of ensuring that PISA results are representative of the target population in all adjudicated countries/economies. PISA implements a data adjudication process each cycle through which each national dataset is reviewed and a judgement about the appropriateness of the data for the main reporting goals is formed (OECD, 2020, p. Chapter 14[21]). However, even when these standards are met, comparisons should still remain with reference to the target population only. For example, PISA results cannot be generalised to the entire population of 15-year-olds, particularly in countries where many young people of that age are not enrolled in lower or upper secondary school (see Chapter 3 in (OECD, 2019[3])).

85. It is also important to consider that the size and composition of the students' and schools' populations could change over time in ways that affect trend comparisons. The most obvious and recent example of a phenomenon that could significantly change the composition of student samples are school closures during the Covid-19 pandemic,

which may vary between districts/states/countries. Data from COVID-disrupted school years are not likely to be valid for accountability purposes but may be valuable for making decisions at the individual student level (Sireci and Suarez-Alvarez, 2022[48]). Other examples include the "opt-out" movement where parents elect to excuse their children from taking standardised assessments, and the large influx of refugees in some countries which may lead to higher school enrolment.

4.2.2. Variations in translations and adaptations

86. PISA 2022 is carried out in more than 80 countries and educational jurisdictions. This means PISA assessment instruments and administration manuals need to be translated and adapted to a large number of languages. PISA Translation and Adaptation Guidelines (OECD, 2019[59]) and PISA Technical Standards (OECD, 2020b) establish the requirements participating jurisdictions in PISA need to meet to develop equivalent national versions of the assessment instruments that enable collecting internationally comparable data. These guidelines are based on professional bodies' guidelines such as the International Test Commission (ITC, 2013[60]; ITC, 2017[54]; ITC, 2019[49]), and aim to avoid biases introduced during the translation process that could distort international comparisons. For example, comparability can be compromised if the difficulty of the items is unintentionally modified, or the administration manuals are translated to the national context in ways that change the data collection. Some of the most frequent systematic errors (see Chapter on Reliability) include translating words or expressions that exist in the English version, but they do not exist in the language of the translated version. Other examples include adapting the names of people, animals, or places to equivalent terms that are not as frequent or easy to read as in the English version. Due to the psychological nature of the constructs included in the contextual questionnaires, translation, and adaptation of these types of items might be sensitive to different understanding across countries, e.g. raising the hand in class could be seen differently depending on the culture. To minimise these sources of errors it is recommended to include subject-matter experts in addition to linguists and translators when translating these types of items to avoid introducing construct-irrelevant variance. With the increasing number of countries and jurisdictions participating in PISA, it is also critical that these guidelines and standards ensure test fairness (see Chapter 4 on Fairness) and score comparability to support meaningful inferences in culturally and linguistically diverse context (ITC, 2019[49]).

4.2.3. Variations in construct coverage

87. PISA administers a large number of items in a limited time. To achieve this goal, PISA uses different strategies. One of them is the use of a rotated test design in which students take different but overlapping tasks. Another one is the introduction of the multi-stage adaptive testing (MSAT) for the reading domain in PISA 2018 (OECD, 2020[21]) and for mathematics in PISA 2022. In traditional testing, all participants are provided a fixed test form regardless of their ability level. In adaptive testing, all participants are provided with a set of items that match their estimated ability level as the test proceeds (Zenisky, Hambleton and Luecht, 2009[61]). Unlike item-level adaptive tests, the smallest unit of adaptability in MSAT is a set of items instead of a single item.

88. Importantly, these test design decisions mean that a relatively small number of items are administered to each student and, consequently, the accuracy of the measurement at the individual level is poor (in comparison with the reliability of high-stakes individual assessments, for example). Therefore, PISA scores should not be used to compare individuals. However, the use of plausible values, which account for these

test design features, allow obtaining unbiased group-level estimates (von Davier, Gonzalez and Mislevy, 2009[62]). Last but not least, the introduction of the MSAT may impose additional comparability issues with results from prior non-adaptive PISA cycles ((OECD, 2020[21]), Chapter 9). From a construct coverage perspective, introducing adaptive testing may also challenge cross-country comparability if the adaptive algorithm does not ensure equivalent construct coverage across countries – although this was treated to a certain degree with the introduction of probabilistic misrouting. In PISA 2018, the adaptive design provided the minimum number of responses per item needed for IRT scaling and an appropriate item coverage across all range of item difficulties and framework aspects cycles ((OECD, 2020[21]), Chapter 9). Variations in construct coverage between PISA and PISA for Development may also occur as a result of extending the framework and assessment items to the lower end of the proficiency scale to provide better coverage of basic processes.

4.2.4. Variations in assessment conditions

89. Variations in assessment conditions could influence group-level comparability, especially when these variations are not distributed at random across countries and jurisdictions (Berman, Haertel and Pellegrino, 2020[63]). The origin of these variations can be very diverse. For example, the value of PISA might be more generally accepted by the education community in some countries than others, leading to more motivated students or even teachers using PISA assessment frameworks and released items as instructional materials. Other times, variations are direct consequence of changes in the test design. For example, beginning in 2015, PISA became a computer-based assessment with a paper-based option for a small number of countries. To address possible effects associated with this change, a mode effect study was conducted to examine whether tasks presented in one format function differently when presented in another format (OECD, 2017[46]). As digital assessment continues to grow, another increasingly important source of variation is student's familiarity with technology. Some students may be more familiar with responding to technology-enhanced innovative items than others because their national assessments use similar items, or they have been more exposed to technology. In that case, the resulting aggregated performance differences might be confounded by familiarity with the item format and access to technology - i.e. students with low technology familiarity may struggle more in a computer-based assessment than their equals on a paper-based assessment but also compared to more tech savvy students on the computer-based assessment.

4.2.5. Variations in test administration

90. A specific case of assessment conditions is test administration. Even when more general assessment conditions are met, like two countries administering the computer-based assessment, differences in how the test is administered may still compromise the comparability between them. These variations are particularly important in the case of PISA due to the great number of participating countries and jurisdictions. For example, students at the end of the testing window may have comparatively higher knowledge and skills than students at the beginning of the testing window. The same phenomenon applies to differences in the general environment or context in which the test is administered. For example, in 2018, some regions in Spain conducted their high-stakes exams for tenth-grade students earlier in the year than in the past, which resulted in the testing period for these exams coinciding with the end of the PISA testing window. Because of this overlap, students in Spain may have been negatively disposed towards the PISA test and as a result may not have tried their best to demonstrate their proficiency.

4.2.6. Variations in scoring criteria

91. Equally important is to ensure that the scoring criteria after data collection meet PISA Technical Standards (OECD, 2020[50]). For example, differences in human-scored constructed-response items across countries and jurisdictions may reduce the reliability of the scores (see Chapter 2 on Reliability) and compromise scores' comparability (see Chapter 13 in (OECD, 2020[21])).

4.3. Guidelines for ensuring comparability in PISA

92. Having in mind PISA goals and assessment design features, the following guidelines concerning comparability are proposed to inform evidence-based interpretations of PISA test scores. These guidelines apply to both cognitive assessment data and contextual questionnaire data. Each of the following guidelines should be ensured for the interpretation of each intended score use and adequately documented for each intended user.

4.3.1. Provide appropriate evidence of country-by-language equivalency for each participating jurisdiction and adjudicated region

93. Assessing invariance in PISA main survey cognitive assessment data means assessing equivalency across country-by-language groups and over time (Chapter 12, (OECD, 2020[21])). Equivalence across countries, languages, and cycles can be documented through statistical tests of equality constraints on item parameters. For example, the proportion of invariant items in PISA 2018 computer-based assessment was near or above 90% for all domains (OECD, 2020a, p. 2). Generally, this means the scores across countries and cycles can be compared and the measurements have reached a reasonably good level of comparability. When strict equality constraints are rejected for some parameters, these violations should be addressed in order to preserve comparability, and the rationale and consequences of the choices shall be documented. For example, in PISA 2018, a partial-invariance model which allowed for country/language-specific item parameters was used in the cognitive assessment data (von Davier et al., 2019[64]) and in the contextual questionnaire data (Buchholz and Hartig, 2019[65]). When reaching measurement invariance requires allowing item-by-country interactions, then a further analysis of how these statistical decisions impact the construct coverage should be fully examined and documented. Indeed, when parameters are freed for many variables/items, it is possible that the international scale does not have sufficient items either at some specific levels or for some sub-domains.

94. A promising avenue for providing additional evidence of country-by-language equivalency of cognitive assessment data is the use of process data. For example, response times can be used to identify differences in the response process between different cultural and linguistic groups to improve the validity of interpretations that might otherwise be overlooked in differential item functioning (Ercikan, Guo and He, 2020[66]). Other examples of promising avenues include researching and implementing new methods to enhance contextual questionnaire comparability, such as forced-choice items (Brown and Maydeu-Olivares, 2011[67]), situational judgement items (McDaniel et al., 2007[68]), and the triangulation between different assessment methods (Walton et al., 2022[69]).

In practice: The Technical reports provide information on the scaling procedures and outcomes, including group-specific item parameters for different language versions (country-by-language groups model fit) (Standards 9.3). The standards on constructing and validating scales further detail the procedure to follow when misfit is detected (Standards 8.3).

4.3.2. Report quality indicators for population coverage and response rates for each participating countries/economies and adjudicated regions

95. Together with country-by-language equivalency, the quality of the sampling outcomes is the most important quality indicator to interpret group-mean comparisons in PISA. While the sampling standards ensure that PISA results are representative of the target population in all adjudicated countries/economies, they should not be readily generalised to the entire population of 15-year-olds in countries where many young people of that age are not enrolled in lower or upper secondary school. For example, although the population coverage exceeds 80% in most OECD countries, it is still comparatively low in OECD countries such as Colombia (62%), Mexico (66%), and Turkey (73%), and it is below 50% in Baku (Azerbaijan) and below 60% in Jordan and Panama (OECD, 2019d, p. 48). To minimise the risk of PISA results being misinterpreted, this information should be reported in the technical documentation and when reporting PISA results.

96. A wider use of the out-of-school component of PISA for Development assessing the skills of 15-year-olds who are not currently in school could mitigate this problem (OECD, 2018[70]; OECD, 2017[71]). Yet, because the relevance of PISA for Development relies on the comparability with international PISA results, both assessments need to be developed using the same assessment frameworks and should share sufficient items in common to enable these comparisons. This means that differences in construct coverage and assessment conditions (e.g. paper- versus computer-based administration) between PISA and PISA for Development should be evaluated, maintained, and documented periodically (i.e. every cycle or every time assessment frameworks and items are updated).

In practice: The initial reports of results present the population coverage as part of the data quality indicators that should accompany the results (Standards 9.1). The Technical Report includes further details on the quality of the sampling design, including a presentation of the target population and sampling frame, how sampled schools were identified and tracked, special school sampling situations (e.g. small schools) and additional sampling options (e.g.oversampling, grade-based sampling), as well as students and teachers selection procedures. The Technical reports reports also provide sampling outcomes such as quality indicators for population coverage, school and student response rates (by country and adjudicated region), teacher response rates, and design effects (Standards 9.3).

4.3.3. Contextualise PISA results by providing information about concurrent changes or differences which are relevant to the result

97. This guideline does not address a comparability issue per se, but an issue related to the validity of the conclusions that are drawn from scores comparisons. It is critical to put results in perspective to evaluate the magnitude of the differences. Multiple reasons might explain why a country scores above another in a ranking, and not all these reasons are connected to the quality of education systems. Differences in performance between students within the same country are, in general, larger than between-country differences

in performance (OECD, 2019[3]). Small country-mean differences that are not statistically significant or practically meaningful should not be emphasised.

98. Similarly, country differences should be interpreted considering the contextual factors that might explain these differences. For instance, whether students commonly speak the language of instruction at home is associated to students' performance in reading in that language (OECD, 2019[72]). Therefore, not speaking the language of instruction represents an additional barrier to attaining high proficiency in reading. Essentially, the same would apply to any factor that is closely related to performance such as the country's national income.

99. Finally, PISA results should be interpreted considering differences in how education is organised across grade levels, particularly in school systems where students progress through different types of educational institutions at the pre-primary, primary, lower secondary and upper secondary levels. In many cases, 15-year-old students have been in their current school for only two to three years. This means that much of their academic development took place earlier, in other schools, which may have little or no connection with the school in which they were enrolled when they sat the PISA test.

In practice: The standards on producing the initial reports recommends that, whenever possible, PISA results are contextualised by presenting the factors that might explain the observed differences across countries and economies and across time. In doing so, it is important to avoid any causal claim that are not supported by the data (Standards 9.1).

100. Guidelines regarding translation are developed in the next Chapter on Fairness.

4.3.4. Examine and document variations in the quality of the reported PISA results

101. The PISA data adjudication process ensures a minimal appropriateness of the data for the main reporting goals ((OECD, 2020[21]), Chapter 14). However, these decisions often leave data users with no idea about the extent to which PISA results are comparable. The fact that summary statistics based on test-score data are good enough to be included in a PISA report does not mean that all data have the same quality or that the quality is sufficient for other goals. Comparisons based on PISA data should be accompanied by indications of the extent to which sample or translation quality, or issues related to scaling and test administration, may affect their validity. Identification of these issues or problems is also needed.

102. For example, there are two types of indices derived from questionnaires administered to students, school principals, teachers and parents. Simple indices are variables constructed through the arithmetic transformation or recoding of one or more items (e.g. highest parents' education); and scaled indices are those constructed through the scaling of multiple items (e.g. household possessions, or student's self-efficacy). Unfortunately, measurement invariance cannot be tested on simple indices such as self-reported effort spent on the PISA test because they are too small a unit of analysis. Consequently, cross-cultural comparability in simple indices cannot be statistically put to the test. However, measurement invariance can be tested in composite indices such as household possessions or student's self-efficacy. Like for cognitive data, assessing invariance in PISA context questionnaire data means testing country-by-language equivalency (Chapter 16, (OECD, 2020[21]). This means that for each item and scale analyses on the invariance of item parameters across countries, and across languages within

a country, need to be conducted and documented. The cross-country comparability of each scaled index in PISA should be documented in PISA Technical reports (e.g. (OECD, 2020[21])). Generally, this means that scale index scores across countries and cycles can be compared but, again, this is not a given, and caution is advised when interpreting comparisons.

103. Finally, indicators based on students', parents', teachers', and principals' reports are susceptible to several possible measurement errors: memory decay; social desirability (the tendency to respond in a manner that is more acceptable in one's own social and cultural context); reference-group bias (what the comparison group is); and response-style bias (e.g. straight-lining, over-reporting, modesty, heaping, acquiescence). These biases can operate differently in different cultural contexts and reading comprehension levels, thus limiting the cross-country comparability of responses (Benítez, Van de Vijver and Padilla, 2019[73]; Buchholz, 2022[74]; Lee, 2020[75]; Suárez-Álvarez et al., 2018[76]; Vigil-Colet, Navarro-González and Morales-Vives, 2020[77]). For example, one potential quality indicator could be that system-level relationships hold within-country. Another potential quality indicator could reflect on the extent to which cultural possessions are indicative of socioeconomic status in each country (Avvisati, 2020[78]; Lee and Borgonovi, 2022[79]).

104. Information about variations in the quality of PISA data that can affect comparisons should be documented for each intended user (e.g. policy makers, practitioners, researchers) in a language the user is familiar with and for each intended score use (e.g. within-country group-mean comparisons, between-country group-mean comparisons, percentage of students below minimum proficiency level, trend comparisons, etc.).

In practice: The standards on producing the initial reports state that data quality indicators shall be developed and reported along with the results, which include quality indicators showing the extent to which sampling (including population coverage), translation, scaling or test-administration issues affect the validity of comparisons. Both the quality of country data and quality of measures which do not meet comparability and reliability requirements should be acknowledged and made visible in an appropriate way, for example in the notes to the figures or tables, or in a reader's guide at the beginning of each report. (Standards 9.1).

Chapter 5. Fairness

5.1. Introduction and definitions of fairness

105. Oxford Languages defines fairness as "impartial and just treatment or behaviour without favouritism or discrimination." (Oxford English Dictionary, 2023). PISA data are used to make inferences about groups of individuals, and inform decisions (such as reforms) that impact entire education systems. For such assessments to be considered "fair," the assessment system should not favour or privilege one type of student, culture, or country over another; and should provide an objective and unbiased platform for measuring the intended constructs. Fairness in testing involves impartiality and unbiasedness with respect to the content of the assessment, the conditions under which an assessment is administered, the way in which test takers can provide responses to the assessment, the way the assessment is scored, and the way assessment results are reported. For these reasons, concerns of test fairness overlap with those of validity (Chapter 2) and comparability (Chapter 4), and fairness issues shall be considered from the earliest stages of test development and continue throughout the entire life cycle of an assessment.

106. The *Standards for Educational and Psychological Assessment* (AERA, APA and NCME, 2014[1]), which apply to all types of assessments (e.g. individual and group-level, high-stakes and low-stakes, mandatory or voluntary, summative, diagnostic, formative, etc.) describe a fair test as one that "…reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (p. 50). This description of fairness in testing emphasises consistency in score interpretation across all individuals within a tested population. As described in the *Comparability* chapter of this document, such consistency is particularly important in large-scale international assessments such as PISA, because the assessment population comprises great diversity with respect to ethnicity, language, culture, and other personal and social characteristics both within and across countries/economies.

107. In addition to consistency in score interpretation, fairness in testing also involves ensuring broad representation in decisions regarding what is tested, how it is tested, and the proper use of test results. Emphasising the perspectives or values of one culture or personal characteristic over another in one or more aspects of the assessment process may introduce elements of unfairness that may bias assessment results for some groups of test takers (Randall, 2021[55]). The complexity and comprehensiveness of PISA also brings up other fairness considerations, such as how well the group of students tested in a particular country appropriately represents the entire intended student population, as well as the welfare of students who participate in PISA.

108. Fairness in assessment involves consideration of a multitude of issues, and for this reason the AERA et al. (AERA, APA and NCME, 2014[1]) *Standards* provide four "views" of fairness; specifically, fairness (a) in treatment during the testing process, (b) as lack of measurement bias, (c) in access to the construct(s) measured, and (d) as validity of individual test score interpretations for the intended uses. Fairness can also address equity issues through an assessment program that is designed so that all students can demonstrate their proficiencies and capabilities in unobstructed fashion (Elliott, 2016; Randall, 2021).

109. Concerns for fairness in testing also overlap with concerns for validity of test score interpretations; however, validity and fairness are not synonymous (Helms, 2006[80]). Validity refers to the degree to which evidence and theory support the use of test scores for

specific, intended purposes. Such evidence and theory provide important starting points for evaluating and ensuring fairness, but the concept of fairness is broader, and involves critiquing all aspects of the assessment process to ensure all test takers are uninhibited by aspects of the testing process and can fully demonstrate their proficiencies, opinions, and other attributes measured. Similar to validity evidence based on testing consequences (see *Validity* chapter in this volume), fairness involves evaluating the decisions and actions that are made on the basis of test results. Evaluating the consequences of testing should include whether uses and interpretations of test scores may have consequences that are unfair for individuals or groups. For example, if a country administers PISA assessments only to samples of students who are enrolled in highly affluent schools, and curricular changes based on PISA results are only made in those schools, the lack of improvement in the curriculum at other schools not represented in the sampling would be unfair to those students. Relatedly, if PISA is conducted in a particular type of schools or region, and policy changes are made to the entire country/economy based on PISA results, these changes may not be adapted to other types of schools and have unfair consequences for students in these schools.

110. The perspective of fairness in testing as a lack of measurement bias also has conspicuous overlap with validity; for example, when evaluating validity evidence based on internal structure using differential item functioning or other analyses of measurement invariance. Essentially, all validity analyses can be conceptualised as addressing or facilitating fairness. In this document, fairness is treated as a separate concept because it provides an important additional lens for ensuring the validity research, policies, and all other aspects of a testing program promote positive, intended outcomes, instead of negative ones. Data protection and confidentiality are also issues related to fairness in testing. Thus, validity and comparability are important components of fairness, but do not address all issues of fairness in assessment. The following sections identify threats to fairness and suggest guidelines for promoting fairness in all aspects of PISA.

5.2. Threats to fairness

111. Given that in this chapter fairness in assessment is defined as "impartiality and unbiasedness with respect to the content of the assessment, the conditions under which an assessment is administered, the way in which test takers can provide responses to the assessment, the way the assessment is scored, and the way assessment results are reported;" each of these aspects of fairness is described next. The descriptions are oriented by focusing on threats to fairness that should be avoided.

5.2.1. Unfairness with respect to test content

112. Fairness with respect to the content of an assessment refers to the way in which the construct to be measured is defined and how well the items on the assessment represent that construct. Fairness issues that should be considered in defining and describing the constructs measured include ensuring the diversity reflected in the population of students tested is represented by the experts who define what is tested, determine how test scores will be used, develop the test specifications, and set the criteria for item development and evaluation (International Test Commission and Association of Test Publishers, 2022[81]). Fairness considerations are also relevant to defining constructs associated with student demographics and surveys (e.g. binary versus non-binary categories for sex or gender, definitions of a "family," etc.). In international assessments such as PISA, the language in which students are tested is a particularly important fairness issue in that students should be tested in the language in which they are instructed. Such a goal cannot always be realised in large-scale international testing programs, but

matching the language of the assessment to the language in which students are proficient remains an important goal. Even when the match is made, concerns for fairness require ensuring one language version of an assessment is not more difficult, or less precise, than another (ITC, 2019[49]; ITC, 2017[54]; Sireci, 1997[82]).

113. Other important fairness issues related to test content refer to the appropriateness of the specific items and other content on an assessment for the population tested. Although it is never intentional, some test content may overly advantage or disadvantage some cultural or other groups of students, or may offend some types of students. For this reason, PISA participating countries conduct "sensitivity reviews" to screen for inappropriate test content (including reading passages, data tables, and other stimuli) and to ensure cultural diversity is represented on the assessment (OECD, 2019[59]; OECD, 2020, p. Chapter 5[21]). Such concerns for fairness should be made not only to screen out content, but also to explicitly include material that can illustrate different cultures within the tested population are valued (Randall, 2021[55]). While this may be challenging to implement in an international large-scale assessment such as PISA and international empirical examples are currently lacking, efforts should be made in this direction. Another important fairness issue related to test content is the degree to which students taking an assessment have had an *opportunity to learn* what is tested. However, some assessments, like PISA, focus on problem-solving skills that are not explicitly linked to curricula, and so fairness issues with respect to opportunity to learn are less clear. On the one hand, assessing how students solve problems, rather than assessing what they remember from school, is intended to enhance the validity of interpretations by making easier to extrapolate PISA results to real-life situations. On the other hand, it may create a potential threat to fairness if, for example, more privileged students had more opportunities to learn and practice the skills tested in PISA than less privileged students. Ideally, students will have had instruction in the areas tested and have had experience with the item formats used on the test so the material on the assessment will be familiar to them. Although it may always be the case that some students will have more opportunity to learn than others, the fairness issue to be addressed here is whether the material tested may be so unfamiliar that it may trigger anxiety or negative academic self-concept in students taking the assessment. In such cases, testing a particular group of students may not be warranted until after they have been exposed to the material tested or had opportunities to take practice tests – for instance, working more closely with countries and making available exemplar tasks ahead of the assessment.

5.2.2. Unfairness with respect to test administration and scoring

114. Assessments are typically *standardised* to ensure students are tested on the same content, complete the test under the same test administration conditions, and are scored in the same way. Such uniformity in test content, administration, and scoring is designed to *promote fairness* in testing by providing a "level playing field" (i.e. the same rules) for everyone. Standardisation also promotes score comparability. Clearly, if the content of a test differed markedly across students, or if some students had more time or better resources while taking a test, those differences would not be fair. Similarly, if some graders scoring students' work are more harsh or more lenient, the scoring process would be considered unfair. For these reasons, test content, administration, and scoring are carefully scripted (standardised) to guard against any variations from standard protocols that could be considered unfair.

115. Although standardisation is designed to promote fairness in testing, the characteristics of a testing program designed to be fair for everyone can introduce aspects of unfairness for some groups of students. Students with disabilities (SWD) and linguistic minorities are two groups that may often be disadvantaged by overly strict

standardisation conditions, but fairness issues with respect to test administration apply to other types of students, too. For this reason, careful consideration of the differential needs and characteristics of all students within a tested population shall be considered in designing the test and in establishing the test administration conditions and scoring procedures.

116. With respect to SWD, alternate forms of assessments, such as read-aloud (audio) versions for visually impaired students, or sign language interpretation for hearing impaired students, can be provided. Another technique for improving assessment fairness for SWD is by providing accommodations to the assessment in a way that makes it possible, or easier, for SWD to access and interact with the assessment than they can under the normal standardised conditions. Accommodations to educational tests can relate to the way the content is presented (e.g. a Braille or audio version for visually impaired students), the setting in which the test is given (e.g. a private room for students with emotional disorders), the timing associated with test administration (e.g. extended time for students with learning disabilities), or the way in which students respond to test items (a scribe for students with motor limitations). Specific accommodations have also been given to linguistic minorities to address unnecessary linguistic complexity that may hinder their performance on an assessment. Such accommodations include translation of test content, providing bilingual dictionaries, and extended time (Faulkner-Bond and Soland, 2020[83]). Sireci and O'Riordan (2020[84]) noted that some educational assessment programs provide over 30 accommodations for SWD and linguistic minorities. Abedi and Ewers (2013[85]) reviewed these and many other accommodations and classified them with respect to the degree to which they are likely to improve the validity of the assessment results.

117. Abedi and Ewers (2013[85]) and others have pointed out that providing accommodations to standard testing conditions can be an effective way to promote fairness for SWD and linguistic minorities; however, the degree to which the accommodations may change the construct measured (and hence the interpretation of the test score) shall also be considered. For this reason, the possibility an accommodation may change what is measured should be studied to investigate (a) whether tests taken with specific accommodations provide results that are non-comparable to tests taken under standard conditions, and (b) whether the accommodations may also be advantageous to students who do not typically have the opportunity to take the test with the same accommodation (Sireci, Scarpati and Li, 2005[86]; Sireci, 2005[87]). In some cases, legal requirements at national or sub-national level may specify the degree to which SWD can participate in assessments and how they should be accommodated. These regulations should be considered alongside the welfare of the students and uses and interpretation of the assessment results to ensure students who can be tested are tested and are reflected in the assessment results.

118. Given that accommodations to standard testing conditions might lead to non-comparable scores across students who take a test with and without accommodations, it is recommended to use universal test design principles (AERA, APA and NCME, 2014[1]; Thurlow et al., 2016[88]) to minimise aspects of the testing process that may impede the performance of SWD and linguistic minorities. The idea underlying universal test design is to create the test and administration conditions in such a way that accommodations are unnecessary. For example, allowing screen-reading software for all students may remove the need to provide it as a read-aloud accommodation to students who request it, provided that it does not distract or impede the performance of other students in any way. A related idea is to make what is typically considered to be an accommodation available to all students, so the support provided by the accommodation (e.g. bilingual dictionary or screen-reading software) is part of standard test conditions. Technology available for computer-based tests makes provision of such supports for all students possible (e.g. (Crotts-Roohr and Sireci, 2017[89]). Similarly, the concept of "understand-ardisation"

involves understanding the diversity of student characteristics within the population to inform test design, scoring, and administration conditions so that universal test design principles are used to promote fairness for *all* students, not just SWD and linguistic minorities (Sireci, 2020[57]).

5.2.3. Unfairness in reporting

119. Fairness is also important to consider in the reporting of test results. Students' privacy should be protected in reporting results and in the publication of underlying data; risks associated with school re-identification should also be considered. However, less obvious fairness issues in reporting may emerge based on how results are disaggregated and what results are emphasised in published reports. If results are reported for certain subgroups in contexts where such disaggregation is not usual, it may facilitate unfair, racist, xenophobic or classist interpretations. Reports of test results may label some student groups in ways that make them look inferior to others, or make interpretations that sound correct from a statistical perspective, but are derogatory. For example, a statement like "on average, 9-year olds in country X have learned about the same amount as 7-year olds in country Y" extends a difference in average scale scores on a particular assessment to what students have learned. Such a statement is unfounded because no test can capture the totality of what any student, or group of students, has learned. Another example is using PISA typical grade gain across countries without considering the gain tends to be larger in high-income countries compared to middle-income countries (Avvisati and Givord, 2021[28]). Concerns for fairness in reporting of test results will emphasise factors that are likely to affect group differences such as differential curriculum emphasis, opportunity to learn, differences in the representativeness of student samples to national populations, subpopulations who are in and out of formal schooling, and other factors (see *Comparability* chapter in this volume).

5.3. Guidelines for ensuring fairness in PISA

120. Given the potential threats to fairness in assessment, several guidelines for promoting fairness can be offered. These suggestions focus on the areas of data privacy, test development (content), test administration, scoring, and reporting of results.

5.3.1. Use culturally sustaining assessment principles in decision making, construct definition and all aspects of test development

121. Culturally sustaining assessment principles require involving a diversity of stakeholders in the decision-making processes of an assessment, and consideration of the needs of the most marginalised students in the tested population (Lyons, Johnson and Hinds, 2021[90]; Randall, 2021[55]; Randall et al., 2022[56]). All PISA participating countries and economies should be provided with opportunities to express their views, and their voices should be considered in the decision-making process. When defining the constructs measured in an assessment, different perspectives should be invited and debated before forming consensus on what is to be tested and how it should be tested. PISA works closely with representatives from participating countries in the test development process. In addition to working with the institutional settings in which students within a country are educated, advocates for historically marginalised groups within these countries should also be included in the discussion to the greatest extent feasible, particularly those groups that are included in reporting of the results. In addition to providing input on construct definition and test design, this community of partners should also advise on test delivery, student sampling, administration, scoring, and score

reporting. Admittedly, some of these issues are technical and will require representatives from these groups with experience in these areas.

In practice: PISA attaches high importance to ensuring that all PISA participating countries and economies can express their views and that their voices are taken into account throughout the assessment cycle. At the design stage, participating countries and economies have the opportunity to propose potential innovative domains through a consultation. Furthermore, the innovative domain for each cycle is selected following a participatory process where PGB members can vote for proposals during a written consultation, and the proposal with the highest support across countries is proposed for approval at the PGB meetings. The selected proposal is further refined by consulting subject-matter experts and practitioners across as many countries/economies as possible (Standards 6.1). The process for the definition of research and development projects in PISA is similarly participative (Standards 6.2). Besides, the PGB is to be consulted and to approve any decision regarding the development of new optional test and questionnaire module (Standards 6.1).

Turning to the development of data collection instruments, participating countries and economies shall have the opportunity to review drafts of the frameworks and contribute to their improvement, and the PGB has to approved the final versions. Besides, the standards expressly mention that the expert group guiding the development of new frameworks or innovations in the frameworks should be nationally and culturally diverse, and national experts from participating countries/economies shall be consulted multiple time during framework development (Standards 7.1). Similarly, national experts from all participating countries/economies are encouraged to contribute to cognitive test item development early in the process by submitting stimulus material and associated items. This is facilitated through the use of an online item-submission tool and the organisation of item-writing workshops with national experts. Developed items are then required to undergo a review by national experts, and their views on item improvement shall be respected by item developers (Standards 7.2). National experts are also consulted for the development of questionnaire items (Standards 7.3).

Finally, at the analysis and reporting stage, experts from participating countries/economies are invited to review the analysis plan and report outlines, and to review drafts of the reports. Their comments should be incorporated in the reports and justifications should be provided when comments cannot be addressed (Standards 9.1).

5.3.2. The universality of the constructs measured on an assessment should be established across all participating countries and jurisdictions tested

122. Many constructs in educational and psychological testing stem from dominant culture orientations that may not generalise to other cultures (Matsumoto and van de Vijver, 2011[91]). Governments should consider the universality of the construct(s) measured in their decision to participate in PISA, and in its optional components. Curricula across various countries and cultures can be studied for the generality and appropriateness for the tested population. Even if PISA does not measure curriculum per se, this type of analysis can shed light on the familiarity, generality and relevance of the constructs measured on PISA across all countries and cultures. Both quantitative research (e.g. differential item functioning) and qualitative research based on expert review (e.g. local educational psychologists familiar with the assessed population) and a review of the literature are recommended to ensure the relevance and equivalence of the construct

across countries and jurisdictions, and ideally, across cultures and other relevant subgroups within countries/jurisdictions.

In practice: The standards applying to the development of frameworks specify that measurement of new constructs shall be carefully evaluated with international comparability in mind, through an extensive literature review, inputs from participating countries/economies, and review of material from previous PISA cycles (Standards 7.1).

5.3.3. Ensure standardised testing conditions are appropriately flexible to allow for diversity in the way students access test content, receive appropriate support during the assessment, and provide their responses to test items

123. Standardised test conditions should be the same for everyone, but what is specified as "the same" should be sufficiently flexible to accommodate the diversity of student needs and reduce obstacles that may prevent students from demonstrating their true proficiencies. For example, alternate language versions of a test allow for students who operate using different languages to participate in a testing program (suggestions follow for fair translation/adaptation of test material). Flexible standardised conditions should also allow for students to take the assessments using familiar devices and with supports they typically use in the classroom. For instance, updating the paper-based form to make it as similar to the computer-based version as possible. Research to understand the diversity of students' needs and how they learn should be used to inform what become "standard" test administration conditions. These conditions should allow sufficient time for students to answer all test questions, without feeling rushed. On the other hand, the time students are asked to spend on assessments should also be minimised to reduce disruption to instruction and other activities.

In practice: PISA currently offers limited flexibility in testing conditions, with the exception of the paper-based test, and some accommodations for students with SEN (developed in guideline 5.3.4). In order to improve on this aspect, the standards on the functionality of PISA digital tools specify that the performance of PISA digital tools should be regularly evaluated to understand concerns from various actors and anticipate requirements for the forthcoming cycles, acknowledging that countries/economies participating in PISA have diverse resources and requirements, and while most countries/economies have started using the computer-based mode for administering the assessment versus the paper-based version, countries/economies can require different logistical modes and devices to complete PISA-related tasks. It is thus recommended that the PISA digital tools progressively develop the capability of delivering the assessment on various devices, including tablets (Standards 6.4).

Besides, the standards on test and questionnaire design state that the assessment should be constructed such that most students in each participating country/economy can finish the test within the limited time, without exceeding 60 minutes (i.e. the test should not be speeded) (Standards 6.6).

5.3.4. Where standardised test conditions, even flexible ones, interact with student characteristics such as disabilities or limited language proficiency, provide appropriate accommodations to students to eliminate any barriers related to the testing situation that impede their performance

124. Such accommodations can include changing the ways in which test content is presented, clarifying instructions, changing the way students provide their responses, increasing the time allotted for testing, providing a separate testing location, or allowing the use of technology tools students use in the classroom when receiving instruction. Students should not be excluded from assessments based solely on their disability, unless it is clear that participation in the assessment would be detrimental to the student in some way or if the severity of a student's disability prevents participation (e.g. severe cognitive impairment).

In practice: The standards on PISA digital tools indicate that PISA should offer accommodations to increase the accessibility of the assessment for test takers with special education needs (SEN) (Standards 6.4). Currently, PISA allows a reduced range of accommodations which would enable countries/economies to limit exclusions (small group, one-on-one, certain type of special equipment like lighting devices, directions read aloud in sign language, and auditory amplification). In addition, for students with an official SEN classification and who would be excluded from taking the regular assessment with or without allowable PISA accommodations, an adapted version of the PISA test, the Une Heure (UH) form, is offered. The UH option consists in the administration of a shorter computer-based test (two 30-minutes sections) and questionnaire (15 minutes), offered as a separate session for the National Centre. Students taking the UH option can benefit from extended time (maximum of 100 minutes in total for the cognitive test, and 25 minutes in total for the questionnaire). This is designed to reduce exclusions. However, national teams have expressed the concern that neither the UH nor the restricted range of allowed accommodations adequately cater to the needs of many students with SEN. The PISA Research, Development and Innovation programme is therefore exploring the feasibility of including a wider range of assistive tools in the platform (such as text-to-speech, zoom or contrast adjustments) and test breaks.

5.3.5. Provide translated/adapted versions of tests to facilitate participation in the assessment.

125. It is crucial that translations/adaptations are of sufficient quality to support valid inferences related to score comparability (see the *Comparability* and *Validity* chapters in this volume). Quality translation/adaptation should follow best practices in test adaptation. In addition, the comparability of different language versions of the assessments should be evaluated. The International Test Commission (2019[49]) and other resources (e.g. (Hambleton, Merenda and Spielberger, 2005[92])) provide comprehensive guidance on how to properly adapt test material for use across languages and how to evaluate different language versions of assessments. To ensure fairness across different language variations of an assessment, accepted translation/adaptation designs should be used (e.g. teams of bilingual translators whose work is evaluated by independent translators local to the assessed population) and both qualitative and statistical analyses should be conducted to evaluate the quality and comparability of the translated and target versions of the assessment.

In practice: These guidelines are reflected in the standards for translations and adaptations (Standards 7.4), covering best practices that should be followed in terms of processes (such as carrying a translatability assessment before translation even starts) and outcomes (e.g. items shall be appropriately adapted to the local context), and qualitative analyses. Quantitative analyses are covered in the standards regarding the scoring and analyses (Standards 8.1, 8.2, 8.3, 8.4, 8.5).

5.3.6. Provide clear and sufficient information to students and their families in advance of the assessment to reduce potential feelings of test anxiety and to mitigate risks to students' academic self-concept

126. The degree to which students have had the opportunity to learn the material to be tested should be determined before administering tests to students. Testing students on material they have not had the opportunity to learn can be stressful and can lead to reduced feelings of academic self-concept. It should be communicated to students that their performance on PISA has no consequences for them and that their responses are completely confidential. Students should also be familiarised with the format of the assessment (e.g. item formats, testing interface) and informed about the expected level of difficulty at which the assessment was built. A fair test includes content familiar to students. In some cases, it may be helpful to document the degree to which students have been exposed to the content tested when interpreting test results. Students should also be informed about how their work on the test will be evaluated: this is particularly important in technology-enhanced tests using process data, where students might not be aware that they are being evaluated also on their response processes.

In practice: The standards on developing cognitive test items mention that task contexts should be familiar, appropriate and accessible to all PISA students. Furthermore, the standards stress the importance of carrying validation studies including analyses of event/process data (e.g. think aloud processes in cognitive labs, computer log-files) in order to detect potential differences in familiarity with the content and engagement across countries/economies and groups of students (Standards 7.2). Students' exposure to the test content can be evaluated when relevant through the contextual questionnaires, notably in the innovative domain to contextualise results (Standards 6.1). Finally, the standards require that any use of process data in scoring is restricted to cases where students are informed that they will be scored on their response processes and not only on their final answers (Standards 7.2).

5.3.7. Consider differences in the ways in which students respond to assessment tasks when developing scoring rubrics for constructed-response items

127. Constructed-response items require students to write or type their answers or create some other response to an assessment task. Students may differ with respect to response style, abbreviations, slang, colloquial terms, and other aspects of their responses. These differences in student response characteristics should be considered in the scoring of their responses to determine how and whether they affect points that are awarded or taken off. For example, an African American English colloquialism like "aks" used in the place of "ask" may be considered appropriate in a scoring rubric that emphasises fairness, but may be considered an inappropriate or misspelled word in a more strict rubric that emphasises specific word forms.

In practice: The standards on developing cognitive test items and scoring rubrics specify that the scoring guides should present general principles for scoring, including how to treat spelling and grammar mistakes (Standards 7.2).

5.3.8. Consider the degree to which students have experience in interacting with digital technology when developing the test design and delivery

128. Fairness in testing requires the demands associated with a testing program be appropriate with respect to the tested construct, and that unnecessary complexity be eliminated. Unnecessary complexity can be due to digital interface and other factors such as overly complex text associated with instructions or test questions. When tests are presented in a digital format, unfamiliarity with the technology should not impede students' performance on the test. It is therefore important to develop test items that have an intuitive design that can be understood easily by students, without the need for introductions, practice items, or technical explanations. When necessary, providing practice tests and other materials to familiarise students with how to interact with the assessment can reduce the construct-irrelevant effects of familiarity with digital devices on test performance, and so tutorials and practice tests are recommended in situations where tests will be administered in ways that are unfamiliar to students.

In practice: The standards on developing cognitive test items specify PISA computer-based items need to be designed so as to minimise the impact of familiarity with digital tools and devices on student performance, which can be done by including tutorials when relevant. (Standards 7.2).

5.3.9. Conduct research to evaluate the fairness of test results with respect to consistency and appropriateness of tests and items across subgroups of students in the tested population

129. As mentioned in the *Comparability* chapter, there are many statistical and qualitative methods that can be used to evaluate the comparability of constructs, assessments, and items across groups defined by language, culture, sex, and other important characteristics (Berman, Haertel and Pellegrino, 2020[53]; ITC, 2019[49]; Sireci, Han and Wells, 2008[93]; van de Vijver, Jude and Kuger, 2019[22]; Winter, 2010[94]). These techniques can be used to evaluate and document the degree to which the psychometric characteristics of PISA assessments generalise across important subgroups of students defined by personal characteristics (e.g. sex, ethnicity, country) or testing condition (e.g. paper-based, computer-based, tablet-based), and so facilitate fairness in the use and interpretation of PISA results.

In practice: These analyses are covered in the Technical reports (Standards 9.3).

5.3.10. Ensure all data and results reported adhere to data protection rights to which all participants in studies are entitled

130. Where PISA operations require the processing of personal data, it is important that these data are processed in a transparent manner and only for the legitimate purposes specified in the PISA program. Personal data should be retained by international contractors only for as long as necessary to fulfil the goals of the Program (depending on

national legislation, National Centres may retain the data for longer periods). Furthermore, it is crucial that international contractors process personal data securely. Finally, participants should be provided with clear information on data protection and their rights to access, rectify or erase their data as facilitated by countries and contractors.

In practice: The PISA sampling contractor shall obey regulations on data confidentiality that apply in participating countries (Standards 6.6). The PISA digital tools similarly need to adhere to strict data protection regulations (Standards 6.4). Besides, the public use data files shall exclude variables which might be used to identify or disclose sensitive information about schools or participants (which concerns responses and identifiers, responses to open-ended questions, but also some type of process data which could be used as biometric data); and participating countries and economies are given the opportunity to suppress variables from their national data in the public use files when it is deemed to present risks to anonymity (Standards 9.2).

Part II. Quality Standards

131. This section includes standards related to the general design and coordination of the programme (Chapter 6), the development of data collection instruments (Chapter 7), the scoring and analysis of PISA data (Chapter 8) and reporting of PISA data (Chapter 9).

132. Individual standards are presented after an introductory text that explains the rationale for developing standards in this domain, and indicates the operations that are covered by the standard. In many cases, the standards themselves are coupled with a clarifying note, possibly providing examples of how the standards might be applied. A concluding section describes the quality control indicators and criteria that should be considered to verify that the standards have been met. These quality assurance indicators refer to both quality evidence on processes and quality evidence that is produced through empirical analysis.

Chapter 6. Design and coordination

6.1. Standards for defining new test domains and questionnaire content

6.1.1. Rationale

133. PISA's primary objective is to evaluate the capabilities of 15-year-old students in applying their knowledge and skills they need to fully participate and thrive in their communities and the global environment. Consequently, it is crucial that PISA remains aligned with the changing landscape of education and the demands of modern societies and economies. While all PISA cycles have followed a similar organisation since 2000, each new cycle is expected to respond to emerging trends and policy questions, incorporating evidence on emerging competencies that are essential for lifelong learning and active participation in today's society. The domain covered in PISA are regularly expanded through the innovative domains and the optional modules. In addition, PISA regularly updates the definition and the focus of items of the reading, mathematics and science assessments to reflect changes in their nature and applications. Finally, the questionnaires are also revised at each cycle in order to accurately depict the current learning environment and provide timely responses to emerging policy questions.

134. PISA has included an innovative domain assessment in every cycle since 2012. The innovative domain assessments aim to provide PISA countries/economies with a more comprehensive outlook on their students' readiness for life, by focusing on transversal or 21st century competencies. These assessments drive innovation in PISA by extending its focus beyond the core literacies of reading, mathematics and science and by fostering technological and methodological innovations in the design of the items, analysis and reporting. In prior cycles, PISA has integrated the following innovative domains: Creative Problem Solving in 2012, Collaborative Problem Solving in 2015, Global Competence in 2018, Creative Thinking in 2022. The innovative domain for the 2025 cycle will be Learning in the Digital World. The innovative domains are an integral part of PISA, so all PISA countries are expected to participate in these assessments at no additional costs, unless they express their motivation not to.

135. Beyond the innovative domains which change every cycle, it is necessary that the core literacies assessed every cycle in PISA are continuously reviewed to take into account changes in how these competences are used for work and for life, innovations in how these competences are taught, as well as technical and methodological developments in large-scale assessments. Since its beginning in 2000, PISA has had three core domains: reading, mathematics, and science. These domains assess how students apply core disciplinary knowledge to solve real-life problems. While these domains are assessed on a regular basis to monitor change over time, the expected learning outcomes, the technology and the real-life contexts in which these competences are applied necessarily change over time. There is thus a trade-off that PISA has to manage given this double goal of minimising changes to obtain reliable trends and keeping the constructs and tests up to date. The PISA long term strategy for 2024-2033 thus stresses that PISA needs to update the assessment framework and instruments of each domain every 12 years, ensuring that the share of trend measures leaves sufficient space to introduce updates and new development in these domains.

136. Optional cognitive test modules represent another instrument PISA has to respond to new demands for information on educational systems. Differently from the innovative domains, participation in an optional module requires additional costs, so not all countries participating in PISA typically take these assessments. The purpose of the optional modules is to provide more flexibility to participating countries in the content of the assessment,

while preserving the key goal of comparing results internationally. Since 2012, PISA has included an optional module on financial literacy, that aims at assessing the engagement of young people with financial issues and their skills to address the challenges posed by an evolving financial landscape. Starting from 2025, PISA will include a new optional foreign language assessment. In the 2025 cycle, the foreign language assessment will focus on students' capacity to listen, read and speak in English. The coverage of languages and skills may progressively widen in future cycles. Further optional modules might be developed in the future. However, only one optional cognitive test will be offered in each cycle, in order to limit the demand on students' time and to ensure that a sufficient number of countries participate in the assessment.

137. Finally, innovations in PISA also take place at the level of the contextual questionnaires. Beginning with the questionnaire framework used for the PISA 2009 assessment, questionnaire content in PISA was explicitly linked to different levels of the education system: the student level, level of instruction in the classroom, school level, and system level. The 2012 questionnaire framework conceptualises contextual variables for understanding education systems 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 different levels of the education system. 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 constructs important for understanding differences in achievement that are not tied to a specific subject area.

138. In addition to the core student- and school-level questionnaires, PISA makes available different optional questionnaires: a teacher questionnaire, an ICT questionnaire, a parent questionnaire and a well-being questionnaire. In order to have a sufficient number of participating countries and so increase the value of international comparisons, only a subset of optional questionnaires are offered in each cycle. To further enhance PISA's capacity to tackle emerging policy questions, the PISA 2024-2030 long-term strategy proposes to develop a module of innovative context questionnaire items, as a new international option, in order to trial new measures of student background, attitude and behaviour that would be included in future main student questionnaires. These innovations could cover new areas or adopt new formats, such as experience sampling or social networks analysis.

139. The standards in this section present the quality criteria that should guide the definition of new test domains and questionnaire content in PISA, covering aspects related to the choice of innovative domains, building consensus, and documentation.

6.1.2. Standards for selecting the innovative domains

Standard 6.1.1. Each PISA cycle shall assess students' competences in an innovative domain to address emerging policy concerns and yield new insights into educational systems' effectiveness.

140. Note: The innovative domains selected should have significance and address evidence gaps in emerging education policy areas. They should provide relevant and internationally comparable information on students' learning and problem-solving competences, including their social and emotional facets. This is primarily achieved

through the development of complex performance tasks in the cognitive tests, but also the development of new questionnaire modules for each innovative domain.

Standard 6.1.2. Proposals for new innovative domains shall be built on insights from previous cycles to ensure assessment robustness and relevance.

141. Note: Previously included innovative domains in PISA encompass Creative Problem Solving (2012), Collaborative Problem Solving (2015), Global Competence (2018), Creative Thinking (2022), and Learning in the Digital World (2025). Proposals on innovative domains should leverage insights from previous cycles, including lessons on challenges related to measuring certain constructs in PISA.

Standard 6.1.3. Proposals for innovative assessments shall demonstrate the existence of a solid research basis in the domain.

142. Note: Given the limited time and resource to develop an assessment for a new domains, and the high quality requirements in PISA, it is important that each proposed domain can rely on solid theoretical literature on the target construct and its development for children and adolescents, including its malleability through formal education. It is also recommended that any proposed domain has already been tested in published research, and the existing evidence demonstrates that it can be applied for international comparisons and for large-scale assessment.

Standard 6.1.4. Proposed innovative domain assessments shall consider practical implications, including the time available for assessment and administrative feasibility.

143. Note: Feasibility is critical for the innovative domains. Proposed innovations should be practical, manageable, and should seamlessly integrate with PISA operations and regulatory framework without compromising assessment integrity or imposing excessive burdens to national teams.

Standard 6.1.5. Participating countries and economies shall have the opportunity to propose potential innovative domains.

144. Note: The OECD Secretariat should consult PISA participating countries and economies to collect propositions of future innovative domains. This consultation should take place at least 5 years prior to the cycle in question in order to leave sufficient time to framework and test development. Proposals should be well defined in terms of relevance, competencies targeted, methodologies and feasibility.

Standard 6.1.6. The innovative domain for each cycle shall be selected following a participatory process involving participating countries and economies.

145. Note: The PISA Governing Board (PGB) is responsible for selecting the target competencies of the innovative domains from a pool of proposals prepared by the OECD Secretariat, including proposals submitted by countries/economies (Standard 6.1.7). The proposal receiving the most support from participants, both during written consultation and discussions at PGB meetings, should be proposed for approval by the PGB. This selection process should take place at least 5 years before the implementation of the assessment in order to leave sufficient time to framework and test development. Transparent and inclusive stakeholder consultation processes should be emphasised to

refine and enhance the proposed innovative domains, incorporating substantial inputs from subject-matter experts and practitioners across as many participating countries as possible.

6.1.3. Standards for defining new content for the core domains and questionnaires, and optional modules

Standard 6.1.7. A plan for the inclusion of optional test and questionnaire modules shall be developed and regularly updated. The plan shall span multiple cycles and be approved by the PGB.

146. Note: In order to have a sufficient number of countries/economies opting in the optional modules, the number of tests and questionnaire modules offered in each cycle should be limited. Yet, it is important to generate data for all modules at regular intervals. Therefore, the strategy on optional modules should define which ones will be offered in which cycle to balance the need for limited options with the need for trend measures. For instance, in current practice, PISA only includes one optional test module per cycle.

Standard 6.1.8. New questionnaire modules should regularly be proposed to the PGB for possible inclusion in PISA.

147. Note: These modules could be proposed in response to a timely need, such as the Global crisis module in PISA 2022 following the COVID-19 pandemic, or as a way to trial new measures as suggested in the PISA long-term strategy (module of innovative context questionnaire items).

Standard 6.1.9. When a domain is the major domain in a cycle, its framework shall be thoroughly reviewed by international subject-matter experts and diverse stakeholders in order to identify areas of improvement and innovations for the next cycle.

148. Note: Areas of improvement can be identified for instance by means of a qualitative survey on the purpose of education in the core domain and changes in society and new skills needed. This stakeholder consultation should result in a strategic direction and vision detailing the possible areas of development and rationale for proposing them, and what it means for the competencies assessed.

Standard 6.1.10. Clear and meticulous documentation shall be maintained for each proposed innovation's rationale, development process, and implications, with feedback from all stakeholders serving as a comprehensive reference for all.

149. Note: Meticulous documentation of each proposed innovation's rationale, development process, and implications should be maintained. This documentation serves as a comprehensive reference point for all stakeholders, offering clear insights into the underpinning logic, developmental trajectory, and potential impacts of the innovations, thereby contributing to a transparent, informed, and cohesive understanding of the enhancements made to the PISA framework.

Standard 6.1.11. Proposed innovations in the form of new content in the core domains or questionnaires, new optional modules, and new optional questionnaires shall be discussed with and approved by the PISA Governing Board.

150. Note: Countries and economies should have the opportunity to provide feedback on the strategic direction and vision proposed by the expert group, in written form or during a PISA Governing Board meeting.

6.1.4. Quality assurance

- The OECD Secretariat produces an innovation development plan which should describe the process and timelines for gathering ideas on innovative domains from participating countries/economies and for engaging participating countries/economies in selecting the innovative domain for each cycle.
- The OECD Secretariat produces an innovative domains proposal which should present:
  - Details on emerging policy concerns and changes in the global context addressed by the proposed domains;
  - Description of how the proposed domains complement PISA's core literacies of reading, mathematics, and science;
  - Discussion on how the proposed domain includes recommendations from implementing innovations in previous PISA cycles;
  - Summary of existing literature on the proposed constructs and their development for children and adolescents including a discussion on cross-cultural validity and suitability for students with special education needs;
  - Discussion on the practical feasibility of including the proposed domains in PISA.
- The OECD Secretariat prepares a document detailing out the process and timelines to get the major domain reviewed by international subject matter experts and incorporating the recommended changes before each cycle.
- The OECD Secretariat produces a rationale note proposing the development of new questionnaire modules and new optional cognitive tests modules, detailing:
  - Details on emerging policy concerns and changes in the global context addressed by the proposed modules;
  - Summary of existing literature on the proposed constructs including a discussion on cross-cultural validity;
  - Discussion on the practical feasibility of including the proposed modules in PISA.
- At the end of each PISA cycle, the OECD Secretariat, with the support of relevant experts or contractors, produces final reports on the development, implementation, analysis and reporting of all the innovations (innovative domains, innovations in core domains and questionnaires, and new questionnaire and test modules). This report should document major challenges and learnings at each phase and include recommendations for implementing innovations in future PISA cycles. These will be used by the PGB to discuss future innovations.

6.2. Standards on continuous research and development

6.2.1. Rationale

151. As one of the main international benchmarks for assessing and comparing the outcomes of educational systems around the world, it is crucial for PISA to stay relevant and to continuously improve its quality by incorporating research advances. PISA aims at measuring 15-year-old students' ability to use their knowledge and skills to meet real-life challenges, enabling them to fully participate and thrive in their communities and the world. Therefore, it is important that the assessment is in sync with the changing landscape of education and with the demands of modern societies and economies. This means extending the range of competencies to be measured in the innovative domains, and regularly updating how PISA core literacies are defined so as to reflect changes in their nature and applications. Similarly, the questionnaires need to be continuously revised so that they describe the current context of learning and provide timely responses to emerging policy questions. In addition, research in psychometrics and learning sciences, as well as progress in assessment methodologies and technologies, are generating new opportunities for PISA to improve its measures of learning outcomes and contexts.

152. The PISA Research, Development and Innovation (RDI) programme was established in 2019 as a permanent and coordinated programme to respond to these challenges. The programme aims to introduce conceptual, methodological and operational advances in PISA in order to improve the relevance, precision, efficiency and validity of the assessment and to extend the range of competences that are assessed in PISA. The RDI programme is one of the main tools for implementing PISA's long term strategy. It carries research and development projects that follow a two-year cycle and that are thus independent from PISA cycles. RDI projects have focused on improving quality assurance mechanisms, improving the assessment's methodology, technology and operation, or researching and developing the type of items, scoring methods and analytical models for the main and new domains. For instance, past projects have developed measures of test engagement, researched improved models for adaptive testing, or improving accessibility of the assessment for students with special education needs.

153. These standards define the scope of the RDI Programme, as well as the processes to prioritise the research outputs and review the results.

6.2.2. Standards

Standard 6.2.1. The RDI programme shall include projects that aim to improve PISA's methodology and operational processes, and/or to inform the scope of future assessments.

154. Note: RDI project proposals should fall under one of these two following categories: 1) methodological and operational improvements, 2) innovation for future assessments. There should be a balance in the number of projects selected at each RDI cycle for each category, in order to ensure that both aspects of RDI are covered each cycle.

Standard 6.2.2. RDI projects shall be aligned with PISA's long-term strategy.

155. Note: Project proposals should clearly state how they are contributing to PISA's long-term strategy as defined in the current strategy document.

Standard 6.2.3. The selection process for new RDI projects shall take place at least six months prior to the closure of ongoing projects.

156. Note: It is important to select projects sufficiently in advance in order to leave enough time to the development of sound implementation plans.

Standard 6.2.4. PISA Governing Board members shall have an opportunity to propose RDI projects.

157. Note: Members should be invited to share their proposals with the OECD Secretariat by filling in a standardised form detailing the relevance and importance of the proposed project for PISA, its alignment with the PISA long-term strategy, methodology and possible outputs. It should also mention which RDI track the project belongs to (methodological improvements or innovation).

Standard 6.2.5. RDI projects shall be selected following a participatory process involving participating countries and economies.

158. Note: Countries/economies should be invited to select new RDI projects from a set of proposals according to their relevance and importance, separately for each category of projects (methodological improvements and innovation for future assessments). The proposals receiving the highest rating in each category are recommended for selection, and the final choice of the projects is determined by consensus during a PGB meeting.

Standard 6.2.6. The objectives and outcomes of the RDI programme shall be reviewed by the PGB every five years.

159. Note: The projects' results should be reviewed regularly in order to consolidate the programme's relevance and orient project proposals for future cycles.

6.2.3. Quality assurance

- The OECD Secretariat produces a clear and detailed implementation plan for the selected projects detailing projects' alignment with the PISA long term strategy, expected outputs, methodology and timeline.
- The OECD Secretariat produces regular progress reports which transparently inform countries/economies on the developments of the ongoing projects.
- The OECD Secretariat produces every five years a report reviewing the objectives and outcomes of the RDI programme as well as an overall evaluation of the RDI projects for future steps. This report is to be reviewed by countries/economies and discussed in the PISA Governing Board meeting.

6.3. Standards on target population, sampling design and sampling operations

6.3.1. Rationale

160. PISA aims to measure achievement of 15-year-old students in the core domains of mathematics, reading and science and to collect information on the contexts of learning in order to provide accurate assessments of the quality and equity of education systems. PISA also aims to enable countries and economies to compare results among them and over

time. Assessing all individuals belonging to the target populations would be too costly and labour-intensive in most countries, therefore PISA relies on representative student samples. To meet the intended purpose, it is crucial that PISA samples allow the estimation of population parameters that are accurate and sufficiently close to their true values. This means that the sampled data are a valid basis for inference to the target population within given precision levels. It needs to do so without relying on strong assumptions (embedded in models) about the distribution of the variables of interest in the population and their intercorrelations. Survey practitioners often use the expressions "design-unbiasedness" and "sampling precision" to summarise those characteristics of a sampling design (Dumais and Gough, 2012[95]). If a sample fails to meet these conditions, the main principles of PISA regarding validity, reliability, and comparability are violated (see also Chapters 3, 4 and 5).

161. A crucial first step in any survey is defining the target population. This seems to be a simple task at first glance, but often it turns out that details are significant, which is especially true for cross-national comparative studies (Meinck, 2020[96]). PISA defines its target population as all 15-year-old students attending educational institutions in grades 7 and higher (OECD, 2020[21]). To allow for a cross-country comparison of results, it is important to detail what are "educational institutions" (hereafter referred to as "schools"), which national grades constitute grade 7, and how are eligible students identified so that they are 15 years old indeed during the testing period ("15-year-olds" itself is operationally defined in PISA to mean "between 15 years and 3 months old and 16 years and 2 months old"). If other survey populations are added, exact definitions are needed, too. As a general principle, PISA's core target population is 15-year-old students, therefore the survey sampling design is optimised for this target population.

162. While PISA aims for covering all 15-year-old students in grade 7 and higher, it is often reasonable to allow for some exclusions of eligible students from the survey for practical and budgetary reasons. It is crucial that valid criteria and acceptable thresholds for such exclusions are agreed-upon, controlled for, and documented (Martin, Rust and Adams, 1999[51]). The standard (arbitrary) threshold for acceptable exclusion rates in PISA is five percent. Further, guidelines for maximum exclusion of specific groups should be provided. Higher rates should be annotated when presenting survey results to highlight potential bias risks.

163. Design-unbiased samples are generally probabilistic random samples (Kish, 1965[97]; Lohr, 1999[98]). There is a variety of methods available on how to select such random samples. For surveys in the field of education, samples applying specific features such as multiple-stage sampling, stratification, and sampling with unequal selection probabilities are considered as practical and efficient. The optimal design is determined by the PISA core research questions, aiming for a description of the performance and contexts of learning of a specific age cohort. However, it also varies depending on the characteristics of the different education systems, national research interests, and availability of information for eligible schools to inform stratification. Given its complex nature, PISA sample designs are substantially less efficient than a simple random sample. A uniform sample size requirement, particularly if expressed only in terms of the final sampling unit (students), would not guarantee the achievement of a target precision, and thus a minimum sample size requirement of schools and students within schools needs to be established. Specific conditions such as rotated test and questionnaire designs, optional add-on surveys, requirements regarding sufficient sample sizes for subgroup reporting (for example, for "adjudicated" subnational entities), high between school or high population variance may require larger school or student sample sizes (Meinck and Vandenplas, 2021[99]).

164. This section presents the standards that should be followed to define PISA target populations, to sample schools for the Main Survey and to sample students within schools, as well as standards regarding sampling for the Field Trial. Since data collected from the Field Trial is not used to estimate population parameters, less strict rules apply regarding sampling for this step of the survey. It should be noted that the following standards do not cover the standards and procedures related with sampling to be followed by National Project Managers, as these are already detailed in the PISA Technical Standards. It is recommended that this section is read in conjunction with the Sampling Technical Standards for the current cycle.

6.3.2. Standards on defining target populations

Standard 6.3.1. Exact definitions of the target populations shall be provided that can be applied in all national contexts in order to ensure comparable estimates of population parameters across countries/economies and over time for 15-year-old students attending educational institutions in grade 7 or higher.

165. Note: Due to cross-national contextual differences, the sampling contractor shall ensure through inquiry, consultation, and clarification with the national team that the nationally defined target populations are comparable across countries and over time. The definition shall enable national teams to correctly identify, and list all schools accommodating the targeted students, and to identify and list all eligible students within sampled and participating schools. Any deviations in the nationally defined target population or changes in between cycles shall be fully documented and in accordance with the international exclusion guidelines and standards, and, if necessary, be considered for reporting and interpretation of results. To identify the correct target grades to be included, a standardised system such as the International Standard Classification of Education (ISCED) should be used.

Standard 6.3.2. Full coverage of the PISA core target population shall be achieved in all participating countries/economies. Requests for exclusions need to be discussed, agreed-upon, and documented. Exclusions exceeding 5% should be well justified and documented.

166. Note: Countries shall receive guidelines on acceptable exclusion criteria and thresholds, both exclusions of whole schools and exclusions of students within sampled schools. Exclusion rates should be computed as the sum of (1) school-level exclusions based on population statistics and (2) within-school exclusions which are estimated using weighted sample data. Guidelines may further cover maximum exclusion rates of specific school types, and students within schools. The fundamental principle should be maximum inclusion of all eligible students in the defined target population.

Standard 6.3.3. Exact definitions shall be provided for additional target populations decided by the PISA Governing Board. Sampling for these additional populations shall not jeopardise the effectiveness and efficiency of the core target population sample.

167. Note: The sampling design shall always be optimised for the core target population. The sampling contractor shall advise the PISA principal investigators on the feasibility to achieve a sufficiently precise, representative, and comparable sample for potential additional populations given this main principle. PISA has added for example

the possibility to survey teachers within PISA-sampled schools in 2018, however, the definition for this target population has changed in 2021. Consequences regarding cross-country and cross-cycle comparability shall be stated and discussed if such changes occur.

6.3.3. Standards on sampling schools for the Main Survey

Standard 6.3.4. School samples shall be selected for each participating country/economy by the sampling contractor.

168. Note: This allows effective monitoring of adherence to all standards described below.

Standard 6.3.5. The PISA sampling design shall, for all participating countries/economies, support the unbiased estimation of population parameters. A probabilistic sampling design, which defines a known and non-zero ex-ante probability of sampling to each final sampling unit, shall be applied to achieve this goal.

169. Note: Characteristics of national education systems and specific research interests of national teams should be considered for identifying an optimal sampling design for each participating country. Besides collecting respective information via sampling forms, consultations with national teams shall be conducted to discuss sampling design aspects.

Standard 6.3.6. The PISA student sample should be selected in two stages (schools at first and students within schools at second stage). Adding further sampling stages should be avoided.

170. Note: In the context of PISA, a multiple-stage sampling usually implies the selection of schools first, and second the selection of a group of students within participating schools. The two-stage nature of such design leads to a significant loss in sampling efficiency due to a clustering effect; that is, students within the same school are more alike than students in different schools. However, this approach allows to survey students and collect context information from principals (and potentially teachers) in a cost-efficient way. Adding another sampling stage usually causes another significant loss in sampling precision and should therefore be applied only in well justified exceptions, for example, if a high-quality school sampling frame is not available for the whole educational system. Exact definitions need to be provided for national teams for the sampling units at each stage to allow the correct identification and listing of these units.

Standard 6.3.7. PISA schools should be selected with probabilities proportional to their enrolment of PISA-eligible students. Exceptions shall apply to schools where all students are asked to participate.

171. Note: In PISA, schools are usually selected with probabilities proportional to their size (commonly known as PPS sampling). This method leads under specific circumstances to approximately self-weighted samples of students, reducing the uncertainty in the estimation introduced by the variation of sampling weights. These circumstances are: (i) a subsample of a fixed number of students within schools is selected, (ii) sample allocation is proportional in all strata, (iii) nonresponse is evenly distributed across strata and within schools. PPS may not be feasible for very large schools, i.e. those with a measure of size (MOS) larger than the sampling interval, and very small schools,

i.e. those with a MOS smaller than the within-school student sample size. Very large schools should either all be selected, or, allowing for some replacements, separated in a "large-school stratum" and undersampled (i.e. in smaller proportions compared to a strict proportional sample allocation) using systematic simple random sampling (SRS). This approach can also help avoid these schools being sampled for every PISA cycle, and other sample surveys, a situation that may imply a decrease in willingness to participate. Very small schools shall also be selected with SRS, and may be undersampled to avoid having too many schools in the sample with small student sample yield – a situation that significantly increases the costs of the PISA data collection with limited benefits for sample quality. If a country cannot provide the expected number of PISA-eligible students for all schools, another size measure can be used instead.

Standard 6.3.8. Stratification should be applied as possible and reasonable, without compromising comparability. The number of schools selected per explicit stratum shall be eight schools at the minimum.

172. Note: Stratification will increase the efficiency of a sampling design if the stratification variables are correlated with the outcome variable. In PISA, (statistically) efficient stratification variables are those predicting achievement. For example, average school performance in national tests may be an excellent stratification variable for school sampling. Stratification variables can also be used to improve cost-efficiency (often through geographical stratification variables). Explicit stratification implies grouping the population of schools according to some observable characteristics and then selecting independent samples for each group (stratum). This allows a full control on the allocation of sample sizes across strata and each stratum is seen as a population of interest by itself. Selecting a minimum of eight schools per explicit stratum will ensure reliable estimation of sampling variance in each stratum. Selecting fewer schools bears a high risk of large nonresponse adjustment factors related with an increase in sampling variance, and also imprecise estimates of sampling variance in affected strata. It can further lead to situations where, due to nonresponse or unreliable frame data, only one or no school participates in that stratum, making methodologically sound variance estimation impossible. Implicit stratification refers to sorting a sampling frame by some observed characteristic, allowing, in conjunction with systematic sampling, an approximately proportional sample allocation across strata. The sampling contractor shall identify efficient stratification variables by analysing past PISA cycles or similar survey data, or by consulting with National Centres. It should be explicitly noted that a change in the sampling design between cycles, including the use of stratification variables, does not jeopardise sample comparability across time or across countries.

Standard 6.3.9. Minimum sample size requirements for schools and students shall be established.

173. Note: These requirements should aim at reaching a precision level for students population estimates equivalent to a predefined effective sample size. All sampling design features affect the achieved precision levels. Usually, stratification increases precision, while clustering and varying weights due to unequal selection probabilities and uneven nonresponse patterns decrease precision. Moreover, population variance also influences sampling efficiency. It is impossible to predict the exact effects of the design features for each country (i.e. the design effect), which is why it is reasonable to determine minimum sample sizes for both, schools and students for each PISA cycle. These requirements may vary by cycle because they depend on other study design features and envisaged outcomes. For example, a bridge study or plans for in-depth subgroup reporting may call for higher sample sizes. The sampling contractor shall accordingly advise the PISA Governing Board regarding reasonable minimum sample sizes. Adjustments for expected nonresponse at all sampling stages shall be considered. Further, investigating sampling designs and their effects in earlier cycles or similar studies can help determining optimal sample sizes for each country. For example, countries where large population variance was detected in previous cycles should consider sample sizes above the minimum. Depending on whether the variance occurs rather at the between - or within - school level, larger school samples, or larger within-school samples should be considered.

Standard 6.3.10. For each sampled school up to two replacement schools shall be assigned.

174. Note: Not all schools will be willing or able to participate in PISA. The sampling contractor shall assign during the sampling process up to two replacement schools that are similar to the sampled school regarding their allocation to strata, and their size. Using replacement schools bears increased bias risks and shall be limited to a minimum. Respective rules are presented in Standards 8.2 on computing survey weights. Schools that turn out to be ineligible for PISA, for example because they do not accommodate any (eligible) 15-year old student, shall not be replaced. The sampling contractor shall ensure this rule is followed by national teams.

Standard 6.3.11. The sampling contractor shall act client-oriented, where clients are national teams.

175. Note: The sampling contractor shall (i) educate national teams to fulfil their roles and responsibilities in an informed manner, (ii) provide all information needed in a format that can be understood by non-statisticians (e.g. the sampling manual should refrain from highly technical language), (iii) discuss and explain the advantages and disadvantages of specific design features, and (iv) inquire and accommodate specific requests by countries, as long as they do not jeopardise the PISA core study. Some countries may wish to use the PISA data collection to inform national research questions beyond the international scope. For example, a country may wish to compare public and private schools. The sampling contractor can accommodate this request by appropriate stratification schemes and oversampling. Further, overlap control procedures can be applied to reduce the likelihood that a school is part of two surveys running in parallel. Or, a country may wish to retrieve estimates for class-based samples, a request that can be embedded into PISA without biasing the regular PISA sample.

Standard 6.3.12. The sampling contractor shall obey regulations on data confidentiality that apply in the participating countries.

176. Note: The sampling contractor shall ensure that all data is by default exchanged and stored in a secure digital environment. Regulations regarding the protection of personal data such as, for example, the General Data Protection Regulation (GDPR) of the EU shall be enforced and considered in all procedures. For example, countries shall be requested to refrain from sending any personal data with their school frames, such as names, telephone numbers or addresses of principals etc. Further, the contractor shall ensure to receive only pseudonymised information on within-school sampling, i.e. all personal data stored in within-school sampling software shall be automatically disguised or removed before it is uploaded to the contractor.

6.3.4. Standards on within-school sampling

Standard 6.3.13. All sampling routines in the software for within-school sampling shall follow sound sampling methodology, and selections shall be tracked.

177. Note: All features of complex sampling designs can also occur at this sampling step. For example, explicit and implicit stratification may be used to select proportional or disproportional numbers of students with specific features, such as gender, age, belonging to tracks or minorities, etc. As an international option, instead of direct student sampling, classes may be selected causing a second-order clustering effect, and another sampling stage (e.g. students from within classes) may be added. Direct student sampling works best though in most countries as 15-year olds are usually distributed across different grades. The same methodological principles apply as for school sampling. As the national teams operate the within-school sampling software, all algorithms shall be determined, programmed, and checked prior to software release. Sampling probabilities and within-school exclusions shall be tracked, so that the sampling contractor can derive accurate weights and nonresponse adjustments using this information.

Standard 6.3.14. Selected students shall not be replaced.

178. Note: Replacing individual students bears too large a bias risk (for instance, there is a risk that low ability students are replaced with higher ability students) and would complicate survey operations unduly.

6.3.5. Standards on sampling for the Field trial

Standard 6.3.15. The Field trial should enable countries and the contractor to assess the feasibility of all sampling-related procedures, identify issues of concern and develop strategies on how to address them in the Main survey.

179. Note: A central objective of the Field trial is to test all operations and make sound plans on how to scale them up for the Main survey. A key issue in achieving representative samples is ensuring high response rates of schools and students. The Field trial design should enable national teams to identify issues with operations and response rates, allowing to devise strategies to reduce the likelihood of such problems during the Main survey. All operations required for the probabilistic sampling of schools and students in the Main survey should be trialed during the field trial. However, the use of non-probabilistic sampling methods for the selection of actual field-trial respondents is allowed, in particular through purposive sampling or quota sampling.

Standard 6.3.16. The Field trial sample shall assure a good coverage of the variation in the target population and be large enough for accurate psychometric assessment of (new) items, including for different languages.

180. Note: The second sampling-related objective of the Field trial is testing (new) items. The Field trial sample should reflect the variation of responses in the target population, so that the fit between items and respondents can be investigated. The Field trial sample shall further be large enough to ensure a minimum number of responses to each item, also per test language. The minimum may vary depending on the character of the item (completely new, new for many countries, new to PISA but tested in other studies, trend, questionnaire vs. test item, etc.). The contractor for psychometric analysis shall determine the minimum number of responses needed. The sample size is then determined by (i) this minimum number, (ii) consideration of a rotated test or questionnaire design, where not every respondent is administered all items, (iii) expected nonparticipation of schools and nonresponse of individuals, (iv) budgetary constraints. It may for example not feasible to achieve enough responses on school questionnaire items for proper country-level psychometric analysis.

6.3.6. Quality assurance

- The responsible contractor will produce an international sampling plan for the given PISA cycle which presents evidence to ensure that all the standards on sampling are met. The plan should be presented to the PISA Technical Advisory Group (TAG). The TAG should evaluate the plan according to the above standards. The TAG could suggest changes to the design if one or more of the standards are not entirely met.
- The responsible contractor provides a sampling manual that details each step of the sampling process relevant to national teams and specifies their tasks and responsibilities. Sampling forms are used to formally document information needed for the sampling.
- The responsible contractor prepares sampling plans for each country and presents them to the principal investigators for approval. An external expert (sampling referee) or the TAG is called upon for adjudicating contested cases.
- The responsible contractor checks the quality of the sampling frames provided by national teams rigorously for consistency and plausibility. Checking routines include the search for duplicates, and comparison of frame information with information on sampling forms, estimated population sizes from earlier cycles or other studies, and officially available data, such as enrolment figures and birth statistics from reliable online sources (Meinck, 2020).
- The responsible contractor implements internal quality control by applying the four-eye principle, that is, one team member selects the sample, another team member checks the sample using both common sense and a dedicated checklist.
- The responsible contractor produces meticulous documentation on sample selection in order to allow full reproduction. The documentation is sent to the country and made available, upon request, to the principal investigators at OECD, PISA TAG, and the sampling referee. The documentation is stored for a minimum of ten years after the release of the respective PISA cycle.
- The responsible contractor presents all methods applied for sampling of schools and students, including the core characteristics of the national samples in the PISA Technical Report.

6.4. Standards on the functionality of PISA digital tools

6.4.1. Rationale

181. PISA is a large-scale collaborative effort. Several actors are involved in designing and successfully delivering PISA in each cycle. These actors include the PISA Governing Board, National Project Managers, Expert Groups, the Technical Advisory Group, the OECD Secretariat, school coordinators, and the PISA International Contractors. Digital tools in PISA enable these different actors to collaborate, develop and implement PISA. These tools support item and test authoring, item review and approval, adaptation and translation, sampling of students, test delivery and monitoring. In addition to the actual development and delivery, PISA digital tools also allow actors spread across several geographies and time zones to collaborate and communicate effectively to complete various PISA-related tasks in a time-effective and quality-conscious manner.

182. Since PISA 2015, the main test delivery mode in schools has been by computer and all new test items in PISA 2015, 2018 and 2022 were developed for computer delivery only. For computer-based assessment participants, all questionnaires, except the optional parent questionnaire, were delivered on a computer. Following the move to computer-based assessments, PISA has started incorporating technology-enhanced items (TEIs) in the assessment. These items require that the PISA digital tools are programmed using modern, state-of-the-art technology for websites and online applications to integrate interactive items with multi-media features easily. It is also important that they incorporate a system to efficiently collect, parse, store and export all the process and response data as requested by assessment designers.

183. Trend items are needed to make PISA results directly comparable across cycles (see Reliability chapter). These items should be easily transferrable should there be a change in the platform used for PISA. Therefore, the digital tools need to include item authoring tools that facilitate the creation of QTI (Question and Test Interoperability specification)-compliant items that can be easily exported to third-party systems.

184. PISA's computer-based implementation also allows for delivering tests adaptively where items or blocks of items (multistage design) are personalised for each test taker. The PISA digital tools should have features that support the multi-stage adaptive design currently implemented in PISA and more complex adaptive algorithms that might be integrated in the future.

185. In addition to the test design (linear fixed-length vs adaptive), the coding of responses also has a bearing on the error and reliability of an assessment. Technological advancements have facilitated the emergence of novel and improved functionalities for automated scoring of assessments, including scoring of selected responses, technology-assisted human scoring of constructed responses, and fully automated scoring of constructed-response assessments (International Test Commission and Association of Test Publishers, 2022[81]). The PISA digital tools should facilitate automated and human coding of responses in online or offline modes for both computer-based as well as for paper-based implementation. The system should provide a feature for supervisors to implement monitoring checks to improve the coding quality and reliability of test scores.

186. Another essential aspect related to PISA's computer-based assessment is the testing environment. Both online and offline options are available for administering computer-based tests. These different environments necessitate thoroughly evaluating possible disruptions during test-taking (International Test Commission and Association of Test Publishers, 2022[81]). To ensure comparability and fairness, online and offline delivery systems within the PISA digital tools should provide the same testing experience for

students. The system should also offer the same testing experience across devices and web browsers used in PISA.

187. A crucial advantage of computer-based tests is the expanded scope for enhancing the accessibility of test items and interfaces. Accessibility during the administration of tests can be categorised into content access and interaction, response production, and interface navigation. The first two categories pertain to specific test items and encompass the phases test takers go through when engaging with a test item. The final category applies more broadly to the test delivery system. It focuses on how test takers utilise the various functionalities embedded in the delivery platform to engage with individual items and the entire test (International Test Commission and Association of Test Publishers, 2022[81]). The PISA digital tools should have a feature to offer accommodations for students with a variety of special needs to widen accessibility and inclusivity.

188. Lastly, efforts in the design of the tools should be made to ensure seamless collaboration between the various PISA actors – the OECD Secretariat, National Centres, and contractors. This section presents the standards which should guide the development and deployment of digital tools for PISA. These standards include digital tools for item and test development, delivery, and communication among different actors across the PISA implementation cycle. In some specific instances, the word "platform" or "system" is used instead of "tools" but the overall meaning remains the same.

6.4.2. Standards on general functionalities of the digital tools

Standard 6.4.1. The PISA digital tools shall offer functionalities required to develop and deliver PISA and provide a central point for communication across PISA actors.

189. Note: PISA's end-to-end test cycle involves several stages and actors (including the OECD Secretariat, the National Centres, the contractors). Efficient and effective communication is also crucial for the successful delivery of PISA. For instance, the platform for PISA 2025 includes a core group of integrated purpose-specific applications that together cover authoring, translation, sampling, test delivery, scoring, communications and helpdesk.

Standard 6.4.2. The performance of PISA digital tools should be regularly evaluated to understand concerns from various actors and anticipate requirements for the forthcoming cycles.

190. Note: The countries/economies participating in PISA have diverse resources and requirements. Most countries/economies have started using the computer-based mode for administering the assessment. Countries/economies use different logistical modes and devices to complete PISA-related tasks. Feedback from the participating countries/economies provides crucial inputs to improve PISA digital tools. Systematic and detailed feedback through surveys or other methods should be gathered to understand issues with the current version of digital tools and anticipate requirements for the forthcoming cycles. Documentation on the current capabilities and limitations should be maintained, updated and shared with all PISA actors. In particular, the group responsible for framework and item development should be engaged in the discussion on the capabilities and limitations of functionalities related to item development.

Standard 6.4.3. The PISA digital tools shall be accessible to all the actors.

191. Note: Accessibility for the PISA digital tools refers to the ease by which all users can access them, regardless of their background and physical and intellectual capabilities. The PISA digital tools should at least be WCAG 2.2 compliant and aim to deliver the AA standard on each WCAG criterion. In addition, the PISA digital tools used to deliver PISA should support multiple language, including those with a right-to-left display.

Standard 6.4.4. Training material explaining the different functionalities of the PISA digital tools should be developed and made available to PISA actors.

192. Note: Training of national teams is critical for the successful development and implementation of PISA. To ensure quality, extensive training material on using PISA digital tools should be developed and made available to PISA actors. The training materials can be developed as bite-sized modules explaining all critical functionalities offered within the digital tools.

Standard 6.4.5. The PISA digital tools shall be continuously monitored for any security and malfunction issues.

193. Note: An initial security analysis and regular security scans as well as end-to-end testing should be implemented to evaluate potential security threats and malfunctions. Necessary vulnerability remediation, application of security patches and other upgrades to the PISA digital tools should be documented and communicated to the OECD regularly.

Standard 6.4.6. The PISA digital tools shall include a prompt and ongoing maintenance program that responds to technical troubleshooting by users.

194. Note: PISA digital tools should offer sufficient and timely technical support to users. The support system can be included within the digital tool, where various actors can communicate with technical support staff by providing details on their issues and challenges. Along with the technical support team, a comprehensive list of frequently asked questions by topic should be developed, maintained, updated and made available to all users of PISA digital tools.

6.4.3. Standards on assessment development functionalities

Standard 6.4.7. The PISA digital tools shall allow authoring a range of item types and response modes using pre-built customisable templates and components. The items created in the PISA digital tools should also be available in a format that can be migrated to other systems/tools/platforms.

195. Note: Several of the PISA items include simple interaction formats that are consistent with the item types included in existing standards, such as the Question and Test Interoperability (QTI) specifications (for example, single- and multiple-selection multiple-choice items, fixed as well as unlimited response text entry items, and items with drag and drop function, drop-down menus, sliders, and radio buttons). The availability of an authoring tool for these standardised items can reduce the costs of item authoring and enable easier collaborations across different organisations that might be responsible for item development. The more complex items that include custom interactions and interactive multimedia should also be designed following industry standards for the design of web-based content (e.g. use of HTML 5, JavaScript, ECMAScript 6, etc.).

The procedure for authoring items in the testing platform should be documented to enable item development even by third parties.

Standard 6.4.8. The PISA digital tools should include a simple-to-use item editing interface for collaboration across actors writing and reviewing cognitive tests and questionnaire items.

196. Note: The editing interface should allow National Centres, the OECD and other contractors to review and modify the items for the different domains directly into the platform. A robust workflow should be defined to allow item authors worldwide to collaborate, comment, edit and publish items without the risk of version control issues. Responsibilities for different elements of the item creation process should be assigned to different user roles, with notifications sent to user groups at defined points in the workflow. This can include multiple authoring and review loops to ensure a robust quality model is applied throughout the item writing process. A granular permissions model should be deployed so that users only get to see content they are authorised for, thus maintaining the integrity of the items. The digital tools used for item development should limit as little as possible the organisational discretion of the National Centres.

Standard 6.4.9. The PISA digital tools shall integrate translation and adaptation functionalities with the cognitive test and questionnaire item authoring system.

197. Note: Translation and adaptation are critical when developing national versions of the PISA assessment instruments. The PISA digital tools should allow seamless integration with the translation service, allowing in-country translators to translate and preview translations whilst maintaining an authoritative, single master version of an assessment item. The PISA digital tools should be able to deliver all approved master instruments, ready for in-country translations, in an industry-standard format like XLIFF. The translation file should be at the right level of granularity with all the essential tags for accurate translations. The translation system should allow a side-by-side view of the source language and translated items, ensuring the translator can thoroughly compare the layout. In addition, the questionnaire authoring tool should enable the management of approved national adaptations and translation: national versions of questionnaire items that have been adapted, translated and verified in previous cycles should be controlled centrally. The digital tools used for translation should limit as little as possible the organisational discretion of the National Centres.

6.4.4. Standards on assessment implementation functionalities

Standard 6.4.10. The PISA digital tools shall be capable of delivering the assessment in ways that are adapted to countries' IT infrastructure. The testing experience shall be the same across delivery modes.

198. Note: Infrastructure diversity and the availability of IT support vary widely across different PISA countries/economies and schools within a single country/economy. The platform should provide a secure, easy-to-use online delivery model, supplemented by cost-effective offline delivery modes (does not require internet connectivity to deliver tests) or hybrid modes (e.g. local networks) for those countries/economies that need it. In the online mode, the digital tool should be developed to have low bandwidth requirements. The countries opting for online testing should undertake extensive testing before the main study to ensure that they have the necessary internet infrastructure in every participating school. Whether offline or online, the test delivery system should not

negatively impact the testing experience. Students shall have the same test experience regardless of the test delivery mode.

Standard 6.4.11. The PISA digital tools should progressively develop the capability of delivering the assessment on various devices and browsers.

199. Note: The PISA 2025 assessment is available for computers using Windows 7, MacOS 10.11 or later, Linux/Unix and Chrome OS. Keeping up with technological advances, developing and trialling the PISA digital tools across various devices (including tablets), operating systems and web browsers should be planned and implemented continuously. The selection of accepted devices, operating systems and web browsers should be based on feedback from participating countries/economies.

Standard 6.4.12. The PISA digital tools shall include a software to implement sampling that shall be capable of handling varying sample sizes, special samples, oversampling, different number of stages and stratification variables across participating countries/economies.

200. Note: The software monitors the sampling frames, list of schools and list of students so that weighting and/or non-response bias analysis can be calculated. The digital tools used for sampling should limit as little as possible the organisational discretion of the National Centres.

Standard 6.4.13. The PISA digital tools shall be capable of delivering the test in adaptive formats.

201. Note: Adaptive testing is an algorithm that personalises how an assessment is delivered to each test taker. It is coded into a software platform to select items and score test takers. The algorithm proceeds in a loop until the test is complete. In a multistage adaptive test, item blocks are authored into testlets with items of comparable difficulty in parallel sets. The resultant score from a testlet block determines how the student is routed to the next stage in conjunction with a probability layer. The PISA digital tools should be capable of delivering the test in adaptive formats following cut scores, probability values or other rules set by the test authors when creating the test. Keeping up with advances in adaptive assessment practices, developing and trialling the PISA digital tools to accommodate innovative adaptive algorithms should be planned and implemented continuously.

Standard 6.4.14. The PISA digital tools should offer Universal Design elements and accommodations to increase the accessibility of the assessment for test takers with special education needs.

202. Note: The OECD aims to widen access to PISA for students with disabilities and other special education needs. The PISA digital tools should offer accessibility options that can be implemented at the test or student level. Some desired accessibility tools that should be included in the PISA platform are zooming in and out to see the content better, the ability to change the background colour to improve contrast, reading aloud the assessment items, allowing extra time, allowing the test to be paused, use of screen readers supported by alt tags for images, and allowing the candidate to record their verbal response, in replacement of physically entering the answer where this might not be possible. Based on needs identified through accessibility surveys, the PISA digital tools

should develop and trial functionalities to include other accommodations for presenting the items and recording student responses in accessible ways.

Standard 6.4.15. The PISA digital tools should be able to create interactive dashboards to offer real-time monitoring updates to actors at different levels.

203. Note: The PISA digital tools should have a feature to track the delivery of the test and various levels. At the school/class level, invigilators should be notified if a candidate loses connection or is stuck in the system for some reason. At the country/economy level, automated and customisable dashboards should be created within the PISA digital tools to monitor if the tests are being implemented as scheduled and will help keep a check on students who are being excluded.

Standard 6.4.16. The PISA digital tools shall support automatic and human coding of responses.

204. Note: The PISA digital tools should include an in-built coding module. The system should support auto-coding of selected-response items and closed constructed-response items. As artificial intelligence and machine learning methods in assessments evolve, the system might be further developed to include AI-based coding of constructed response items by connecting the PISA digital tools to external tools without releasing access to student responses. A system for human coding should also be offered. The human coding functionality should have a feature where coders and their supervisors can interact in case of specific questions about particular responses. The coding system should include backreading function so coding supervisors can monitor inter-rater reliability in real-time and take corrective actions (e.g. retraining) as soon as possible.

Standard 6.4.17. The PISA digital tools shall be capable of recording and parsing process and response data.

205. Note: Construct-relevant process data in the context of a PISA test include total time spent in each item; time of each interaction with the user interface; interactions with menu items, navigation buttons and blocks; instances of tests and trials in simulation tasks; state of workspace each time a given action is performed; instances of support material being accessed; etc. Continuous efforts should be made to collect, parse, and export process data following industry standards to facilitate secondary research using these important sources of evidence.

Standard 6.4.18. The PISA digital tools shall integrate mechanisms to prevent the loss of student test data.

206. Note: In a fully online assessment, data losses can arise because of problems in the reliability of the internet connection. For test delivery scenarios that use offline modes like local client servers, the risk of data loss is if the USB sticks or local server device is lost or damaged before the data is uploaded to the database. The PISA digital tools should be developed to include functionalities to minimise the risk of such data loss.

Standard 6.4.19. The PISA digital tools shall adhere to strict data protection regulations mandated across different countries/economies.

207. Note: PISA test security, data protection and confidentiality are essential to the programme's success. The PISA digital tools should comply with applicable legal and regulatory obligations across different countries/economies, including the General Data Protection Regulation (GDPR). The digital platform should have physical and virtual data protection following industry standards and be capable of deploying secure transfer protocols where data can be encrypted in transit. Technical support engineers should have access to data on a role basis only, managed through user login credentials, access privileges and regular review processes.

Standard 6.4.20. A School Readiness Tool (SRT) shall be developed and implemented as an integral part of the PISA digital platform to check that all the required computer systems for implementing PISA are set up and ready for use.

208. Note: A School Readiness Tool (SRT) should be designed and implemented, which National Centres can use to assess the capabilities of sampled schools to implement the computer-based assessment. The school readiness tool should clone the actual PISA student delivery system. The SRT should help check browser compatibility, internet speed, audio functions, hardware issues, etc., and create a report to alert National Centres when any requirement is unmet.

6.4.5. Standards on communication and management functionalities

Standard 6.4.21. The PISA digital tools should include a communication platform to foster greater collaboration and transparency among PISA actors using a single-sign-on functionality.

209. Note: The single sign-on system should be developed at multiple levels so that users can access only the material and the information they have the right to access.

Standard 6.4.22. The PISA digital tools should allow information sharing through several modes.

210. Note: The communication platform should include a range of tools that enable seamless communication and information sharing between the contractors, national teams and the OECD Secretariat. This should include sharing information the overall calendar and timelines listing tasks across the current test cycle, regular updates on progress across different tasks for the PISA contractors and each participating country and what different actors can expect shortly. The system should also enable members to interact with each other. For instance, communication could be managed through a community and personal messaging system; through email and newsletters; or through audio/video meetings.

Standard 6.4.23. The PISA digital tools shall include features for file sharing and archiving.

211. Note: The PISA communication platform should include a file management feature allowing National Centre teams and contractors to share files securely. The platform should enable users to upload or download multiple and large files. Files should be organised hierarchically and easily searchable within the platform. The platform should have a functionality to archive documents and material.

6.4.6. Quality assurance

- The responsible contractor produces a Digital Tools Development Document. The document should include:
    - Description of functionalities available within the platform(s) and how tools and systems that are available separately will be integrated for improved efficiencies;
    - Analysis of technical issues faced by various PISA actors in the previously completed test cycles, possible solutions to mitigate issues and timelines when missing features can be developed.
- The responsible contractor produces documentation and training material on all critical functionalities of the digital tools and shares a plan on how training can be integrated with the PISA digital tools.
- The responsible contractor provides documentation of the interface, features, templates and customisations available in the item authoring tool and describes the features available to accommodate students with special needs.
- The responsible contractor documents how complex items are integrated into the PISA platform and programme them using modern web-based tools to facilitate their use in other platforms.
- The responsible contractor integrates the translation system within the item authoring tool or describes how the translation system will work.
- The responsible contractor develops a system for online coding with details on the coding process.
- The responsible contractor should provide a document indicating the delivery modes, devices and browsers supported by the PISA digital tools.
- The contractor should develop a system to mitigate the risk of data loss for each delivery mode and identify and report on any data loss.

6.5. Standards on the use of alternative forms

6.5.1. Rationale

212. Alternative forms of PISA allow experimenting with major modifications of the key components of PISA in order to address policy questions or assess populations for which the main PISA is inadequate. Alternative forms should nevertheless maintain essential features of the main PISA study, such as its emphasis on international comparability and on assessing students' readiness for life.

213. An alternative form of PISA should thus be comparable to the main PISA in its salient components. Typically, alternative forms include a cognitive test, focus on the same age cohort (specifically, students at the end of compulsory education), and include some of the main contextual indicators at the student level, such as an indicator of students' socioeconomic characteristics. Major modifications can occur in the context, in the technology or in the coverage of the study. For instance, unique elements that could be developed for alternative forms might include the enhancement of cognitive item banks (e.g. including more items in the lower end of the ability distribution, or adapting items to address students' specific needs); the assessment of new cognitive domains; the enhancement of questionnaires for unique contexts, such as vocational programmes or the learning environment of out-of-school youth; or the modification of forms

and technology in data collection, for instance through household survey, collection of data from other informants, or through the use of visual data collection or GPS coordinates.

214. For example, PISA extended its instruments to different population of students and to diverse group of participating countries with the PISA for Development (PISA-D), an alternative form of PISA that targeted not only students in school but also about 15-year-olds out of school. To better assess this extended target population, PISA-D offered enriched cognitive and contextual assessments to describe the context of learning and the skills in reading and mathematics of fifteen-year-olds, in and out of school, in eight low- and middle-income countries. As an alternative form to PISA, the PISA-D experimented alterations and integrations to three aspects of the PISA study: the target population and sample design (14 to 16-year olds, out of school, interviewed and tested in a household survey); the cognitive tests (including more easier items in the PISA scales); and the contextual assessment (with new items and new informants on life and resources in lower income families and schools). PISA-D allowed countries and economies to compare themselves with PISA countries using the cognitive scores and trend indicators from the contextual assessment.

215. Another example is the existing Une-Heure (UH) option for students with special education needs, which involves the administration of a shorter test and questionnaire. The UH form was developed to reduce exclusions in PISA and to help participating countries and economies achieve the PISA coverage standard. It is offered as a separate session with logistical and cost implications for the National Centres. The cognitive assessment includes a 60-minute single testing form comprised of 20 minutes of trend material from each of the three core domains of Reading, Mathematics and Science. Students taking the UH instrument take an approximately 20-minute UH version of the Student Questionnaire, which includes a subset of items from the regular 35-minute Student Questionnaire. The UH option can therefore be seen as an alternative form of PISA as it provides a modified version of the PISA test and questionnaire designed to cater the needs of students with special education needs. Future versions of the UH option might include the possibility of having assistive tools (such as text-to-speech, zoom or contrast adjustments), as explored by the PISA Research, Development and Innovation programme on improving accessibility.

216. Alternative forms of PISA shall meet the core principles of validity, reliability, comparability, and fairness as described for the main PISA study (see Chapters 2 to 5). These principles need to be established and met at four different levels: during the process of establishment of an alternative form; at the onset and during the study design; when constructing and administering the instruments for data collection (both the cognitive and the contextual assessments); and at the stage of analysis and reporting. Notably, particular attention should be paid to the terminology employed for the communication around alternative forms of PISA, especially concerning usage and applications of terms such as "PISA-linked" or “PISA-equivalent”, for which comparability with the main PISA shall be validated. The communication and reporting of results should not create confusion when comparisons are established and discussed.

217. This section presents the standards that should guide the processes to establishing, designing, analysing, and reporting results from an alternative form of PISA.

6.5.2. Standards for processes to establishing and designing an alternative form

Standard 6.5.1. The need for the development and use of an alternative form of PISA for a particular context shall be clearly documented and justified.

218. Note: Before any alternative is explored, proposed, or developed, a clear case should be made for why the main PISA is not fit for purpose for the context in question. The case for the need of an alternative form can be raised by the OECD Secretariat or the PISA Governing Board. Evidence for the inadequateness of the main PISA and the need of an alternative form can consist, for instance, in research findings exposing gaps in the PISA main offering, or policy questions unanswered by the main PISA.

Standard 6.5.2. The objectives and the conditions requiring the development of an alternative form shall be sufficiently different to justify the investment in study, development, and implementation of the alternative form.

219. Note: A review committee, joined by experts nominated by PGB countries, should evaluate the case for the investment in the alternative form. The case should be based on the assessment of evidence (Standard 6.6.1) and of capacity and requirements to develop and implement the new form.

Standard 6.5.3. The development of an alternative form shall be approved by the PGB.

Standard 6.5.4. The creation of an alternative form of PISA should involve the development of a cognitive or a contextual framework that are specific for the alternative form of PISA.

6.5.3. Standards for study design

Standard 6.5.5. The alternative study shall be designed to provide answers to specific questions that are not addressed by the main PISA.

Standard 6.5.6. The study design shall clearly identify the data-collection instruments or parts of those instruments that need to be adapted or developed.

220. Note: The study design should outline the scope of the alterations to the main PISA components, and be prescriptive and precise about what elements will be affected and how. An assessment of the impacts/effects of the modifications is also necessary.

Standard 6.5.7. Participating countries or economies shall be consulted and involved in the development of the study design through regular communication, timely provision of technical documentation, training, and collection of feedback.

221. Note: Involvement of participating countries and economies includes technical guidance, support, and training for the correct implementation of the study. In addition, feedback from participating countries and economies should be collected and used to improve the following iterations of the alternative form.

6.5.4. Standards for the creation and implementation of new data collection instruments

Standard 6.5.8. When developing new test or questionnaires, the alternative form of PISA shall comply with the same standards that apply to the main PISA.

222. Note: See Chapter 7 on the development of data collection instruments.

Standard 6.5.9. If the alternative form involves minority groups, vulnerable groups, or understudied groups, the data collection instruments shall be fully accessible to these groups.

223. Note: For instance, the formulation of the items should ensure that these groups can fully understand and respond easily and truthfully.

Standard 6.5.10. When tests and questionnaires include questions or items from the main PISA contextual assessment tools, these questions or items shall be evaluated on their adequateness for inclusion in the alternative test and questionnaires.

Standard 6.5.11. When questionnaires include modified versions of main PISA trend items, these modifications shall be justified and documented.

Standard 6.5.12. When tests and questionnaires use main PISA items and trend indicators within instruments that are administered differently, a field trial shall assess the behaviour of the items and trend indicators under new conditions of administration.

224. Note: Examples of different administration that would warrant piloting include, for instance, administration on tablets, or as part of interviews with a trained interviewer. After a field trial, a document should describe and discuss any alterations in distributions and characteristics of the items and trend indicators.

6.5.5. Standards for analysis and reporting

Standard 6.5.13. Clear documentation shall guide readers and analysts on whether and how to establish comparisons between countries participating in the alternative form of PISA and countries participating in main PISA. The documentation shall explicitly address matters of linkage and comparability with the main PISA.

225. Note: The documentation accompanying the development and the implementation of an alternative form of PISA should consist of a framework document that describes its purposes and main goals, as well as the theoretical approach adopted for its design. A technical document should describe the implementation of the instrument of the alternative form and their comparability with main PISA instruments.

Standard 6.5.14. Reports of the results from the main data collection shall clearly communicate the differences of the alternative form with the main PISA, the rationale for establishing it, and its implementation.

Standard 6.5.15. The release of data and assessment material for the alternative forms shall comply with the standards of the main PISA (see Standards 9.2)

6.5.6. Quality assurance

- The OECD Secretariat produces a clear and detailed implementation plan for the alternative form detailing its alignment with the main PISA, and with the expected outputs and planned modifications. The implementation plan should include details on the rationale for establishing the alternative form, the methodology and timeline.

- The responsible party produces regular progress reports which transparently inform on the developments of the ongoing project. The recipient of the reports are the parties or the countries and economies participating in the study, as well as funding partners or agencies.
- Following the pilot of the alternative form, the responsible party produces a report that presents and discusses how the study design effectively met the objectives of the alternative form. For instance, the report would explain how effectively the alternative form included the target vulnerable or understudied groups.
- The responsible party produces every cycle of data collection a review of the objectives and outcomes of the alternative forms. The report discusses if and how to integrate elements of the alternative forms in the main PISA. The report should also be reviewed by countries/economies and discussed in the PISA Governing Board meeting.

6.6. Standards on test and questionnaire design

6.6.1. Rationale

226. Designing tests and questionnaires means assembling items from a pool into test and questionnaire forms, thereby determining how, in what order, and for how long students and other questionnaire respondents will interact with questions, tasks, and other types of items that PISA uses to measure the characteristics of students, schools, and education systems. A first important goal of the PISA design is to provide participating countries/economies with estimates of population distributions of student proficiency for 1) the three core domains of Reading, Mathematics, and Science, 2) the subscores for the major domain, and 3) an innovative domain. A second important goal consists in linking student learning outcomes across domains, across cycles, and with students' background and attitudes towards learning across countries/economies. Furthermore, the PISA design should also accommodate the international options that are decided for each cycle, which requires a certain flexibility (e.g. varying participation in Financial Literacy, Foreign Language and additional questionnaires). Choices on what test and questionnaire design to use in PISA have impact on reliability, comparability, and fairness. The standards in this section help to ensure the expected quality level for the above goals and to maintain this level across participating countries/economies and PISA cycles. However, the different goals may sometimes compete in terms of design requirements, and it is therefore important that the different pieces of the test and questionnaire designs not only work well in isolation, but also fit well together.

227. Test design choices, such as time limits or whether the test includes adaptive paths, have important implications for the size and the composition of the item pool; in turn, the amount of available items sets limits on how a construct can be assessed. The size of the item pool denotes the number of available items. The composition of the item pool describes characteristics in terms of the assessment framework, item format, content specification, psychometric properties, and timing information. Questionnaire design choices, such as time limits or whether questionnaire forms include advanced routing options or rotation of items, also have important implications for development, administration, and analyses.

228. Until PISA 2015, the PISA tests were based on a multiple matrix sampling design in which each student was assessed on multiple domains and was administered a subset of items from the total item pool for each assessed domain (Messick, Beaton and Lord, 1983[100]; Rutkowski et al., 2014[101]). Starting from 2018, PISA has introduced a multistage

adaptive testing (MSAT) design for the major domain in which test difficulty is adapted to student proficiency during the administration. This was done to improve measurement precision, particularly in the tails of the proficiency distribution.

229. In general, test design needs to control three factors that could impact test validity and the quality of linking within and between domains: balance, speededness, and connectedness. A test design is balanced if items appear in all positions in the test, which is important to mitigate position effects. Speededness relates to the question whether the test design allows sufficient time for students to reach the end of the test. In the context of test design, connectedness refers to the question of whether pairs of items and pairs of domains are observed, so that the design is properly linked and correlations can be estimated. Optimal test design (van der Linden, 2005[102]) provides a formal approach to optimise a certain objective under a set of constraints. However, given that the PISA test design has multiple goals, there are multiple ways to define objectives and constraints, which may lead to conflicts. For example, the main survey design should support both the estimation of item parameters and the estimation of student proficiency. Since the PISA field trial does not provide sufficient data to finalise item calibration, item parameters are estimated using the PISA main survey data. For this reason, the test design should be such that sufficient data is collected on all items in all participating countries/economies to ensure the quality of the PISA scale. However, when the goal is to develop an optimal adaptive test, the requirement of having sufficient data on all items provides a serious restriction on the adaptive algorithm and the freedom to select items from the pool. Furthermore, significant efforts have been made to progressively reduce the distinction between major and minor domains, which led to substantially larger item pools across domains while the total testing time per student has remained fixed at 2 hours. In addition, the test design should be able to integrate the innovative domain and optional domains (e.g.\ , financial literacy in the past and foreign language in the future).

230. The shift towards adaptive testing has required a different look on PISA test design, because the number of students following each adaptive path cannot be controlled (in contrast to controlling the number of students that is administered each test form in a non-adaptive test). Nevertheless, elements such as balancing item positions remain relevant in an adaptive testing context as well.

231. The PISA questionnaire design has similarly evolved across cycles, taking advantage of the shift to computer-based testing. In 2022, PISA applied deterministic routing to delimit the questions asked to each student based on their answers to previous questions and used a within-construct matrix sampling design whereby individual students answer a subset of items from a larger set of items for each construct. The standards on questionnaire design described below apply to the core student and school questionnaires and the optional questionnaires (for students: ICT familiarity, well-being, and financial literacy; for teacher and parents).

232. This section defines the quality standards that should guide the test and questionnaire design, for the field trial and the main study.

6.6.2. Standards on test design

Standard 6.6.1. The PISA test design shall, for all participating countries/economies, support the estimation of population distributions of student proficiency for 1) all three core domains of Reading, Mathematics, and Science, 2) subdomains of the major domain, and 3) the innovative domain, as well as estimates of joint population distributions of student proficiencies for all pairs of domains.

233. Note: The proportion and minimum number of respondents per domain, and per combination of domains, should be set in view of this goal. To evaluate this standard, the precision of summary statistics of the proficiency distributions can be used, such as the standard errors of the mean and standard deviation.

Standard 6.6.2. The test design shall be integrated such that trends over time can be accurately measured for the three core domains, student burden is minimised while the information obtained from each domain is maximised, and student proficiency can be associated with policy-relevant indicators.

234. Note: The test and questionnaire design and the associated standards should not only work well in isolation but should also provide an integrated design in which they work well together. For example, information should be provided on how certain standards can determine limitations for other aspects of the design.

Standard 6.6.3. The size and composition of the item pool for each domain shall be set in relation to the number of students completing a domain and the length of the corresponding test section, and such that all required test forms and adaptive paths can be assembled without exhausting the pool.

235. Note: The item pools for the core domains should have a size between five and ten times the average test length. The item pool for the innovative domain should have a size between three and six times the average test length. The rationale is that these numbers are intended to strike a balance between framework coverage (both content and difficulty), item production, scale stability, and test security. The recommended size of an item pool for adaptive testing is generally said to be between five and ten times the test length (Davey, Pitoniak and Slater, 2016[103]). For example, for a test length of 30 items, the recommended item pool size is between 150 and 300 items. However, since each item needs a minimum number of student responses in each participating country/economy and test forms need to have a sufficient number of common items, the item pool cannot be too large. For example, assuming 6,000 students, a test length of 30 items, the maximum number of items for the item pool is 600 to reach at least 300 student responses per item per country/economy (Zwick, 2012[104]) (see Standard 6.6.7). The composition of all item pools in terms of the assessment framework, item format, content specification, psychometric properties, and timing information should be such that all required test forms and paths can be assembled to meet the standards without exhausting the pool. For adaptive testing, this means that sufficient numbers of items should be available to create paths of varying difficulty. Even well-designed item pools may not work well for all participating countries/economies. Changes in the student population across cycles may require changes to the item pool requirements. In addition, the overlap between the distribution of student proficiency and the information function of the item pool can be used to determine how well the item pool's difficulty matches with student proficiency levels. For example, with the addition of new countries/economies in each cycle, it may be needed to add more items of a specific type (e.g. more easy items may be needed if there is a mismatch between item pool difficulty and student proficiency levels). Finally, updates to the framework may also require changes to the item pool (e.g. if subscales are changed or added).

Standard 6.6.4. The content of a single domain shall be sufficient to provide enough measurement precision while allowing most students in each participating country/economy to finish the test within the limited time (no more than 60 minutes).

236. Note: It is currently recommended that the test for a single domain should not exceed 60 minutes of testing. PISA test forms and adaptive paths can have different test lengths, both within and between domains, as long as the large majority of students in each country/economy can finish the test within the allotted time. That is, the tests should not be speeded - if the test is speeded, validity is at stake (Lu and Sireci, 2007[105]). An often-used definition of speededness is found in Swineford (1974, pp. 8-9[106]), which states that a test is "essentially unspeeded if at least 80 per cent of the group reach the last item and if virtually everyone reaches at least three-quarters of the items". In terms of test design, results of analyses on item responses and response times from data of previous cycles or the field trial should be used to ensure (to the extent possible) that the cognitive tests in the main survey are not speeded for all participating countries/economies. Since response time can be related to proficiency, special care should be taken in adaptive testing designs so that students performing at a wide range of proficiency levels are able to finish the test within the allotted time.

Standard 6.6.5. The proportion of the existing item pool maintained from cycle to cycle shall be large enough to ensure the quality of trend measurement.

237. Note: At least two thirds of the item pool should be maintained when a domain moves from a major to a minor domain cycle in order to balance framework coverage (both content and difficulty), scale stability, and test security. In addition, for a major domain, no more than two thirds of items can be newly introduced ones, and at least one third of items should be maintained from previous cycles. The following example illustrates how the standard works: If the item pool consists of 240 items in a major domain cycle, the item pool can be reduced by a maximum of 80 items to a minimum of 160 items when moving to a minor domain cycle. Similarly, at least 50% of the item pool should be retained when a domain moves from a minor to a major domain. For the example with 160 items in the item pool for a minor domain cycle, at least 80 items should be kept and at most 160 new items can be introduced. The item pool can then contain 240 items again for the next major domain cycle.

Standard 6.6.6. When subscores are to be reported for a domain, the number of items administered per subscale to each student shall be large enough to ensure a minimum level of measurement precision for each subscore.

238. Note: For instance, it is currently required that each student should be administered at least 5 items per subscore. With fewer items, issues related to complete separation in the latent regression model can occur. This phenomenon occurs if the predictors can perfectly predict all proficiency values (which may happen if the number of patterns in the predictors is equal to or larger than the number of item-response patterns) (e.g. (Zeng and Zeng, 2021[107])).

Standard 6.6.7. The number of student responses per item shall be sufficient to detect and allow potential item-by-country interactions in the measurement model.

239. Note: For instance, the number of student responses per item should be at least 300 in each participating country/economy (see Zwick (2012[104])).

Standard 6.6.8. Within each domain, the test design should be balanced such that position effects can be evaluated for all items at each key position in the test. Between core domains, the test design should be balanced such that each domain is assessed through equal numbers of students in each position (e.g. first hour, second hour) in every participating country/economy.

240. Note: The first part of this standard means that each item should appear at each key position in the test. Historically, this meant that each item appeared in each of four 30-minute cluster positions across the two-hour cognitive assessment. With multistage adaptive testing, design balance within domains means that each item should appear in every stage. Note that this does not mean that each item has equal numbers of student responses across stages, because, depending on proficiency, smaller or larger numbers of students can be routed to paths of lower and higher difficulty. However, design balance is an important aspect to maintain in future cycles in which it may be possible to create more extensive adaptive designs (e.g. unit-level or within-unit adaptivity). Maintaining design balance across evolutions of adaptive testing designs ensures that item position effects can be evaluated within and between cycles. The second part of this standard means that proficiency estimates are based on equal proportions of students across the two hours.

Standard 6.6.9. Sufficient data on item pairs within a domain shall be collected to ensure the connectedness of the design and the quality of item calibration.

241. Note: For instance, currently, it is required that at least two thirds of all item pairs within a domain should have at least 100 student responses in each participating country/economy. Having data on an item pair means that their covariance can be estimated, conditional independence assumptions in the measurement model can be evaluated, and correlations between their item parameters can be estimated. This standard should hold within trend items, within new items, and between trend and new items. Linking error can be impacted if this standard is not met.

Standard 6.6.10. The test design shall accommodate the international options that are decided for each cycle.

242. Note: For instance, the financial literacy assessment has been an international option from PISA 2012 to PISA 2022. This had an impact on sampling in that an additional number of students was required. It also impacted on the field trial and main survey design in that these should accommodate the assessment of financial literacy in combination with the other domains. The foreign language assessment is an international option in PISA 2025 and will need to be integrated in a similar fashion.

Standard 6.6.11. The impact of changes in the test design on comparability over time and across countries should be assessed and documented, for instance, through simulation studies, field-trial experiments, and/or bridge studies.

243. Note: Changes which may impact comparability over time (trend measurement) include moving from paper to computer, moving from linear or incomplete block design to adaptive designs, or moving from 3-domain to 2-domain forms. Changes which may impact comparability across countries include, for instance, differences in test administration (paper or computer-based assessment), or designs with and without the innovative domain.

6.6.3. Standards on questionnaire design

Standard 6.6.12. The PISA questionnaire design shall, for all participating countries/economies, support the estimation of an imputation model for student proficiency, allow for analyses which relate student proficiency to their background and learning experiences, and support comparisons over time and across countries of questionnaire-based indicators.

244. Note: For optimal support of the estimation of the imputation model, it is important that all students are administered the student questionnaire and that the questionnaire contains all relevant constructs. Questionnaire rotation between constructs should not be used as this affects the quality of the imputation model (see PISA 2012 Technical Report, pp. 376-384). Questionnaire design changes which may impact comparability over time should be clearly and carefully documented (e.g. major changes in question order, or changes in the way certain data are collected).

Standard 6.6.13. The total length of questionnaires should be limited to reduce respondent burden.

245. Note: It should be clear to respondents how long it will take to complete the questionnaire(s). Data on respondent burden should be collected, including timing information, evidence of disengagement, and qualitative feedback from respondents. Field trial data should be evaluated to ensure that the assigned administration time for each questionnaire is adequate. Filter questions with appropriate routing rules should be implemented to the extent possible to delimit the questions asked to each respondent based on their answers to previous questions and to limit administration time. For example, questions related to parental occupation and immigration to the country where the test is taking place have been examples where the use of filter questions (branching) made the questionnaire design more efficient.

Standard 6.6.14. For constructs assessed in any of the questionnaires using within-construct matrix sampling, the number of items administered to each student shall be large enough to ensure a minimum level of measurement precision when scaling the construct.

246. Note: For instance, the current requirement is that each student should be administered at least five items. See standard 8.4.7 in the standards on constructing and validating scales from the questionnaires.

Standard 6.6.15. For constructs whose measurement relies on questions with mixed wording and within-construct matrix sampling, each respondent shall be administered balanced numbers of positively and negatively worded questions.

247. Note: This standard enables mitigating response-style effects (i.e. students' tendencies to give construct-irrelevant responses, e.g. straight lining, extreme, midpoint) to the extent possible.

Standard 6.6.16. The conditioning variables for the imputation model(s) for student proficiency are derived from all questionnaires combined. The number of principal components extracted from these, and which enter the conditioning model directly, shall be determined to ensure that they account for the majority of the variance in the conditioning variables, while avoiding overfitting in the imputation model.

248. Note: This standard indirectly affects the number of conditioning variables that can be considered in the main imputation model used by PISA for reporting students' performance in the international database. In PISA, the principal components that account for at least 80% of the variance in the conditioning variables after sweeping collinearities and low variances (Goodnight, 1979[108]) for each country/economy are used in the imputation model. However, to avoid over-fitting, a heuristic 5%- rule is currently given priority which does not allow more than 315 principal components to be entered for a student sample of 6 300 (i.e. 315 is 5% of 6 300).

6.6.4. Standards on field trial design

Standard 6.6.17. The field trial design shall be close to the main survey design in its operational implications, such as test length, the number of breaks, or the length of the questionnaire.

249. Note: The field trial design should remain close to the main survey design, to achieve the operational testing goals of the field trial. In particular, no changes should be introduced to the presentation of items between the field trial and the main study.

Standard 6.6.18. The design shall enable the estimation of item parameters and item-level diagnostic analyses.

250. Note: The field trial design shall support item-level diagnostic analyses for all major country-by-language groups, as well as the estimation of item parameters and the detection of item-by-cycle interactions for trend items at the international level.

Standard 6.6.19. Where appropriate, the design should enable the comparison of alternative designs by including adequately powered experimental studies.

251. Note: In order to guide the final design choice for the main study or to confirm the comparability of various possible designs, the field trial should include well-powered experimental studies that compare the alternative designs.

6.6.5. Quality assurance

- The responsible contractor produces a test design plan which presents evidence to ensure that all the standards on test design are met. The test design plan should be presented to the PISA technical advisory group (TAG). The TAG should evaluate the plan according to the above standards. The TAG could ask for more evidence (e.g. request additional analysis) and/or suggest changes to the design if one or more of the standards are not sufficiently met.
- The responsible contractor produces a questionnaire design plan which presents evidence to ensure all the standards on questionnaire design are met. The questionnaire design plan should be presented to the PISA TAG and questionnaire expert group (QEG). The TAG and QEG should evaluate the plan according to the above standards. The TAG could ask for more evidence (e.g. request additional analysis) and/or suggest changes to the design if one or more of the standards are not sufficiently met.
- The responsible contractor produces a detailed description of the test and questionnaire design in the PISA technical reports.

Chapter 7. Development of data collection instruments

7.1. Standards for developing assessment frameworks and specifications for tests and questionnaires

7.1.1. Rationale

252. Assessment frameworks define the constructs assessed by the different PISA instruments (including both test instruments and contextual questionnaires), describe their overall theoretical foundations, and detail how the constructs will be measured in the assessment (see for example Ainley & Schulz (2020[109]); van de Vijver, Jude & Kuger (2019[22]); OECD (2022[110])). Thus, frameworks are of central importance in ensuring that PISA assessments are fair and provide valid and internationally comparable information (see Validity, Comparability and Fairness chapters). In the context of PISA, the frameworks also provide justification for measuring certain constructs, in terms of expected insights for education policies and practices. By defining the focus of measurement, the frameworks also ensure that PISA remains relevant and responds to the questions of policymakers and educators. Frameworks regarding the three main domains and context questionnaires (with the exception of 2003) have been published since the onset of PISA, as well as additional frameworks for the innovative domains (Problem Solving, Collaborative Problem Solving, Global Competence, Creative Thinking) and optional modules (Financial literacy, Well-Being, ICT).

253. The frameworks for the cognitive tests define the knowledge, skills and attitudes that are to be measured in each domain, and how these will be measured. They act as a blueprint which will guide the full item development process as well as analysis and reporting, by laying out the types of tasks that should be included to assess each dimension of the target construct, drivers of difficulties in the tasks, the testing time that should be dedicated to each dimension, how students' responses should be interpreted and scored, and how the evidence from the full set of tasks should be aggregated in order to support valid claims. Newly developed test frameworks should follow a principled assessment design approach, such as the evidence-centred design (ECD), and establish clear evidentiary links between claims, assessment tasks and measurement models (Mislevy, Almond and Lukas, 2003[111]). This is particularly relevant for new tests which aim to assess complex competencies, such as the PISA innovative domains. For the core domains, each iteration of the assessment framework should provide justifications for changes in how reading, mathematics and scientific literacy are conceptualised and explain the implications of these changes on the instruments and on trend measures. Moreover, the frameworks should elaborate on how the domain is taught and learned in school settings and out-of-school settings, and how progression can be shaped by curriculum and pedagogical choices.

254. The questionnaire frameworks define the contextual elements which influence outcomes from the PISA tests, and how they should be assessed (including the type of questions and the choice of respondents). In order to facilitate the alignment of the constructs to specific areas in educational research, constructs are structured according to specific taxonomies and allocated to content areas (i.e. "modules" in the PISA context; (Kuger and Klieme, 2016[112])). Previous questionnaire frameworks for PISA have thus distinguished between levels of educational governance, as well as areas of education related to all stakeholders in the context of schooling (i.e. students, principals, teachers, families). Out-of-school experiences that are likely to influence learning and well-being outcomes are also covered in PISA questionnaire frameworks. Similarly to test

frameworks, questionnaire frameworks provide guidance to the item development process by including specifications on the testing time that should be allocated to each module, the type of questions and response formats that should be used, as well as information on design choices, such as the use of filter questions or rotations.

255. Given the purpose of PISA, both tests and questionnaire frameworks shall take into account how constructs can be measured in a global study (van de Vijver, Jude and Kuger, 2019[22]) (see Comparability and Fairness chapters). Hence, in-depth review of the research findings addressing how indicators function differently across cultural contexts is required before proposing new constructs. In particular, frameworks for the PISA innovative domains should carefully consider to what extent the target construct is similarly defined and understood across cultural contexts, and whether evidence of mastery in the target competence looks similar in different countries.

256. The standards in this section present the quality criteria that should guide the development of tests and questionnaires frameworks in PISA, covering aspects related to the development process and the content of these frameworks, including tests and questionnaires specifications.

7.1.2. Standards on the framework development process

Standard 7.1.1. The frameworks should be completed before item development starts.

257. Note: This is in order to ensure that item developers can design items and evidence rules that follow the specifications in the framework.

Standard 7.1.2. Participating countries/economies shall have the opportunity to review drafts of the frameworks and contribute to their improvement. The final version of the frameworks shall be approved by the PISA Governing Board.

258. Note: The final version of the frameworks has to be shared with countries/economies well before test administration starts, to ensure that stakeholders have a good understanding of what will be assessed and can base participation decision on that.

Standard 7.1.3. The development of new frameworks and innovations in the existing frameworks shall be guided by an expert group.

259. Note: The expert group for a cognitive test should be composed by experts in the domain and experts in assessment development and measurement. The expert group from the questionnaire should include experts in education research and experts in survey design and measurement methods. The expert group should be nationally and culturally diverse.

Standard 7.1.4. The development of a new framework shall be based on a stakeholder consultation and policy review process.

260. Note: National experts from participating countries and economies should be consulted multiple times during the development of the framework. An early review should focus on the definition of the target construct for the innovative domain and on the direction of innovation in the conceptualisation of reading, mathematics and scientific literacy.

Standard 7.1.5. The development of new areas of a framework shall be based on an extensive review of research on the target construct, and be supported by new research whenever important gaps in the existing literature are identified.

261. Note: Development should start with a thorough review of existing assessments and of the scientific literature in the areas and constructs assessed. This literature review should be summarised in the framework. For the PISA innovative domains it is possible that there is no consolidated understanding of the target competences or no well-established method to collect the relevant evidence in the existing research. In those cases or in any other case where gaps in existing knowledge are identified, sufficient time and resources should be dedicated to undertaking an in-depth analysis and modelling of the domain: this might include observational studies on how students in the PISA target age develop and demonstrate expertise in the domain.

Standard 7.1.6. The introduction of new constructs shall be based on a review of previous PISA assessments.

262. Note: Material from previous PISA cycles such as the frameworks, Field Trial instruments and findings, and Main Survey findings, should be reviewed in order to evaluate whether related constructs were assessed in the past and with what results (in terms of measurement quality and relevance of the findings).

Standard 7.1.7. Measurement of new constructs shall be carefully evaluated with international comparability in mind.

263. Note: This evaluation should be based on the literature review, inputs from participating countries and economies, and the review of material from previous PISA cycles.

Standard 7.1.8. Any change in construct definition that impact trend measures should be justified, approved by the PISA Governing Board, and documented in the frameworks.

264. Note: Changes in trend measures should be based on sound scientific evidence and/or policy argument (e.g. if they are not relevant anymore).

7.1.3. Standards on the tests and questionnaire frameworks' content

Standard 7.1.9. The frameworks shall clearly define the domains and constructs that will be assessed in each PISA cycle.

265. Note: The framework should justify the importance of comparative data in the target domain for education policy and practice. In addition, given the objectives of PISA, the test frameworks should explain why and how the target skills are important for students to thrive in and contribute to modern societies. The frameworks should then present the theoretical foundations underpinning the domain, including definitions of the domain and associated subdomains, and clearly articulate what claims will be made about students' knowledge, skills, and attitudes and contexts of learning with respect to each domain. The framework should also detail the extent to which it is expected that the constructs are similarly understood and conceptualised across national contexts and cultures. The frameworks should be sufficiently clear and specific to guide all stages of the item development process.

Standard 7.1.10. The frameworks shall include test and questionnaire specifications and exemplar items that are sufficiently detailed to inform the item development process.

266. Note: The test specifications should include the relative weight with which each subdomain will be represented in the test, as well as the types of items required to assess each subdomain at varying levels of cognitive demand, including information on scoring methods. Exemplar items should also be developed for each subdomain, level of cognitive demand, and item type to further clarify expectations. Furthermore, the specifications should explain how the results of the test are expected to be reported. Questionnaire specifications should indicate the relative weight and associated testing time that should be dedicated to each module of the questionnaire framework, and provide information on design choices, such as the use of filter questions or rotation.

Standard 7.1.11. The questionnaire frameworks shall indicate which respondents are best suited to provide information on each module.

267. Note: The frameworks need to clearly articulate which constructs are best measured using a specific source of information (e.g. teachers rather than students), and which ones require triangulation of data from multiple sources. The justification regarding which respondent should provide information on each module should be based on evidence from research in the target domain and review of material from previous PISA cycles.

Standard 7.1.12. The cognitive test frameworks should clearly present the measurement approach, following an evidence-centred design (ECD) framework.

268. Note: The ECD approach (Mislevy, Almond and Lukas, 2003[111]; Goldhammer et al., 2021[113]) involves developing a competency, task and evidence model in order to produce accurate inferences about a student's competencies. The framework should carefully consider which target competencies would be best measured through event/process data, following a "process data by design" approach (Maddox, 2023[114]).

7.1.4. Quality assurance

- The responsible contractor or research team produces a framework development plan indicating the milestones of the framework development process. The plan should:
  - Detail how the expert groups will be selected and consulted during the framework's development process;
  - Detail the development team's composition, including their qualifications and relevant experiences and diversity (e.g. nationality, gender);
  - Detail the process for seeking inputs from education stakeholders in different countries on updates to existing framework (for main domains and questionnaire framework). For the innovative domains there should be documented evidence that most countries agree with the relevance and cultural appropriateness of the new assessments;
  - Allocate sufficient time to countries' reviews of the frameworks and specifications and subsequent revisions;
  - Allocate sufficient time for content-matter experts' review of the frameworks and specifications.

- The frameworks should include:
  - Definitions of the target constructs.
  - Detailed test specifications and exemplar items.
  - A synthesis of an extensive literature review that reflects experience gained from previous assessments and measurement research in the area, as well as research in cognitive development, teaching and learning in the areas being assessed.
  - A justification of all changes to the previous frameworks, based on scientific evidence and policy arguments.
  - A section that specifically addresses the extent to which constructs are likely to manifest differently across countries.
- The responsible contractor produces a final report at the end of the process. The report should demonstrate that the frameworks development plan was implemented, providing information on:
  - A summary of the engagement and consultation with countries, including indications of main revisions undertaken following countries' feedback and documenting their approval for any change in trend measures;
  - A summary of the contribution of the expert group.
- The responsible contractors produces a detailed description of the test and questionnaire frameworks development process in the PISA technical reports.

7.2. Standards for developing, reviewing, evaluating and revising test items and scoring rubrics

7.2.1. Rationale

269. The cognitive tests items are the cornerstone of PISA. Test items need to ensure that PISA effectively assesses "the extent to which [students] have acquired key knowledge and skills essential for full participation in social and economic life" and "how well [they] can extrapolate from what they have learned and apply their knowledge in unfamiliar settings, both in and outside of school." (OECD, 2019[3]). A sound item development process is thus critical to fostering validity in PISA. This process needs to ensure that the test aligns with the framework and specifications and minimises construct-underrepresentation and construct-irrelevant variance (see Validity chapter).

270. Item development in PISA should be informed by evidence and respect rigorous validation protocols. The process should involve several cycles of item design, review by domain experts, piloting, and revisions.

271. Careful item development is the first step in ensuring cross-country comparability. In order to ensure that items are not biased towards a particular socio-economic context or culture, national experts from a wide range of countries should be involved both in providing content for new items and reviewing that the context and cognitive demand of tems are appropriate for their students.

272. Item design and review should also pay particular attention to fairness and accessibility issues (see Fairness chapter). In order to avoid discriminating between students on characteristics other than the targeted skills, items should not contain features and material that create irrelevant cognitive, emotional and physical barriers for groups

of students (Educational Testing Service, 2016[115]). Rather, the item development process should follow Universal Design principles, which is "an approach to assessment development that attempts to maximise accessibility of a test for all of its intended test takers" (AERA, APA and NCME, 2014[1]).

273. Lastly, a sound item development process should also ensure that the PISA test assesses the target skills reliably (see Reliability chapter). This implies, in particular, that specific attention is paid to the development of clear and consistent scoring rubrics for constructed responses, to ensure that the same items are scored the same way by different raters.

274. Following the move to computer-based assessments, PISA has started to incorporate technology-enhanced items (TEIs) in its assessments. TEIs denote computer-delivered items that allow students to interact with the test environment, going beyond simply selecting or typing a response. Among other advantages, TEIs can better measure complex constructs, enhance student's engagement, and offer test experiences that are closer to real-world tasks and learning. Moreover, TEIs open the possibility of collecting and exploiting rich data on students' response processes ("process data"), which can be used to improve the evaluation of students' performance in interactive tasks, describe solution strategies or produce indicators of metacognitive processes and engagement, and may thus inform teachers and educators on how students solve items. Process data are also useful as sources of validity evidence, for instance by enabling to identify items prone to guessing and which need to be revised. However, designing TEIs is a more complex endeavour than designing traditional item types. In particular, it requires an interdisciplinary team of test developers, including content experts, experts in user experience and user interface (UX/UI) design, software developers, as well as experts in innovative measurement methodologies. More time and iterations should be allocated to their development (including additional validation steps such as usability studies). TEIs' item developers need to be careful not to introduce construct-irrelevant variance resulting from differences in familiarity with digital tools, if digital literacy is not part of the target construct. An evidence-centred design, where student model, an evidence model, and a task model are formulated in a systematic way, needs to be followed to ensure that process data are collected and used according to evidence rules defined by experts and validated through cognitive laboratories and pilots.

275. This section presents the standards which should guide the item development process in PISA as well as those describing the properties of a good item, including standards which are specific to the development of technology-enhanced items.

7.2.2. Standards on the item development process

Standard 7.2.1. The item development team shall include domain experts, experts in test development, UI and measurement experts.

276. Note: Item development is a collaborative process which needs to involve a diverse team of experts. Item developers need to have proved experience in developing test items for students around the PISA target age, and in the target domain (or experience in developing innovative item types if they work on the PISA innovative domains). For the development of TEIs, the team should also include measurement experts with expertise in process data, data scientists and data architects, software developers, as well as UI/UX designers. Item developers have to be trained by the framework developers on the target content, skills and specifications.

Standard 7.2.2. Item development shall build on the test framework and specifications.

277. Note: The item development process should start when the framework (including the test specifications) are in their final stage. Each test item should be mapped to the constructs defined in the framework, and this mapping should be confirmed by content experts, including those involved in developing the framework and independent experts.

Standard 7.2.3. The item development process shall be iterative and engage a team of item developers over several months.

278. Note: The item development process should leave sufficient time for multiple iterations of the development cycle (design, review, pilot and revise). This is especially the case for technology-enhanced items. Therefore, the time planned should be proportional to the scope of change in trend domains, or the innovative nature of new domains.

Standard 7.2.4. National experts from a wide range of participating countries shall contribute to the item development process.

279. Note: Participating countries and economies need to be involved in the item development process by submitting stimulus material and associated items. This should be facilitated through use of an online item-submission tool and the organisation of item-writing workshops with national experts. Item submission guidelines need to be developed to guide national experts in this process. Guideline documents should include information such as the type of material that is expected, a summary of the test framework and specifications, as well as model items. Items proposed by national experts should undergo the same review process as items developed by the PISA contractors.

Standard 7.2.5. Items shall undergo a review by national experts under appropriate confidentiality conditions. Their views on item improvement shall be taken into account by item developers.

280. Note: Participating countries and economies should be provided with the opportunity to review the items once they have been developed. Sufficient time should be granted to countries for this review and allocated to the subsequent revisions. The review process should be structured such that countries are able to provide ratings, comments and recommendations for revisions for each item. Country reviews should assess the relevance and appropriateness of the items instructions and context for their socio-economic and cultural context and for their education system. Following these reviews, transparent documentation should be produced explaining how the recommendations from national experts have been addressed. The criteria for addressing the recommendation should be transparent and communicated to the PGB. The review should be carried in a secure environment.

Standard 7.2.6. Items shall undergo a review by an international group of experts.

281. Note: The development process needs to include sufficient time and opportunities for regular item review and advice by an independent expert group. Experts' reviews should assess the alignment of the item pool to the test framework and specifications, the quality of distractors for multiple-choice items, the clarity and completeness of scoring rubrics for open response items, the quality of design of the interface in TEIs, and the validity of evidence rules for items with more complex scoring (e.g. TEIs).

Standard 7.2.7. Newly developed items shall be tested in qualitative pilot studies, including cognitive laboratories, and in quantitative pilot studies. Qualitative studies shall notably focus on user experience. These pilots shall be administered in multiple countries and languages.

282. Note: Cognitive laboratories focus on the processes that an assessment-taker engages in when presented with an assessment task. The laboratories can reveal instructions that are difficult to understand, items that are unclear or are subject to different interpretations, and vocabulary that is unnecessarily complex. Usability studies focus specifically on issues related to interface and task design in TEIs, looking at how intuitive a student finds the interactions within the digital environment. Quantitative pilot studies aim to verify at an early stage issues such as the accuracy of data capture, item/task difficulty and duration, dimensionality of the constructs, issues with the implementation of scoring rules and with the interpretation of process data sequences, empirical relationships between item scores within and between units and item types. These studies are particularly important for validating innovative domains and for complex, technology-enhanced tasks, because they help refine scoring models. As much as possible, cognitive laboratories usability studies and pilot data collections should be conducted in several participating countries and involve students of various levels and background in order to ensure that the population tested is as representative as feasible. The contractor should produce material ensuring a standardised study protocol across countries, and, when relevant, should explore the feasibility of conducting these studies using multi-modal data collection technologies, such as eye-tracking devices. The validation of items should include analysis of response process data, such as reports of thinking aloud processes in cognitive labs and log-files. This analysis is particularly important to detect differences in familiarity with the content and engagement across countries and groups of students.

Standard 7.2.8. Innovative item types and response formats shall be theoretically justified and tested.

283. Note: PISA should strive to improve its measurement of high-order thinking skills by introducing innovative item types (for instance, technology-based items enabling to assess process-based constructs). Innovative item types and response formats, including those taking advantage of digital technologies, should be carefully tested before inclusion in the cognitive tests. The theoretical motivations for choosing an innovative format should be elaborated, and an appropriate testing strategy should be implemented (see Standard 7.2.7). Such justification should be shared with countries/economies.

Standard 7.2.9. Items should be fully transferable to other software platforms.

284. Notes: The contractor should program the computer-based items in a standard format (QTI) or ensure that they are otherwise fully transferable to other software platforms (for example through the use of Portable Customised Interactions [PCIs]). Technical documentation should be produced to facilitate the transfer of items to another testing platform.

7.2.3. Standards on item properties

Standard 7.2.10. Items shall measure the competences (knowledge, skills and attitudes) defined in the framework.

285. Note: Each item is specifically designed to provide evidence on one or more facets of the competency model defined in the assessment framework.

Standard 7.2.11. The item pool shall meet the test specifications.

286. Note: The item pool reflects the targets defined in the test specifications about the distribution of items by facet/dimension of the test construct, item format (e.g. constructed-response and MCQ), and inclusion of TEIs. The items are designed so that they cover a range of cognitive demand that is aligned with experts' expectations about students' performance.

Standard 7.2.12. The instructions shall not be unnecessarily complex.

287. Note: The instructions for each unit and items are appropriate, complete, and not confusing. The wording is clear and appropriate for the whole PISA population. The material does not assume prior knowledge that is inappropriate or irrelevant given the target construct assessed. The reading load (i.e. text and sentence length, syntax, vocabulary) is deliberately minimised for tests which do not aim to assess reading ability.

Standard 7.2.13. The typography, colours, format, layout and response method shall not constitute irrelevant barriers for students.

288. Note: Item content should not include unnecessary physical barriers, i.e. physical aspects of material such as the typography, colours, format, layout or response method which are not helpful to assess the target construct and which may be unnecessarily difficult for students, and especially those with sensory or motor problems. For instance, this covers the inclusion of visual stimuli that are not needed to assess the construct, or unnecessarily complex or cluttered; low level of contrast; hard-to-read fonts. The response format is the most suitable for the target construct.

Standard 7.2.14. Task contexts should be familiar, appropriate, and accessible to all PISA students.

289. Note: The stimulus material is clear and accessible to all PISA students, avoiding as much as possible topics that are only familiar to students in some regions or cultures. Given that cultural references cannot be completely eliminated in any assessment material, it is important to make an effort to ensure that multiple cultures are represented, for example by including several items that are proposed by national teams.

Standard 7.2.15. The tasks should represent a meaningful application of the domain in a real-world context.

290. Note: Given the focus of PISA on assessing students' capacity to use and apply the knowledge they acquired at school in real-world situations, most items should present students a real-world scenario where they are asked to complete a task that is relevant and plausible to them.

Standard 7.2.16. Items shall not contain words, visual or audio content that is generally regarded as sexist, racist or negative toward cultural groups, or potentially offensive and sensitive.

291. Note: Items should not include irrelevant emotional barriers which can disturb students and make understanding or responding to an item more difficult. The instructions and material (including images and contexts) should use non-discriminatory language, avoid stereotypes and use appropriate terminology to denote people of different groups. Items' visuals should represent diversity when people are depicted.

Standard 7.2.17. The choice of distractors shall be justified.

292. Note: In multiple choice items, the incorrect options proposed, or distractors, should be plausible (e.g. reflecting students' common misconceptions) yet undeniably wrong. The correct answer should not stand out from the distractors. The choice of distractors should be justified and documented in the item submission forms.

Standard 7.2.18. TEIs shall have an intuitive and engaging interface.

Standard 7.2.19. Computer-based items shall be designed in a way that minimises the impact of familiarity with digital tools and devices on student performance.

293. Note: The design of computer-based items, and in particular of TEIs, should take into account the fact that participating students may not have the same experience with and knowledge of digital tools and may be used to different input devices. When relevant, items should contain a training/tutorial section to familiarise students with the tools and motions (e.g. 'drag and drop') required to complete the items.

Standard 7.2.20. The uses of process data to measure cognitive processes in TEIs shall be clearly and consistently identified and defined. PISA students shall be informed before the test that their process data would be collected and scored.

294. Note: In TEIs, any use of process data in scoring needs to be clearly justified through the development of detailed evidence and scoring rules that define how sequences of actions captured in the process data relate to the target cognitive processes in the framework. Test developers define which information has to be extracted from the log files and collaborate with programmers to ensure that all the relevant information is consistently captured and parsed.

Standard 7.2.21. Open-ended response items shall be accompanied by clear and detailed scoring guides and each country/economy shall attend a training on scoring.

295. Note: Scoring guides enable countries to score open-response items reliably, ensuring that they are interpreted in the same way across countries and scorers so to allow for a good level of consistency in the coding. The scoring guides should describe the scores available for each item (e.g. full credit and partial credit score), define the numerical codes associated with each score, include descriptions and examples of responses for each coding category. The guides should also present general principles for scoring, such as how to treat spelling and grammar mistakes, and when to consult a supervisor. Lastly, they should highlight common problems encountered when scoring items and how they should be treated. The guide should be complemented by training on item scoring.

7.2.4. Quality assurance

- The responsible contractor produces an **item-development plan** indicating the milestones of the item-development process. The plan should:
  - Ensure minimal overlap between the framework-development process and the item-development process;
  - Detail the role of the expert group in overseeing the item-development process;
  - Detail the item-development team's composition, including their qualifications and relevant experiences, and the responsibility of each member in the item development process;
  - Include information on overall time (e.g. months of full-time staff, by team member) that will be dedicated to item development and revisions;
  - Allocate sufficient time to countries' reviews of items and subsequent revisions;
  - Allocate sufficient time for content-matter experts' review of items and subsequent revisions;
  - Define the scope and use of cognitive laboratories, quantitative pilots and usability studies and subsequent revisions;
  - In the case of TEIs, allocate sufficient time for usability studies and subsequent revisions.
- The responsible contractor produces regular **progress reports**. The reports should:
  - Indicate the state of advancement of the items;
  - Present the review and validation activities undertaken, documenting how items are being revised (including the criteria for modifying items, or accepting/rejecting items) and including any deviations from targets indicated in the item development plan;
  - Detail the time invested by each staff, indicating deviations from targets in the plan;
  - Indicate changes in the staff;
  - Note any other adjustment to the timeline.
- The responsible contractor produces a **final report** at the end of the process. The report should prove that the item development plan was implemented, providing information on:
  - Staff time allocated to the activities specified in the plan;
  - The final item pool and its alignment with the test specifications;
  - A summary of findings from pilot studies and subsequent revisions to items;
  - A summary of the engagement and consultation with countries, including indications of main revisions undertaken following countries' feedback;
  - A proof that the expert group has reviewed and approved the final item pool.
- The responsible contractor puts in place an easy-to-use **online item-submission tool** for countries and issues **item submission guidelines**

- The responsible contractor produces **items in QTI format** or otherwise ensures that they are **transferable** to other platforms
- The responsible contractor produces **reports** following the cognitive laboratories, usability studies and quantitative pilots. The reports should:
  - Describe the design of the studies and supporting documentation;
  - Present the results of the studies;
  - Examine what the results reveal of students' behaviour and what they mean for item quality;
  - Explain how items will be revised in light of these results.
- The responsible contractor produces documentation on how items will be coded, defining which process data will be collected for each item and how the data will be logged.
- The responsible contractor produces coding rubrics for human-scored items, and undertakes training of national teams.
- The responsible contractor produces a detailed description of the cognitive test items development process in the PISA technical reports.

7.3. Standards for developing questionnaire items

7.3.1. Rationale

296. Since the first round of PISA in 2000, the programme has included questionnaires for students and the principals of their schools to assess contextual factors of teaching and learning. The questions focus on different characteristics of the students and different aspects of their learning environment, considering a multi-level perspective (i.e. the context of the individual student; the home, peer and neighbourhood environment; the school context; the wider cultural community). This includes information on the individual background and learning behaviour of students, learning settings and opportunities at home and in the classroom, information on what students are taught and how, as well as aspects of educational policies, such as the availability of different resources at school or the existence of certain school-specific programs. Increasingly, the questionnaires have also been used to collect information on students' well-being in and out of school, and on several social and emotional skills such as resilience or growth mindset. Additionally, PISA offers optional questionnaire modules to collect information from teachers and parents, as well as to inform specific analyses, such as on educational pathways and expectations, or on the impact of the COVID-19 pandemic. At each cycle, new questionnaire material is developed to better contextualise and interpret students' performance and analyse opportunities to learn in the innovative domain.

297. These questionnaires are designed to provide context for interpreting PISA results, by describing differences between and within systems, highlighting areas of policy intervention, or contributing to explain the impact of specific educational policies over time (on the importance of context indicators in large scale assessment, see for example (Kuger and Klieme, 2016[112])). Consequently, findings from the questionnaires occupy a central space in PISA reports and secondary analysis, and attract considerable attention from policymakers, researchers, and the general public. Contextual information assessed by the questionnaires is also an essential input to the measurement models used in the scaling of the cognitive indicators of PISA.

298. As for the cognitive test items, quality standards need to be followed when developing new questionnaire content. This is particularly important in an international study like PISA, as any technical errors might impact the measurement quality and affect international comparability. First, it is essential that the questions and scales included in the PISA questionnaires build on established and validated research. At the same time, it is necessary to track and incorporate developments in survey research methodology that can allow to better measure new constructs (such as socio-emotional skills) and reduce the potential impact of cultural differences in response style (i.e. students' response patterns which are unrelated to the target construct, such as acquiescence bias).

299. Moreover, questionnaire indicators need to adhere to measurement standards for assessing and providing good quality information over time. This includes a continuous evaluation of adequateness and relevance of every indicator, as well as assurance mechanisms to maintain measurement quality for new or updated indicators (on the need of high-quality questionnaires and advanced methodological approaches, see for example (Rutkowski and Rutkowski, 2018[116])).

300. Lastly, trend measurement occupies a central place in the questionnaires. PISA is a monitoring study based on the continuous assessment of so called "trend indicators" to enable for analysing the change in teaching and learning contexts as well as outcome measures over time. These can be indicators which are assessed in every cycle, or only related to the rotating major domains (trend over several but not all timepoints of the assessment). Measuring trends reliably requires minimising changes in the measures. However, trend items sometimes need to be adapted to changes in society (e.g. home possessions, or how parents or guardians are defined). Changes in trend indicators should be carefully monitored (Teltemann and Jude, 2019[117]).

301. The standards included in this section define quality criteria for the development of questionnaire scales and individual items, providing guidelines on how to address different types of response bias. They also define the processes for reviewing and revising the items, and for documenting the whole process of item development.

7.3.2. Standards on the item development and selection process

Standard 7.3.1. Questionnaire items shall be developed and selected based on criteria outlined in the questionnaire framework.

302. Note: The assessment framework defines the content areas of the questionnaires which should be included in the cycle, including measures that should be assessed in every cycle (trend across four years) or within the major domain (trend across twelve years), as well as the areas where new content is expected to be developed. All modules and constructs within modules in the framework should be sufficiently covered, following guidance from the Questionnaire Expert Group (QEG) on the selection of scales and items, and from the Technical Advisory Group (TAG) on measurement standards.

Standard 7.3.2. Questionnaire development shall be based on a thorough review of the existing literature and scales.

303. Note: Development should start with a thorough review of existing literature on the assessment of the construct in question. This should include analysing existing scales' usability in PISA. Existing scales should ideally have been already implemented in studies with a comparable age-group and have documentation available on measurement quality in different language versions and cultural contexts (e.g. analysis of latent structure, reliability, and invariance).

Standard 7.3.3. Items from previous PISA cycles shall be evaluated on their adequateness for inclusion in the questionnaire.

304. Note: Existing items from all previous cycles should be evaluated regarding a) their coverage of the constructs as described in the framework, b) their measurement quality, based on empirical research, and c) whether they are up-to-date with respect to education and broader social changes. Evaluation of measurement quality should be based on information in the PISA Technical Reports and additional criteria defined by the QEG and TAG in the respective cycle.

Standard 7.3.4. Any change in trend indicators shall be documented and justified.

305. Note: A trend item or scale becomes outdated if a) it is not relevant anymore for several countries, for example if it refers to an educational track no longer in place in most countries or b) it uses specific terms or classification that are no longer in use (e.g. the term "educational software" was expanded to "educational software or Apps"), c) it is not suitable anymore for capturing variation in a construct (e.g. the possession of an outdated mobile device), d) it is unlikely to be used in international reporting (e.g. it has not been used in the reports for a previous cycle), e) it shows lack of variation (e.g. due to ceiling effects) for many countries in the previous cycle, f) new research has shown that innovative items or changes in wording would improve measurement quality, or g) specifications in the assessment framework call for a reduction in the assessment time for the construct and thus for a shortening of trend scales.

Standard 7.3.5. Newly developed content shall be evaluated by the national experts nominated by the participating countries.

306. Note: The country review should focus on the appropriateness and sensitivity of the question for the target group (e.g. whether it is politically sensitive, risk of cultural bias, compliance with local data privacy regulations), expected quality of the data (e.g. whether the target respondent has the knowledge to answer the question), and needs for adaptation to any national context.

Standard 7.3.6. New content and adaptations of existing items shall be validated through cognitive laboratories in multiple countries and languages.

307. Note: Cognitive laboratories with the target groups (e.g. students, teachers) can indicate if the items are understood and answered correctly by participants, and are not unnecessarily complex or confusing (see for example NCES (2002[118]), Willis (2005[119]), Beatty & Willis (2007[120])). Different methods can be used, such as think aloud interviews, probing, or paraphrasing (see (Brancato et al., 2006[121])). As much as possible, cognitive laboratories should be conducted in several participating countries and involve participants of various background in order to ensure that the population tested is as representative as feasible. The contractor should produce material ensuring a standardised study protocol across countries. In addition to qualitative evidence from cognitive laboratories (e.g. through think aloud interviews), the validation of items should include analysis of response process data in order to identify items with abnormally long response time or those that generate confusion or disengagement among any group of participants. The contractor should produce reports after each study, documenting the analysis and how items are being revised.

Standard 7.3.7. Innovative item types and response formats shall be theoretically justified and tested.

308. Note: PISA should strive to achieve high measurement quality of questionnaire data by introducing innovative item types that have been validated in published research. Innovative item types and response formats, including those taking advantage of digital technologies, should be carefully tested before inclusion in the main study questionnaires. The theoretical motivations for choosing an innovative format should be elaborated in the documentation accompanying the questionnaire. Moreover, an appropriate testing strategy should be defined and implemented for the innovative items: this strategy can include experimental studies integrated in the cognitive laboratories or the field trial (i.e. a random group of students being tested with an innovative item and another group tested with a traditional format on the same construct).

7.3.3. Standards on item properties

Standard 7.3.8. Items shall measure the constructs defined in the framework.

309. Note: Each item is relevant to the research and policy questions asked (as captured in the framework), and specifically designed to provide evidence on the constructs defined in the questionnaire framework.

Standard 7.3.9. Items formulation shall be clear, precise and unambiguous for all PISA students.

310. Note: Items should not be confusing and should not contain complex or unfamiliar words. When the use of complex or unfamiliar words is necessary, these should be defined in the questionnaire itself (additional help can be given by adding an explanation in the test administrators' notes). The wording should be precise, for instance defining the question in time and space. The terminology, length and content should be adapted to the target population and age-group (e.g. students, parents, teachers). Items should be specific, avoiding double-barrelled questions (item should relate to a single idea only), or negatives. The items should be understood in the same way by PISA participants from all countries and socio-economic and cultural contexts (see also the Standards for translations and adaptations for standards related to translatability assessments).

Standard 7.3.10. The questionnaire layout shall be appealing, easy to navigate and clear.

311. Note: In the paper-based version of the questionnaire, items should be appropriately spaced-out on the page. In the computer-based version, participants should be able to easily change their response, including going back to the previous question. The progression from item to item should be smooth and logical.

Standard 7.3.11. The selection of items and scales shall consider the time constraints of PISA.

312. Note: As assessment time is a constraint for measuring constructs in PISA, the development of context questionnaires should consider decisions regarding single items measures versus latent constructs assessed with multi-item scales. This decision should be based on theoretical and methodological discussions and needs to consider the implications for reporting and secondary analysis using PISA data (see for example Diamantopoulos,

A. et al. (2012[122])). Analysis of response time during the Field Trial can guide the selection process (see Standards on the analysis of Field trial data).

Standard 7.3.12. Items shall avoid high cognitive burden on participants.

313. Note: Items should not include too many response categories (Krosnick and Presser, 2009[123]). Items should also avoid asking respondents to recall distant events, especially when the information required is too specific. To support participants in the retrieval process, memory aids and retrieval cues helping to recreate the event context can be included in the item.

Standard 7.3.13. Items shall not contain words or sentences that is generally regarded as sexist, racist or negative toward cultural groups, or potentially offensive/sensitive. Students shall be informed before the questionnaire that they are not required to answer any question they feel personally sensitive.

314. Note: Item questions and response categories should use non-discriminatory language, avoid stereotypes and use appropriate terminology to denote people of different groups. Items should represent diversity when people are depicted (in visuals or text).

Standard 7.3.14. Item formulations shall be objective.

315. Note: The questions asked should not influence participants to answer in a particular way, or reveal the item developer's own opinions (e.g. avoiding leading or unbalanced questions).

Standard 7.3.15. Items should minimise social desirability biases in participants.

316. Note: Item developers should avoid the inclusion of questions that induce participants into giving a positive image of themselves or their environment and peers.

Standard 7.3.16. The questionnaire should minimise order effects.

317. Note: Item developers should be mindful of introducing context effects in the questionnaire such as item or response order effects (i.e. items that appear earlier in the questionnaire may influence participants' interpretation of later items). In particular, item developers should avoid placing questions that may influence the emotional state of respondents (e.g. questions on bullying or on feelings of (in)competence) at the beginning of the questionnaire.

Standard 7.3.17. Item answer categories shall cover all possible responses and avoid overlaps that would create ambiguous results.

318. Note: Items using a Likert scale should reflect the entire measurement continuum. There should be no overlap between response categories when a single answer is expected.

Standard 7.3.18. When relevant, national/regional items should be included in the questionnaire.

319. Note: Participating countries/economies should be able to adapt items which are context-specific and for which comparability cannot be ensured. For scales that are most affected by contextual differences (e.g. socio-economic status) , a set of "optional" items should be developed, from which countries can choose. For instance, the items on household possessions can be adapted to have different versions for groups of countries sharing similar levels of economic development, geography, or cultural background (Avvisati, 2020[78]).

Standard 7.3.19. The skipping and routing paths between the different items shall be defined.

320. Note: If an item is not targeted at all participants, there should be a visible option to skip the item. In order to avoid asking questions to participants when they do not apply to them, branching should be used so that such questions are asked only following specific answers in the previous question.

Standard 7.3.20. Item identifiers shall be standardised and enable the identification of trend items, new items, and changes in trend items.

321. Note: There should be a unique number for trend items, and a cycle specific number for new items for each cycle.

Standard 7.3.21. Items shall be fully transferable to other software platforms.

322. Note: The contractor should program the computer-based items in a standard format or ensure that they are otherwise fully transferable to other software platforms. Technical documentation should be produced to facilitate the transfer of items to another testing platform.

7.3.4. Quality Assurance

- The responsible contractor produces a questionnaire item development plan indicating the milestones of the item development process. The plan should:
  - Ensure no or minimal overlap in the timeframe between the framework development process and the item development process;
  - Detail how the QEG and other relevant expert groups will be consulted during the item-development process;
  - Detail the item development team's composition, including their qualifications and relevant experiences, and the responsibility of each member in the item development process;
  - Include information on overall time (e.g. months of full-time staff, by team member) that will be dedicated to item development and revisions;
  - Allocate sufficient time to countries' reviews of items and subsequent revisions;
  - Allocate sufficient time for content-matter experts' review of items and subsequent revisions;

- Define the scope and organisation of cognitive laboratories and other pilot studies;
- Define experiments and other studies that will be undertaken as part of the Field Trial to validate new items or innovative item types, and how results would be incorporated into the Main Survey;
- Allocate sufficient time for a translatability assessment and subsequent revisions.

- The responsible contractor produces a final report at the end of the process. The report should prove that the item development plan was implemented, providing information on:
  - Staff time allocated to the activities specified in the plan;
  - The final item pool and its alignment with the test specifications. This should consist of a comprehensive overview of how all items in the questionnaires relate to the modules/constructs defined in the questionnaire framework. For existing item used in previous PISA cycles, this also includes how the respective constructs have been assessed in the previous cycles, their measurement quality and a rationale why they were selected for the current cycle. For items and scales from other studies, this should indicate the studies in which the items have been used and their measurement quality;
  - Justifications for changes in the trend indicators;
  - The rationale for including innovative designs and formats;
  - A summary of findings from pilot studies and subsequent revisions to items;
  - A summary of the engagement and consultation with countries, including indications of main revisions undertaken following countries' feedback.
- The responsible contractor produces a detailed description of the questionnaire items development process in the PISA technical reports.

7.4. Standards for translations and adaptations

7.4.1. Rationale

323. In a large-scale comparative assessment such as PISA, cross-linguistic, cross-national and cross-cultural equivalence is not only an objective – it is a fundamental requirement without which the whole notion of quantitative cross-cultural comparison is invalidated (Dept, Ferrari and Halleux, 2017[124]). Therefore, PISA dedicates substantial attention and resources to the localisation process and lays down stringent procedures for its implementation. Localisation is the process by which the international English source or master version (the original text) of the data collection instruments (test batteries, contextual questionnaires, coding guides) is refined and complemented by a French source version and then goes through translation, adaptation and validation to become target versions or national versions in the languages of each participant[1]. This process strives to ensure that localised instruments meet the core principles of the PISA quality target: validity, reliability, comparability, and fairness.

324. The localisation process aims at achieving linguistic equivalence, i.e. semantic congruence, same quantity and quality of information, similar register, as well as replicated reference chains, matches and patterns. For PISA cognitive test items, this means that localised items from the cognitive tests should assess the same competence and skills

and appeal to the same cognitive processes as the source language items. The localisation process should not introduce biases likely to distort international comparisons. This level of functional equivalence and invariance cannot be guaranteed a priori, and direct validation on the basis of purely linguistic criteria is not sufficient. The validity argument shall rely on an indirect validation through analysis of the field trial data (see Standards 8.1). For PISA questionnaire items, linguistic equivalence means that questions are understood in comparable ways across respondent groups and the way the questionnaire author intended them to be.

325. In PISA, the localisation process is a collaborative endeavour: the National Centres are responsible for producing their localised instruments, under the guidance and supervision of the Contractors− in particular the Contractor for Linguistic Quality Assurance & Linguistic Quality Control (LQA-LQC) and the Translation Referee. Linguistic quality assurance is the methodology developed to ensure that localisation procedures will be capable of delivering equivalent instruments in all countries/economies, languages and cultures, while linguistic quality control consists of a set of measures designed to monitor whether the standards are met and propose corrective measures as needed (Dept, Ferrari and Wäyrynen, 2010[125]).

326. The standards in this section describe the processes that should be followed when developing translations and adaptations for the PISA data-collection instruments as well as the characteristics of translated/adapted items (i.e. the outcomes of the localisation process).

7.4.2. Standards on translation and adaptation processes

Standard 7.4.1. Newly developed items shall undergo translatability assessment before translation.

327. Note: The translation/adaptation process should start with source items that are appropriate for a multi-lingual, multi-national and multi-cultural survey. The aim of translatability assessment is to optimise the source material by identifying and remedying potential translation/adaptation issues at the onset, before translation even starts. The translatability assessment process consists of submitting 'mature' draft versions of new items to a pool of experienced linguists covering a broad range of language groups, who report on any hurdles they encounter as they each translate the material to their language; this feedback is collated and sent to the item developers, including suggestions for rewording or for adding translation/adaptation notes (see Standard 7.4.2). Translatability assessment is complementary to cognitive laboratories, usability studies and pilot studies (see the Standards 7.2 on the item development process) and can be performed in parallel with them.

Standard 7.4.2. Newly developed items shall include translation and adaptation notes.

328. Note: Translation and adaptation notes clarify the intended meaning of concepts, phrases and/or terms in the source text and provide information which allows translators, reconcilers/adjudicators, or verifiers to focus on what is meant in survey measurement terms. Their purpose is to guide translators and reconcilers on specific translation and adaptation issues (Survey Research Center, 2016[126]). They are prepared before the start of the translation by a panel of test designers, bilingual domain experts and professional translators, who identify the crucial elements of each item, the psychometric pitfalls (e.g. related to distractors or a synonymic correspondence

between a stimulus and an answer option), and special difficulties. The translatability assessment (see Standard 7.4.1) and the production of a second source version (see Standard 7.4.3) are important wellsprings for relevant translation and adaptation notes. Translation and adaptation notes are best presented "segment-by-segment" i.e. linked to the specific part of the source text to which they refer.

Standard 7.4.3. Newly developed items shall be produced in two source versions: English and French.

329. Note: PISA requires the development of a second source version in French. Two independent translations from English into French are produced and merged into a third version, which is submitted to a set of stringent linguistic control procedures until it is deemed to have the same status as the original English source version (Grisay, 1999[127]). The main advantage of using two source versions (either when double translating or cross-checking, see Standard 7.4.5) is that the second source provides equivalent alternatives of wording and syntax, as well as an illustration of the degree of acceptable translation/adaptation freedom. Another important benefit is that the development of the French source version feeds back to the original English source version (identification and correction of residual defects) and feeds significantly into the production of item-per-item translation/adaptation notes (see Standard 7.4.2).

Standard 7.4.4. All localisation processes shall allow the use of modern translation technology on electronic files.

330. Note: All linguists involved in the localisation process should be enabled to use state-of-the-art translation technology through CAT (computer assisted translation) tools. This makes it possible to fully separate text from layout (allowing translators and reconcilers to focus on content rather than on format issues), to exploit what are known as "language assets" (translation memories, glossaries), and to deploy automated quality checks. The file formats and the platform set-up should also allow the linguists to view their work in progress as it would appear to test-takers (preview functionality). The digital tools used for translation should limit as little as possible the organisational discretion of the National Centres.

Standard 7.4.5. Translated items shall be developed through a double-translation and reconciliation procedure. This is not required for other materials (coding guides, school-level materials).

331. Note: Double translation plus reconciliation is widely recognised as a very robust design -for translating items for an international comparative survey (see e.g. (Dept, Ferrari and Wäyrynen, 2010[125])). Two translators translate the source material into the target language independently. A third person (senior linguist) reconciles these two translations into a single national version, taking the best from each. Double translation from two parallel source versions in two different languages is recommended. When using a double translation from two source versions design (Grisay, 2003[128]) as in PISA, one translation is performed from the first source version and the second translation from the second source version (Dept, Ferrari and Halleux, 2017[124]). In those countries or economies where finding competent translators from both source languages is a problem, an alternative consists of double translation and reconciliation from one of the source languages, followed by extensive crosschecks against the second source language (Grisay, 2003[128]). The standard is relaxed for coding guides and for school-level materials, for which single translation and review is deemed sufficient.

Standard 7.4.6. Adaptation of already translated materials, whenever possible, should be preferred to translation 'from scratch'.

332. Note: This applies of course to participants testing in English or French. Participants who share a language (e.g. Spanish, German, Chinese, Arabic) are encouraged to enter collaborative models, for instance by producing a common reference version, which each participant then adapts to their 'flavour' of the language and their local context. In other cases, a participant may make arrangements to borrow a same-language version developed by another participant. All such arrangements are not only practical operationally, but also enhance cross-linguistic equivalence.

Standard 7.4.7. The National Centres should be guided and assisted throughout the localisation process by the responsible Contractor(s), including in the implementation of a robust validation process.

333. Note: Whether they are translating or adapting, the National Centres are guided, instructed, and assisted by the Contractors, in particular the contractor for LQA-LQC and the translation referee through the development of localisation guidelines, trainings, user guides for the use of the electronic translation tools, as well as support during the localisation process (including technical support for the use of IT tools; e.g. through a helpdesk or Frequently Asked Questions). The validation process on translated/adapted national versions should be carried out in collaboration with the National Centres. Validation should include the following steps:

- Sentence-by-sentence verification by experienced independent verifiers, trained to assess linguistic equivalence of target versus source version(s) and fluency/correctness in the target language. In particular, verifiers check compliance with the item-specific translation and adaptation guidelines.
- The verification feedback (proposed corrections accompanied by explanatory comments) is reviewed by senior staff of the contractor and by the translation referee, who will flag crucial issues for follow-up.
- Post-verification review by the National Centres; possible controversies about crucial issues are negotiated and resolved by the translation referee, who liaises if needed with the item authors.
- Technical and linguistic final check, to ensure crucial issues (layout or linguistic) have been resolved.

Standard 7.4.8. The localisation process carried out for each national version shall be documented.

334. Note: The documentation is traditionally carried out in Excel monitoring worksheets that accompany the instruments throughout their journey across translation/adaptation and validation. They are organised in rows for each segment of the source version with columns for the translation and adaption notes, and columns to record the successive comments/resolutions/agreements/checks by translators, reconcilers, adapters, verifiers, translation referee, country reviewers, final check verifiers. The purpose is to have a recorded history of the localisation of each instrument. For questionnaires and school-level materials, these monitoring instruments are used before localisation to negotiate and agree on adaptations.

Standard 7.4.9. **Trend items shall be centrally managed.**

335. Note: To maintain measurement and comparability over time, trend items (items reused from a previous PISA cycle) should stay identical as administered in the previous cycle. To this end, differently from newly developed items, it is best to implement central management, i.e. to shift control over these materials from the National Centres to the Contractors. The contractors retrieve the national versions of trend items and allow the National Centres to review them but without editing rights. The National Centres may request correction of outright errors, which should be documented and approved before implementation by the contractors. Note that this process does not apply to countries that are new to PISA or did not participate in the cycles where (some of) the trend units were administered. A similar process should be organised for the review of translated/adapted materials between Field Trial and Main Survey, given the similar challenges: participants should only reflect updates to the source materials or correct items that did not perform satisfactorily at Field Trial. It should be noted that in case of mistakes in the translations of trend items, it should be possible to revise and correct the items before the next cycle.

Standard 7.4.10. **Errata corrections shall be made in translated/adapted national versions, up to the point in time where this is possible for all national versions.**

336. Note: After source materials are released for localisation, errors may be identified in these materials. Errata notifications are then made available to the teams responsible for translating/adapting or validating, which do not all work on the same timeline. The "next person in line" is tasked with making sure to reflect the errata correction in the national version in their charge. To safeguard comparability, errata notices that come after a first national version is 'locked' (ready for administration in the field) should be postponed e.g. to Main Survey phase.

7.4.3. Standards on outcomes of the localisation processes

Standard 7.4.11. **Localised items shall have the same format, layout, and graphical elements or illustrations as source items.**

337. Note: Localised PISA items should have the same "look and feel" as the original source items. Items that are open-ended in the source version should never be turned into multiple-choice items in the national version, or vice-versa. The order or the content of headings of responses presented as columns in tables should never be changed. For example, "Yes/No" or "True/False" categories should not be changed into "No/Yes" or "False/True". Illustrations and graphics should appear in the same position, with attention to captions or legends which may be longer in translated versions. Students should not need to scroll to see content if this is not needed in the source item. For right-to-left languages, some illustrations or graphics may need mirroring. For paper-based administration of cognitive test items, the final items need to be reviewed for any alteration that may have occurred when assembling translated materials into booklets.

Standard 7.4.12. Localised cognitive items shall assess the same competence and skills and appeal to the same cognitive processes as the source language items.

338. Note: Translators, reconcilers, verifiers, and reviewers should pay attention to the following aspects (in general but most importantly also in relation to specific passages):

- Unintentional modification of the difficulty level of the questions asked to the students should be avoided. The translation should match the source in consistency, fluency, accuracy, style, register, and use of terminology.
- Translation should not introduce ambiguities that are not intended by the source.
- In multiple choice items, the relative lengths of the 'key' (correct answer) and of the distracters (wrong answers) should reflect the respective lengths in the source − this is a well-known aspect that drives the attractiveness of a response. Similarly the translation should not provide clues that direct the students towards the key (correct answer) or, conversely, that make a distractor (wrong answer) more attractive. The elements that make the difference between responses should not be toned down.
- Literal and synonymous matches between words found in a question and words used in the preceding 'stimulus' (the material that the student needs to examine to find the answer) should be reflected in the translation.

Standard 7.4.13. Localised questionnaire items shall "Ask the Same Question" as the source language items.

339. Note: For PISA questionnaires, linguistic equivalence corresponds to the shorthand "Ask the Same Question" – as known in comparative survey design and implementation (CSDI) (Survey Research Center, 2016[126]). In studies that are administered in more than one language or culture, all the respondents should understand the question in the same way. The focus of the translation and adaptation of questionnaires should thus be on the fluency, accuracy, on the rendering of the scales as equivalent as possible, and on appropriate adaptations.

Standard 7.4.14. Localised items shall be appropriately adapted to the local context.

340. Note: This applies to both cognitive test and questionnaire items. An adaptation is an intentional deviation from the source version(s) made for cultural reasons, to conform to local usage, or, in the case of cognitive items, to avoid putting the respondent at an advantage or a disadvantage. In questionnaires, the comparability objective calls for heavier adaptations.

7.4.4. Quality assurance

- Each National Centre produces and submits a Translation Plan for their national version, to be negotiated and approved by the translation referee. For each survey instrument (cognitive items, questionnaire items, coding guides, school-level materials), the appropriate localisation procedure in relation to the Technical Standards is hence agreed before operations start.
- The responsible contractors provide trainings for the translation teams within the PISA Centres, including webinars and instructional videos.

- The responsible contractors produce translation/adaptation guidelines for the translation teams within the PISA Centres (translators, reconcilers, reviewers, etc). These include:
  - General instructions about the type and nature of the materials that will be translated;
  - Translation approach and its goals, the procedures to be followed;
  - Tasks to be completed;
  - Security requirements.
- The responsible contractors develop user guides for the use of the electronic translation tools by the translation teams within the PISA Centres (translators, reconcilers, reviewers, etc). These include:
  - Instructions for using the CAT tool;
  - Instructions for accessing and using the test administration platform during the translation and adaptation process, in particular for previewing the source version(s) and the translation/adaptation in progress;
  - Organising directories in which to save successive versions of the files; using proofing tools; using an equation editor; documenting the localisation process in Excel monitoring spreadsheets; etc.
- For each national version, the detailed localisation 'history' is recorded in Excel monitoring spreadsheets, which are kept on file and may be consulted e.g. during FT data analysis, MS data adjudication.
- The contractor responsible for LQA-LQC produces a report on localisation activities for inclusion in the technical report. The report provides detailed information on the procedures employed, the outcomes (statistics on identified/corrected errors, illustrative examples), and learnings for future cycles.
- The responsible contractor produces a detailed description of the translation and adaptation process in the PISA technical reports.

Chapter 8. Scoring and analysis

8.1. Standards on analysis of Field Trial data and additional preparatory analyses

8.1.1. Rationale

341. The PISA Field Trial analysis aims to support the development of the set of assessment instruments that can achieve the objectives of the Main Survey. The Field Trial itself has two main goals. The first goal is to identify and solve any operational or platform issues. The second goal is to obtain the information that is necessary to assemble and recommend the complete set of assessment instruments for the Main Survey. Thus, the analysis of Field Trial data is a crucial step to ensure the validity, reliability, comparability and fairness of PISA. The data analyses are driven by the Field Trial design described in the PISA integrated design document. In addition to analyses of Field Trial data, other preparatory analyses may be required to evaluate the final assessment instruments. Furthermore, the Field Trial can allow carefully planned experiments related to changes and innovations in design and methodology.

342. In order to confirm the administration procedure and the psychometric properties of the items, checks on data yield and data quality of the Field Trial should be conducted in all participating countries/economies. Any issues detected and solved in this phase enhance the data quality and smoothen the process in the Main Survey when there is much more data to handle, within short timelines. Problematic items can be removed before the Main Survey, thereby enhancing reliability and validity, and the final test design can be optimised using Field Trial results.

343. Item-response-theory (IRT) models have been used in PISA to create a common scale for reporting the cognitive assessment results in each domain (van der Linden, 2018[129]; Rutkowski, von Davier and Rutkowski, 2013[130]). Although the IRT model in PISA has changed from a one-parameter logistic model in PISA 2000 to PISA 2012 (Adams, Wilson and Wang, 1997[131]) to a two-parameter logistic model in PISA 2015 and beyond (von Davier et al., 2019[132]), as did the associated estimation and fit procedures, it remains important to assess how well the selected IRT model fits the data. IRT models are applied to the Field Trial data from the cognitive tests for each domain to 1) get an indication of the stability of the PISA scale from the trend items, 2) evaluate the quality of the scale resulting from adding newly developed items and/or new participating countries/economies, and 3) assemble the cognitive tests for the main survey. Additional assumptions of the applied IRT models such as those related to dimensionality, local independence, and invariance of item position should be evaluated as well since failure to meet these assumptions affects the quality of statistical inference. The invariance assumption with respect to item position is of particular importance since items may be used in different positions in the Main Survey than in the Field Trial.

344. Because there can be differences in test design between the Field Trial and Main Survey (e.g. the Field Trial can follow a linear test design while the Main Survey is adaptive), additional preparatory analyses, beyond Field Trial analysis, are required. Field Trial data analyses and additional preparatory analyses are important to support and evaluate the assembly of the cognitive tests in the Main Survey. They give insight into what data can be expected from the Main Survey and can be used to evaluate whether the standards on test design will be met (see the Standards 6.6 on test and questionnaire design). Analyses may be based on Field Trial data, but also on additional simulations using Field Trial results and/or Main Survey results from previous PISA cycles.

345. The Field Trial is also used to evaluate the characteristics of background questionnaires (core and optional; see Standards 7.3), facilitate decision making about which questions can be used in the Main Survey, and determine which derived variables can be considered for the database. There are two phases in the analysis of the Field Trial data from the background questionnaires. In the first phase, data quality checks and routines are implemented to ensure that the data is collected in the right way in each country/economy so that it can be compared across countries/economies. This includes checks on the creation of derived variables (DVs, which are based on multiple questionnaire items or on recoded data from individual items) not based on IRT. Data quality checks include checks on the range of open-ended questions, consistency across instruments, and consistency within instruments across questions. Countries/economies need to give feedback as well as to perform validity checks on specific questions and to check the harmonisation of national categories into international ones as to guarantee comparability. Furthermore, data cleaning is to be performed by checking established indicators for rapid responding and response style, recoding reverse-keyed questions, and coding missing data. In the second phase, analyses are conducted related to descriptive statistics, dimensionality, IRT scaling, scaling of the economic, social and cultural status variable (ESCS), relationships between constructs and with proxies of proficiency. After these analyses have been evaluated, the goal is to select questionnaire content for the Main Survey using criteria related to both framework representation and measurement quality.

346. Finally, the PISA Field Trial can allow experiments related to both design and methodology to evaluate enhancements and innovations both with respect to the cognitive tests and the main survey.

347. The standards in this section aim to describe the analyses that are necessary to achieve the goals of the Field Trial for the three core domains, the innovative and optional domains, and the background questionnaires. Furthermore, they also serve to establish the process and responsibilities in between the Field Trial and Main Survey.

8.1.2. Standards for the analysis of data from the cognitive tests

Standard 8.1.1. A Field Trial shall be implemented in all participating countries.

348. Note: The Field Trial is not expected to produce representative data. The number of participating schools and students should be set at the minimum level that allows for the psychometric analysis described in the standards below (see also Standards 6.4 on sampling).

Standard 8.1.2. The data yield and basic quality of the cognitive tests in the Field Trial shall be evaluated in all participating countries/economies.

349. Note: The evaluation of the data yield includes examining data analyses related to quality control, response rates, item statistics, coding reliability, and timing. It adheres to checking survey operations, instrumentation, performance of the computer delivery platform, and response rates in relation to the design of the Field Trial and PISA standards (e.g. sample size requirements). Checks on basic quality of the cognitive tests serve to ensure the reliability and validity of the items and consists of inspecting item statistics related to difficulty, discrimination, distractors, missingness, and response time as well as coding reliability of constructed-response items for both human and automated scoring. Checks should also include the analysis of timing data in order to evaluate potential issues related to test speededness and student engagement.

Standard 8.1.3. Item parameters shall be estimated using the Field Trial data using established procedures, and the fit of the IRT model shall be evaluated in all participating countries/economies.

350. Note: The fit evaluation of the IRT model involves item fit, dimensionality, local independence, and position effects. Item fit should be evaluated for trend items to evaluate invariance of item parameters compared to the PISA scale established in previous cycles and for all items to evaluate potential item-by-country interactions (i.e. evaluating the extent to which items continue to function consistently across cycles and countries/economy). Items that show significant misfit within countries/economies should be reviewed and remedial actions may need to be taken (e.g. removal from the trend measure). Items that show substantial international misfit should be removed for the Main Survey.

Standard 8.1.4. Additional preparatory analyses shall be performed to support and evaluate the assembly of the cognitive tests in the Main Survey.

351. Note: These analyses should be performed regardless of whether the tests are administered in linear or (multistage) adaptive form. They can consist of inspecting psychometric features of Main Survey tests such as the test characteristic curves, the test information functions of each test form and each adaptive path in the proposed test design for the Main Survey, the item exposure rates, routing rules (in case of adaptive testing) in terms of how many students end up in different levels of the adaptive test, and an evaluation of other objectives and constraints that were applied in the assembly. These analyses can also include simulations to evaluate the efficiency of the adaptive test in countries/economies of different proficiency levels. Analyses may be based on Field Trial data, but also on additional simulations using Field Trial results and/or Main Survey results from previous PISA cycles.

Standard 8.1.5. For the cognitive test of the innovative domain in the Field Trial, the same standards 8.1.1 – 8.1.3 shall apply, considering the specific nature of the innovative domain.

352. Note: Differences with the core domains which should be considered include the size and composition of the item pool, test design, and sample sizes. Depending on the substance of the innovative domain, special attention may be required for specific aspects of the innovative domain (e.g. the coding of responses, IRT modeling).

Standard 8.1.6. For the cognitive test of the optional domain(s) in the Field Trial, the same standards 8.1.1 – 8.1.3 shall apply, considering that a subset of countries/economies participates.

353. Note: Analyses of the FT should consider that a subset of countries/economies participates in the optional domain, and take into account differences such as the size and composition of the item pool, test design, and sample sizes. Special attention may be required for aspects of an optional domain (e.g. listening in the foreign language assessment).

Standard 8.1.7. Proxy scores should be provided for each cognitive domain.

354. Note: Proxy scores generally are standardised country-specific proficiency estimates and therefore not comparable across countries/economies. They are however useful to evaluate statistical associations with other variables (e.g. from the background questionnaires) and can thereby help guide the assembly of Main Survey instruments (e.g. by validation and/or selection of questionnaire variables).

Standard 8.1.8. Experiments conducted in the Field Trial related to either design or methodology for the cognitive tests shall be carefully evaluated.

355. Note: Enhancements and innovations in either design or methodology of the cognitive tests should be supported by theoretical and/or empirical results and approved by the PISA TAG and OECD before they can be carried over to the Main Survey. The PGB should be informed of any enhancement or innovations that have been approved.

8.1.3. Standards for the analysis of data from the background questionnaires

Standard 8.1.9. Data quality checks shall be performed for the core questionnaires.

356. Note: Checks related to response rates, recoding of responses, range restrictions for open-ended questions (e.g. out of range values for parental occupation), consistency across questions and instruments should be implemented. Furthermore, checks can be performed related to the number of responses in each category, including missing data. For specific questions, higher-level validity checks can be conducted by each country/economy (e.g. gender and age distributions).

Standard 8.1.10. Routines for creating derived variables for the core questionnaires without using IRT shall be implemented and evaluated in each country/economy.

357. Note: The creation of derived variables without using IRT, referred to as simple indices, either by combining multiple questionnaire items or by recoding data from individual items, can typically be automated. This automation can be checked by creating cross-tables of the derived variables with the original variables.

Standard 8.1.11. Analyses of core questionnaire data in the Field Trial shall be conducted related to descriptive statistics, dimensionality, IRT scaling, scaling of ESCS, and relationships between constructs and with proxies of proficiency.

358. Note: Descriptive statistics of questionnaire items and simple derived variables would include distributions of responses for each country/economy, including missing values. Furthermore, the random form assignment and item selection needs to be verified.

359. For the optional questionnaires, the same standards 8.1.10-8.1.12 should apply, considering that a subset of countries/economies participates.

Standard 8.1.12. Experiments conducted in the Field Trial related to either the design or methodology for the questionnaires shall be carefully evaluated.

360. Note: Enhancements and innovations in either design or methodology of the questionnaires should be supported by theoretical and/or empirical results and approved by the PISA TAG, QEG, and OECD before they can be carried over to

the Main Survey. The PGB should be informed of any enhancement or innovations that have been approved.

8.1.4. Quality assurance

- The responsible contractor produces Field Trial analysis plans for the cognitive tests and the background questionnaires, to be presented to the OECD, the PGB, the TAG, and the SMEG, so that comments and questions can be addressed. The final analysis plans need to be approved by OECD. Separate plans should be produced for the core domains, the innovative domain, and the optional domain; and for the core questionnaires and the optional questionnaires. The contents of the Field Trial analysis plans are to be presented to the OECD, the PGB, the TAG, and the QEG, so that comments and questions can be addressed. The final analysis plans need to be approved by OECD and the TAG. These plans should contain:
  - The steps in the proposed analysis;
  - The deliverables;
  - The timeline, including a schedule for finalising the Main Survey instruments
- The responsible contractor should provide to the OECD and participating countries/economies the Field Trial data files.
- The responsible contractor should produce Field Trial analyses reports for both the cognitive tests and the questionnaires. The reports should include:
  - Data yield and data quality summary tables;
  - Summaries of coding reliability and the performance of automated scoring methods;
  - A summary of IRT model fit;
  - A summary of additional preparatory analyses
- The responsible contractor should provide country/economy-specific analysis reports. The reports should contain:
  - Tables of coding reliability;
  - Item analysis tables;
  - IRT item analysis tables;
  - Tables with summaries of proxy scores.
- The responsible contractor should present summaries of the Field Trial results at relevant meetings (PISA Governing Board meeting, National Project Manager meeting and Technical Advisory Group meetings).
- The PISA Technical Report should describe how the Field Trial informed the assembly of the Main Survey instruments. This should include results from the analysis of Field Trial data, additional preparatory analysis, and the evaluation of experiments conducted in the Field Trail.

8.2. Standards on computing survey weights and preparing datasets for estimating sampling variance

8.2.1. Rationale

361. In simple random sampling (SRS), sample statistics are unbiased estimators of population characteristics (Kish, 1965[97]). However, as presented in Standard 6.4 (relative to sampling), the PISA sampling design is characterised by several features such as multi-stage sampling, stratification, and clustering. These complex design features need to be accounted for in the estimation of population characteristics and their standard errors. If these design-based features are ignored during estimation, the main principles of PISA regarding validity, reliability, and comparability1 would be violated (see Cochran (1977[133]), Lohr (1999[98]) and Särndal, Swensson and Wretman (1992[134]) for more details on the underlying statistical theory). The complex sampling design affects the estimation of two parameters, the population characteristic itself and its sampling error, in different ways.

362. First, varying selection and nonresponse probabilities induce biased estimations of population parameters. Indeed, the selection probabilities are not equal for the sampling units, for example because disproportional samples are selected from different explicit strata (see, e.g. Dumais & Gough (2012[135]), Meinck (2015[136]), Meinck & Vandenplas (2021[99]) for details). In addition, not every sampled individual will respond to the survey. For these reasons, in PISA and other large-scale studies, sample statistics are not unbiased estimators of population features as such. To reflect the varying selection probabilities, design weights are computed. To address unit nonresponse, so-called nonresponse adjustments are derived to make up for the loss in data yield. These adjustments reflect the nonresponse model imposed in the estimation. Nonresponse can lead to substantial bias when three conditions hold (Lohr, 1999[98]; Meinck and Vandenplas, 2021[99]): (1) the response rate to the survey is relatively low; (2) there are relatively large differences between the characteristics of respondents and nonrespondents; and (3) nonresponse is highly correlated with survey outcomes. As opposed to item level nonresponse (i.e. failure of respondents to answer single items), unit-level nonresponse creates higher risk to the validity of the ILSA because usually very little is known about the nonresponding schools and students. Condition (1) is addressed in PISA by determining thresholds for minimum participation rates for schools and students. Conditions (2) and (3) are difficult to assess, which is why PISA employs by default two approaches to address potential issues here. The first measure is to pre-assign replacement schools for nonparticipating sampled schools, which share similar observable characteristics than the ones that they replace (see Standards 6.4). The other measure is to compute nonresponse adjustments within supposedly homogenous adjustment cells, assuming a non-informative response model (i.e. nonresponse occurs at random) within these cells. It should be highlighted that both these measures come with strong and usually unverifiable assumptions, which is why achieving high response rates is of utmost importance. The product of all design weights and nonresponse adjustment factors is called an estimation weight2. Its use statistical analysis with PISA data allows valid conclusions to be drawn about the population features based on the PISA samples.3

363. Second, all methods employed in complex sampling (i.e. systematic and multiple stage sampling, cluster sampling, stratification, and sampling with unequal selection probabilities) have a direct or indirect effect on sampling precision.4 Therefore, standard errors in PISA cannot be estimated assuming SRS methods. In most large-scale assessments in education, resampling techniques such as the Jackknife Repeated Replication method (JRR; (Wolter, 2007[137])) or Balanced Repeated Replication (BRR;

(Kish and Frankel, 1970[138])) are used to achieve unbiased estimates of the sampling distribution of an estimator. Failure to recognise the correct methodology (e.g. using a method applicable only to SRS) will likely lead to considerably underestimated standard errors, inflating the precision level of the results (i.e. underestimating the confidence intervals), and can lead to falsely detecting significant group differences. Both JRR and BRR can be implemented utilising a set of so-called replicate weights that is prepared by the PISA sampling contractor and provided to users together with the public PISA dataset.

364. This section presents the standards that PISA should follow to compute design weights, nonresponse adjustments, and the resulting estimation weights. The standards also cover participation rates and the preparation of datasets for variance estimation5. It should be noted that these standards only apply to the Main survey data collection, and not to the Field Trial (no estimation nor replicate weights are needed for the Field Trial, as this stage of the survey does not aim for estimating population characteristics).

8.2.2. Standards on computing design weights

Standard 8.2.1. Design weights shall be computed as the inverse of the selection probability at each sampling stage.

365. Note: Selection probabilities need to be tracked and documented at each selection stage, within each sampling unit containing a further subsample, and, if applicable, for each target population separately to provide the basis for design weight computation.

8.2.3. Standards on computing nonresponse adjustments

Standard 8.2.2. Nonresponse adjustments shall be computed with the goal to reduce potential nonresponse bias in parameter estimates.

366. Note: It is important to reflect the selection probabilities of sampling units in the analysis to achieve unbiased parameter estimates by using weights. To reflect on the nonresponse model imposed in the estimation, responding units need to carry the weight of nonresponding units. As a principle, participating units should carry the weight of similar nonparticipating units. However, in cross-sectional large-scale assessments only little information is available on nonresponding units. Further, the information available has often only limited predictive power on the main outcome variables. Therefore, while using the available information is reasonable, caution should be placed into the mediative power of such variables in a nonresponse adjustment model. Explicit, and possibly also implicit strata (if large enough), provide usually reasonable adjustment cells to compute nonresponse adjustments for both sampling stages.

Standard 8.2.3. Nonresponse adjustments shall be computed at each selection stage.

367. Note: Nonresponse can occur at each stage of the sampling process: schools, but also students within schools may not participate. If a school or one of its replacements does not participate, other similar schools participating shall carry its weight. That is, the design weight of the nonparticipating school is distributed across participating schools with supposedly similar characteristics (usually those within the same stratum). If a student within a participating school does not participate, its weight is distributed among students within this school who share similar characteristics. Currently, adjustment cells are constituted by gender and grade within schools. If further information is available

(e.g. achievement-related information), more sophisticated adjustment models may become feasible in future.

Standard 8.2.4. Nonresponse adjustments cells shall be large enough to not induce too large variation in weights.

368. Note: Nonresponse adjustments lead to variation in estimation weights whenever nonresponse occurs to a different extent across strata or sampling units. The increase in weight variation is negatively related with the size of the adjustment cell, i.e. the smaller adjustment cells, the larger adjustment factors. Hence, small adjustment cells should be avoided. At the school level, this can be addressed during sampling by avoiding selecting only few schools within a stratum. When within-school participation is relatively low, similar schools can be collapsed to create student nonresponse adjustment cells. In general, non-response adjustments greater than a factor of 2 should be avoided.

8.2.4. Standards on computing and providing estimation weights

Standard 8.2.5. PISA estimation weights shall be computed as the product of all design weights and nonresponse adjustments at the multiple sampling stages.

369. Note: Estimation weights reflect the probability of participating of each sampling unit: it accounts for the probability of selection and the probability of response. The estimation weight of a sampling unit can be interpreted as the number of units in the sampling frame represented by this unit.

Standard 8.2.6. Estimation weights shall be computed and provided separately for each unit of analysis.

370. Note: PISA provides data for schools and students, each comprising data with its own analytical value. Estimation weights shall also be provided for further (optional) populations of special interest (e.g. teachers).

Standard 8.2.7. Trimming weights should be avoided and, when done, well justified.

371. Note: Even though the PISA sampling design is "self-weighting", i.e. aims for a low variability across estimation weights, various conditions can cause relatively large variation in weights, a circumstance adding considerably to the sampling variance. Trimming of weights has been used in previous PISA cycles to address this issue, however, this can lead to both bias in parameter estimates and bias in sampling error estimates. While the risk for parameter bias is usually small (Kish, 1992[139]), the latter issue may be bigger. Large weights correctly reflect the error induced by the sampling method. If the weights are trimmed, the sampling errors will be systematically underestimated. However, conservative trimming in exceptional cases may be defendable.

8.2.5. Standards on participation rates

Standard 8.2.8. **Participation rates thresholds and consequences for not meeting these thresholds, including annotation rules, shall be determined prior to the survey.**

372. Note: PISA aims for participation rates of 100% for schools and students: unit nonresponse creates a high risk of bias and should therefore be avoided. However, meeting this goal is usually practically not feasible, which is why thresholds and consequences for not meeting them need to be established. Data deemed to be potentially less reliable due to low participation rates should be annotated. Annotation rules should be determined prior to data collection and ideally remain stable across cycles. These thresholds and rules should be documented in the PISA Technical Standards agreed by the PISA Governing Board at the commencement of a cycle.

Standard 8.2.9. **Participation rates shall be expressed as population estimates, reflecting selection probabilities and non-response patterns at each sampling stage by means of weights. Participation rates shall be reported separately for schools, for students within participating schools, and combined.**

373. Note: Unweighted participation rates reflect merely data collection yield and therefore do not reflect the unwillingness of the PISA school and student population to participate in the survey.

Standard 8.2.10. **Schools should count as participating if a specified response threshold is met. Data collected from schools not meeting the threshold should be deleted.**

374. Note: Determining an optimal threshold is difficult as it depends on the nature and context of nonresponse. It is important to highlight that, on the current nonresponse model maintained in PISA, deleting data from schools with low response rates implies that the weight of students in the affected schools will be carried by other students in participating schools. That is, the model assumes that other students in participating schools represent students in affected schools better than the (few) students in the schools affected by this rule. On the other hand, if the data is retained, one assumes that participating students in affected schools represent better the nonparticipating students. Both assumptions are strong and cannot be validated. Other disadvantages of dropping data is a loss in statistical power and wasting resources. One option may be enlarging nonresponse adjustment cells across schools to prevent large adjustment factors and reduce potential bias while keeping this data. More research is needed to collect evidence on determining reasonable response thresholds within schools. At the moment, the sampling contractor should carefully investigate occurring issues, consider the options and consequences, and discuss them with the TAG and the Sampling Referee.

Standard 8.2.11. **High quality Nonresponse Bias Analysis (NRBA) results should be considered to provide useful evidence on bias risks.**

375. Note: The nonresponse model described above assumes that within adjustment cells nonresponse is ignorable. Assumptions cannot be tested; however, it is possible to use auxiliary information to try to justify such assumptions. Reliable population statistics or statistics on nonparticipating schools or students that use variables highly correlated with main survey outcomes can help evaluate the risk of bias introduced by such assumptions. For example, if students' mathematics grades are available for responding

and nonresponding students, comparing the distribution of scores between these two groups can inform the credibility that nonresponse occurred at random. Participating countries should be encouraged to collect such information already during the data collection to have it at hand in time for a potential NRBA. Note that NRBA is, for reasons of statistical power, only reasonable if considerable nonresponse occurs.

Standard 8.2.12. The computation of participation rates shall coincide with the retained dataset and nonresponse adjustments.

376. Note: Thresholds for response rates within schools should determine whether a record is retained in the database, and whether or not it contributes to the participation rate. This is important for transparency and reproducibility but will also reflect the assigned credibility level in the data.

8.2.6. Standards on preparing datasets for variance estimation

Standard 8.2.13. PISA uses BRR with Fay's adjustment for estimating sampling variance.

377. Note: BRR with Fay's adjustment (Judkins, 1990[140]) has been traditionally and successfully used since PISA 2000. Other recognised methods exist such as jackknife repeated replication or bootstrapping, that may provide similarly efficient estimators. However, unless new methods with significant advantages are developed, it is recommended to stick to this best practice.

Standard 8.2.14. The sampling contractor shall provide replicate weights that can be used for estimating sampling error in any statistical analysis. Replicate weights should be based on the BRR method described in Judkins (1990[140]).

378. Note: For each target population, separate replicate weights should be provided. When adjusting weights across primary sampling units, adjustments should be re-calculated per each replicate weight.

8.2.7. Quality Assurance

- The responsible contractor develops a weighting plan which presents evidence to ensure that all the standards on weighting are met. The plan should be presented to and evaluated by the PISA technical advisory group (TAG).
- The responsible contractor computes weighted participation rates and collects, if applicable, evidence from countries regarding nonresponse bias risks. This information is presented to the PISA sampling referee, who makes recommendations to the OECD Secretariat on how to treat data with questionable quality.
- The responsible contractors should guide countries in conducting non-response bias analyses. Non-response bias analyses reports should be made available to the public, in a format that does not increase the risk of re-identification of schools and students sampled for PISA.
- The contractor implements internal quality control by applying the four-eye principle, that is, one team member weights the data, another team member checks the weighting process and results using both common sense and a dedicated checklist.

- The contractor documents the weighting and nonresponse adjustment procedures as well as the participation rates for each country and target population in detail in the PISA technical report. The whole weighting process shall be reproducible based on this documentation.
- The contractor produces an annotated database flagging data with low reliability.

8.3. Standards on constructing and validating scales from cognitive tests

8.3.1. Rationale

379. The construction and validation of PISA scales from the cognitive tests are important steps to be able to report the results of the Main Survey. The designs for administrating the PISA cognitive tests are based on multiple matrix sampling so that not all students take all domains and not all students see all items. In addition, for most domains, multistage adaptive testing is employed to enhance measurement precision (see Standards 6.5). This combination of matrix sampling and adaptive testing leads to different groups of students taking different test forms with different difficulty within each country/economy. Statistics based on the number of correct responses can therefore not be used for reporting survey results. For this reason, IRT methods (van der Linden, 2018[129]) are used to construct scales from the cognitive tests administered in the Main Survey. The international scaling approach evolved from using a random sample of 500 students from 27 participating OECD countries in PISA 2000 to using essentially all historical data in PISA 2015. In each cycle, the complexity of scaling increases and choices are made to manage the ever-increasing amount of data.

380. Since students are administered different subsets of items (i.e. test forms) from the item pool, it is important to be able to link these subsets. In addition, it is crucial to allow linking to previous cycles in order to estimate trends. To address these needs, the test design in PISA is such that all items in a domain are linked between test forms within a PISA cycle and linked between PISA cycles by using trend items. This within- and between-cycle linking ensures that concurrent calibration methods can be employed to estimate an IRT model with which the PISA scale can be constructed (Lord, 1980[141]).

381. The between-cycle linking can be performed in a variety of ways (Kolen and Brennan, 2004, p. Chapter 6[142]). Several options are possible, ranging from a separate calibration using only data of the most recent PISA cycle, to a full concurrent calibration that includes data from all PISA cycles (Robitzsch and Lüdtke, 2022[143]). In recent cycles, a common-item equating to a calibrated item pool has been used to establish the link with previous cycles (see PISA 2018 Technical Report, Chapters 9 and 12 (OECD, 2020[21])). This entails fixing the item parameters for trend items to the values from the previous cycle and estimating the item parameters for new items with data from the current cycle.

382. The approach with which item fit is evaluated in PISA has changed over the years, and, since this is an active area of research, will likely change in the future. The invariance approach in PISA essentially evolved from deleting misfitting items at the country/economy level to accommodating item-by-country interactions. Although the definition of item invariance in groups is arbitrary (Bechger and Maris, 2014[144]; Robitzsch and Lüdtke, 2022[143])(Bechger & Maris, 2015; Robitzsch & Lüdtke, 2022), an evaluation of item fit in countries/economies remains necessary according to professional standards (e.g. (AERA, APA and NCME, 2014[1]), Standard 4.10).

383. In addition to scales for each domain, subscales of the core domain should be, and have been reported since PISA 2000. In fact, multiple sets of subscales have been construed using different, yet overlapping, classifications of the item pool (e.g. subscales

related to content and to process have been created in mathematics). Both unidimensional and multidimensional approaches have been employed to construct subscales in PISA. For example, in PISA 2018, item parameters were obtained by applying a univariate IRT model to each domain. Multiple multidimensional conditioning models were then estimated to obtain plausible values for each domain and reported subdomain. Subscales generally consist of fewer items than the full scales, both at the level of the item pool and at the level of the number of administered items to each student. For this reason, subscales can have limitations and it is important to evaluate their measurement properties. For example, methods for evaluating the added value of subscores beyond overall proficiency scores can be applied (Haberman, 2008[145]; Haberman and Sinharay, 2010[146]).

384. In PISA, the IRT scaling is combined with a population model in the form of a latent regression model that extracts information from the responses to the background questionnaire and, more recently, additional data from the cognitive tests such as response times. Thus, a set of conditioning variables is created to predict student proficiencies. Given the large number of potential conditioning variables, principal component analysis has typically been used to obtain a reduced set, either starting from the full set or the full set reduced by a small number of direct covariates deemed important for reporting (e.g. gender). This combination of scaling and conditioning increases measurement precision and facilitates the unbiased estimation of relationships between proficiency and contextual information (congeniality). Once the IRT model and the population model are estimated, plausible values (multiple imputations) can be drawn from the posterior distribution of proficiencies. Given the large number of steps in the creation of the plausible values that make up the eventual PISA scales in the different domains and the number of choices in each step, it is important to document each step.

385. Estimating scores' reliability is crucial at this stage, as developed in Chapter 3. Since reliability is inversely related to the uncertainty in the estimation of country/economy means (Adams, 2005), its evaluation across countries/economies is important. Furthermore, it is also key to estimate linking error, which describes the uncertainty that results from linking between PISA cycles using different items and different data on common items. Due to changes in PISA's methodology, the method to estimate linking error has evolved, and is also likely to evolve in the future. Until PISA 2012, linking error was defined as the standard deviation of equated item difficulty parameter estimates of common items from two separate calibrations.1 Since 2015, linking error has been estimated by the standard deviation of the equated country/economy means from two calibrations (note that if concurrent calibration was used in which common items would be assumed to have equal item parameters, the linking error would be zero). This shift from difficulty to mean is said to better reflect the uncertainty when comparing country/economy means across PISA cycles (Robitzsch and Lüdtke, 2022[143]).

386. Finally, although large-scale assessment methodology is an active area of research in the era of big data (von Davier et al., 2019[132]), machine learning (Immekus, Jeong and Yoo, 2022[147]), and process data (Han, He and von Davier, 2019[148]), it is important to find a balance between maintaining trend with previous cycles and introducing innovations with new cycles of PISA. For example, since PISA moved from a paper-based assessment to computer-based assessment in 2015, additional data (process data) has become available and has been published as part of the international database. These new data should be properly accounted for in the creation of plausible values. For this reason, for example, response time was used in the conditioning model in PISA 2018 (Shin, Jewsbury and van Rijn, 2022[149]). Given the theoretical nature of many of the topics of the standards in this section, the Technical Advisory Group (TAG) has an important role in determining the established procedures for constructing and validating scales for the cognitive tests.

Established procedures refer here either to a well-known statistical method or a special method which is supported by the PISA TAG in the given context.

8.3.2. Standards

Standard 8.3.1. The data yield and basic quality of the cognitive tests in the Main Survey shall be evaluated in all participating countries/economies, and the proper functioning of items shall be verified at national and international levels.

387. Note: This evaluation includes examining data analyses related to quality control, response rates, item statistics, coding reliability, and timing. Furthermore, checks should be implemented to evaluate the matrix sampling and multistage adaptive testing designs.

Standard 8.3.2. Item parameters shall be estimated using the Main Survey data with established procedures and the fit of the IRT model shall be evaluated in all participating countries/economies and selected country-by-language groups.

388. Note: The evaluation of IRT model fit involves item fit, dimensionality, local independence, and position effects. Groups defined by the combination of country/economy and language should be used to evaluate item fit. In the scaling approach used in recent cycles, item fit is evaluated for country-by-language groups given a minimum number of student responses. If misfit is detected, unique item parameters are estimated for the group under scrutiny (for indication, the current limit to the allowed number of country-specific item parameters is of 1/3 of items). For trend items, this uniqueness can pertain to both international and unique item parameters from the previous PISA cycle. For new items, this uniqueness can only relate to the international item parameter from the current PISA cycle. Furthermore, if the misfit of an item is in the same direction for multiple country-by-language groups, unique item parameters would be estimated for this set of groups instead of for each group separately.

Standard 8.3.3. Invariance of item parameters across countries/economies shall be evaluated for all domains with an established procedure.

389. Note: Item parameters determine the shape of the item response function that describes the relation between the probability of a correct item response and the latent proficiency. Invariance of item parameters across countries/economies can thus be defined as the property that these items' response functions are the same across countries/economies while allowing differences between countries/economies in the proficiency distribution. Such invariance is needed to establish the international PISA scale so that it can be used to compare countries/economies in terms of proficiency. Depending on the IRT scaling method, invariance of item parameters could be evaluated using, for instance, partial invariance methods (von Davier et al., 2019[132]) and alignment methods (Muthén and Asparouhov, 2014[150]).

Standard 8.3.4. Measurement precision of the cognitive tests and test reliability shall be evaluated for all domains and reported subdomains across countries/economies.

390. Note: Depending on the IRT scaling method, measurement precision can be evaluated by comparing test information functions for each domain to student proficiency distributions. Information functions can be computed for individual items as well as sets of items (e.g. units, test forms, trend items). Test reliability can be determined by computing test reliability for each domain and subdomain in each country/economy. Issues

with test reliability (e.g. lower values than anticipated) can be due to a mismatch between the test information function and the student proficiency distribution. Furthermore, the source of such issues could also be pinpointed by inspecting information functions of individual items or sets of items.

Standard 8.3.5. Subscales for the core domain shall be evaluated with respect to their measurement properties.

391. Note: Methods for evaluating the added value of subscores beyond overall proficiency scores can be applied.

Standard 8.3.6. The process with which conditioning variables are created from the background questionnaire shall be documented.

392. Note: In particular, how the items in the background questionnaire were coded (e.g. dummy-coded, contrast-coded) should be documented, so that the resulting conditioning variables can be reproduced from the original data. If additional data such as process data is used, how this data was transformed into conditioning variables should be documented.

Standard 8.3.7. The parameters of the population model shall be estimated using the Main Survey data with an established procedure and the fit of this model should be evaluated in all participating countries/economies.

393. Note: All parameters in the population model are specific to each country (i.e. the latent regression models are estimated country by country).

Standard 8.3.8. Plausible values that describe the distributions of student proficiency in each reported domain and subdomain shall be created with an established procedure and their reliability shall be evaluated in all participating countries/economies.

394. Note: Plausible values' (PV) reliability is different from test reliability in that it represents the information in both the item responses and the conditioning variables. That is, PV reliability is expected to be higher than test reliability. It can be evaluated by inspecting the correlations among PVs.

Standard 8.3.9. Linking error between the current and previous PISA cycles shall be estimated for the domains of reading, mathematics, science, and the optional domains using an established procedure.

395. Note: Since 2015, linking error, which is relevant for evaluating trends in proficiency, has been estimated by the standard deviation of the equated country/economy means from two calibrations. Although linking error currently focuses on the mean, it can be computed for other statistics of interest as well (e.g. percentiles). Furthermore, it could theoretically be determined for each country/economy individually, and could also be determined within a single PISA cycle across countries/economies. The latter could be important with multistage adaptive testing because different countries/economies use the item pools differently (e.g. students in higher performing countries more often see harder items than students in lower performing countries and vice versa).

Standard 8.3.10. Enhancements and innovations in the methodology related to scaling and conditioning shall be documented, reviewed, and approved by the PISA TAG and OECD.

396. Note: The PGB should be informed of any of any enhancement or innovations that PISA TAG and OECD approved.

8.3.3. Quality Assurance

- The responsible contractor produces analysis plans describing the steps in creating and validating scales from the cognitive tests using the Main Survey. The contents of these analysis plans are to be presented to the OECD, the PGB, the TAG, and the Subject-Matter Expert Group (SMEG), so that comments and questions can be addressed. The final analysis plans need to be approved by OECD. These plans should contain:
  - The steps in the proposed analyses;
  - The deliverables;
  - The timeline.
- Chapters in the PISA Technical Report should describe the steps for scaling the cognitive tests. These chapters should contain (see Standards 9.3):
  - Analyses of data yield and quality;
  - Description of the psychometric scaling models used, calibration and scaling, as well as the population modelling and multiple imputation procedures employed to obtain plausible values for student performance;
  - Presentation of how these models were applied to the given cycle data and their outcomes;
  - Results of the psychometric scaling and population modelling;
  - Correlations between domains (by country/economy) and domain sub-scales where applicable, as well as the percentage of respondents at each proficiency level (by country/economy).

8.4. Standards on constructing and validating scales from questionnaires

8.4.1. Rationale

397. The PISA questionnaires for students, parents, teachers, and school administrators are developed to provide policy-relevant context to achievement outcomes as well as provide policymakers with other variables relevant to monitoring student learning, well-being, and preparedness for lifelong learning. While some questions are developed as meaningful reporting variables without further aggregation or analysis, the currently employed approach for the large majority of PISA questionnaire items involves the creation of derived variables (DVs) for reporting that are each based on multiple items or on recoded data for individual items. Previous PISA cycles have introduced two types of standard DVs, namely (1.) indices based on Item Response Theory (IRT) (also referred to as "scaled indices") and (2.) indices derived based on direct calculations without using IRT (also referred to as "simple indices"). The index of Economic, Social, and Cultural Status (ESCS) constitutes a third type of DV that combines IRT scaling with other statistical data aggregation methods.

398. Constructing and validating scaled indices involves multiple steps not limited to the final phase of Main Survey analyses. The process to construct valid scales representative of relevant constructs aligned with framework definitions starts with (a) initial development of a sufficiently large item pool for each construct (see Standards 6.5 and 7.3), followed by (b) a first reduction and/or refinement of the item pool based on pre-testing findings from small-scale cognitive interviews (see Standards 7.3), (c) a large-scale empirical evaluation of the envisioned scales for select countries via the international Field Trial (see Standards 8.1), and (d) scaling of questionnaire constructs based on large-scale Main Survey data for all countries.

399. The standards outlined in this chapter focus on the steps recommended for phase (d), most of which apply to phase (c) as well. In particular, the standards below define the procedures that should be used for constructing indicators and scales from the questionnaire data, including methods for deriving trend variables and for excluding groups that either have insufficient data quality or low comparability with other groups. These standards also provide guidance on approaches that should be used for validating questionnaires using the data from the main data collection, including investigating the internal consistency of each scale within and across countries/economy-by-language groups, and analysing the invariance of item parameters across countries and languages. Ensuring that these standards are followed is an important condition for the creation of a robust international data base suitable for policy interpretations.

8.4.2. Standards

Standard 8.4.1. The data yield of the questionnaire data in the Main Survey shall be evaluated in all participating countries/economies.

400. Note: This evaluation should include checks to evaluate the within-construct matrix sampling design, any questionnaire routing rules, and missing rates due to non-reached items. Appropriate missing values should be assigned to each of the different types of missing questionnaire responses (e.g. no response, valid skip, not reached, missing by design).

Standard 8.4.2. The basic quality of the questionnaire data in the Main Survey shall be evaluated in all participating countries/economies, and the plausibility of observed frequency data across question response options as well as the ranges of fill-in responses shall be verified at national and international levels.

401. Note: This evaluation should include examining data analyses related to quality control, response rates, item response category frequencies, coding of any open-ended responses, and timing data. See also standard 8.4.4 for additional data quality checks regarding invalid response patterns.

Standard 8.4.3. Any necessary recoding of questionnaire data shall be applied prior to scaling questionnaire constructs.

402. Note: Recoding should aim at ensuring that data values for all items for a given construct can be interpreted in a consistent way (e.g. in ascending or descending order). For example, “yes/no” questions should be coded so that yes receives the value 1 and no receives the value 0. All items in mixed-valence scales (i.e. scales where responses to some items are related positively to the underlaying construct and other items are negatively related to the underlaying construct) should be recoded so that the data represents the same valence.

Standard 8.4.4. Response data flagged for invalid response patterns shall be excluded when creating scaled indices.

403. Note: This evaluation should include examining of extreme straightlining for mixed-valence scales (i.e. choosing the most extreme response options to the left or right of the scale consistently across all items, regardless of their valence) and potentially other invalid response patterns (e.g. rapid responding). Any plans to apply data filters prior to analyses should be reviewed and approved by the PISA Technical Advisory Group (TAG). Raw data for all respondents should be retained in the international data files (see Standards 9.2).

Standard 8.4.5. Item parameters shall be estimated using the Main Survey data with established procedure and the fit of the multi-group IRT model shall be evaluated in all participating countries/economies and country-by-language groups.

404. Note: The phrase "using an established procedure" means that either a well-known statistical method is used or that a special method is used which is supported by the PISA TAG in the given context. Unless there are specific reasons for a particular construct, a generalised partial credit model (GPCM (Muraki, 1992[151])) should be applied for polytomous data and a two-parameter logistic (2PL; (Birnbaum, 1968[152])) model for dichotomous data, which have been the standard IRT models for questionnaires since the 2015 cycle. The evaluation of IRT model fit should involve item fit, dimensionality, and local independence. Groups defined by the combination of country/economy and language should be used to evaluate item fit.

Standard 8.4.6. Invariance of item parameters across countries/economy and language groups shall be evaluated for all scaled indices with an established procedure.

405. Note: Root mean square deviation [RMSD] model-data item fit statistics or a similar method should be used to assess item-by-group interactions and only items that fall below a predetermined threshold should be used in final scales. The same or stricter thresholds for invariance than in previous PISA cycles should be used.

Standard 8.4.7. Scaled indices shall be created only for constructs with an appropriate number of items ensuring construct representation consistent with published theories about the respective construct.

406. Note: While the number of items needed to adequately represent a latent construct may vary depending on the type of construct, a minimum number of five items per scaled index should be targeted. Although a minimum number of three items is sometimes mentioned to provide minimum construct representation in the factor-analysis literature (e.g. (Hair, 2010[153])), the minimum of five items is targeted in PISA. This is done to have more than minimum construct representation, to be able to do dimensionality analysis, and to prepare for situations in which misfitting items need to be dropped from the scaled index. Misfitting items can be dropped just for a particular country/language group: scaled index scores can be compared across countries as long as there is a minimum of 3 items scaled together with other countries.

Standard 8.4.8. Overall reliability and measurement precision shall be evaluated for all questionnaire scales within and across countries/economy by language groups.

407. Note: Multiple indicators of reliability should be considered, including those based on classical test theory (e.g. Cronbach's Alpha) and marginal reliabilities based on IRT scaling. Measurement precision across the range of each construct can further be evaluated by comparing scale information functions to distributions of person parameter estimates (i.e. thetas) for each scale.

Standard 8.4.9. Trend analyses should be considered for any scaled indices with three or more items that were also administered in previous PISA cycles.

408. Note: Items identified as trend items should be used as linking items. Whether trend can be reported should be evaluated by checking whether three or more items can be identified as linking items that function in the same way in the current cycle as they did in the previous cycle. These checks should be based on the analysis of item parameter invariance across cycle-by-country/economy-by-language groups. To convert scale scores to a meaningful reporting metric, linear transformations should be implemented to ensure that the interpretation of scores is directly comparable to that of the previous cycle(s). The transformation constants should be documented in the technical report. Where required, certain components of scaled indices of previous cycles may be reconstructed in order to ensure comparability across cycles (e.g. as was done for the ESCS due to evolutions in the ISCED and ISCO coding systems over the years).

Standard 8.4.10. Correlations with other related questionnaire constructs as well as relevant outcomes should be considered for additional validity evidence. Correlations should be evaluated based on the total sample, for each country/economy by language group, and at aggregated country/economy by language group-level.

409. Note: Correlations with other questionnaire constructs should at minimum include all scaled indices from the same module per the questionnaire framework. Correlations with outcomes should include achievement scores (PVs or appropriate proxies) as well as alternative outcomes indicative of student well-being (e.g. subjective life satisfaction). Observed relationships inconsistent with theories should be flagged for further investigation.

Standard 8.4.11. Scale scores for the newly constructed scales should be converted to a more convenient reporting metric using linear transformation, so that the resulting distribution across OECD countries has a mean of 0 and a standard deviation of 1.

410. Note**:** In relation with Standard 8.4.9, it is worth noting that trend scales are on the (0,1) distribution of the OECD countries of the initial cycle that the scale was constructed in. This means that they may not have a mean of 0 and standard deviation of 1 in the current cycle.

8.4.3. Quality Assurance

- The responsible contractor produces Field Trial (See Standards 8.1) and Main Survey analysis plans for constructing and validating the scales from the questionnaires. The contents of the Field Trial and Main Survey analysis plans for the questionnaire are to be presented to the OECD, PISA Governing Board (PGB), National Project Managers (NPM), and the relevant technical advisory and expert groups (e.g. TAG, Questionnaire Expert Group (QEG), subject-matter

expert groups on a need base), so that comments and questions can be addressed. The analysis plan needs to be approved by OECD. These plans should contain:

- The steps in the proposed analyses;
- The deliverables;
- The timeline.

- The Technical Report should document (see Standards 9.3):
  - The main steps in constructing and validating the scales from the questionnaires;
  - Data yield and data quality summary tables;
  - Summaries of coding reliability and the performance of automated scoring methods;
  - A summary of IRT model fit;
  - A summary of additional preparatory analyses.

8.5. Standards on constructing proficiency levels and descriptors

8.5.1. Rationale

411. The proficiency levels are a central element of reporting. Although the proficiency scales in PISA are to be considered as continua, the ordered categorical proficiency levels enable readers to understand more concretely what the PISA results uncover in terms of what students know and can do. The proficiency levels are defined by setting cut scores on the continuous PISA scale. The number of proficiency levels increased from five (1, 2, 3, 4, 5) in PISA 2000 to eight (1a, 1b, 1c, 2, 3, 4, 5, 6) in PISA 2018 due to the need for describing students' abilities at the extreme levels of proficiency. Given the importance of proficiency levels for reporting, it is important that the descriptions of knowledge and skills at each proficiency level accurately represent what students know and can do.

412. The proficiency levels in PISA are constructed using standard setting methods (Blömeke and Gustafsson, 2017[154]; Cizek, 2012[155]). The specific method employed in PISA is known as scale anchoring (Beaton and Allen, 1992[156]), which involves two main components. In the statistical component, items that discriminate between proficiency levels are selected using psychometric properties. In the consensus component, the selected items are used by experts to provide descriptions of what students at the different levels know and can do. These descriptions should be comparable over time for the core domains, but are generally updated when the framework of a domain is refreshed.

413. Proficiency levels for the core domains have been defined in previous PISA cycles where each domain was a major domain for the first time. An important preliminary step in the process of defining the proficiency scale is the item classification based on the framework specifications by test developers as part of new item development. In this classification, items are organised according to relevant dimensions of the assessment framework, including subscales if applicable. Then, the exact definition of the proficiency levels, which has subjective elements, is largely determined by choosing values for the three following variables: 1) the success probability of students at a particular proficiency level on a test consisting of items at that level; 2) the width of the proficiency level; 3) the probability that a student in the middle of a level would correctly answer an item of average difficulty for that level. An important role is played here by the response probability (RP) value, which is the value on the proficiency scale associated with

a particular success probability for an item. For example, the RP62 of an item indicates the scale point at which a student has a 62% chance of success. Some additional calculations may be needed because of differences in IRT models across PISA cycles (i.e. without/with item discrimination parameters). Apart from the approach described above, a lower limit is to be determined for the lowest described level, below which no meaningful description of proficiency is possible. In PISA, the floor of the lowest described level is set so that it has the same range as the other described proficiency levels.

414. For the core and existing optional domains that already have established proficiency levels, the levels are generally carried forward from one PISA cycle to the next through the IRT linking of the PISA scale across cycles. That is, when the item parameters and the linear transformation from the IRT scale to the PISA scale for the new cycle are established, the cut scores for the proficiency levels can be applied in the same way as in previous cycles. However, linking error can arise due to changes in items and item parameters across PISA cycles, which is relevant to assess the significance of trends in the distribution of proficiency levels across cycles.

415. For the innovative domain, the proficiency levels are newly developed for each PISA cycle. The experts for the innovative domain should generally define the factors that drive difficulty across the levels of proficiency. They also set the cut scores along the scale that defined each proficiency level of performance.

416. To help the understanding of the proficiency scale of the major domain, item maps are often created by sorting a selection of items from easiest to most difficult and providing brief descriptions for each item together with their position on the scale (Huynh, 1998). With such a map, factors associated with item difficulty can become visible and how they may relate to different dimensions of the framework and identified subscales.

417. The standards in this section define the substantive analysis of PISA assessment items used to develop the proficiency-level descriptions, and the criteria that should inform decisions on the number and width of the levels, methods used to set cut scores, and the identification of the lowest level below which no meaningful description of proficiency is possible. The standards also define the role of the subject matter expert group in the validation of the proficiency-level descriptions.

8.5.2. Standards

Standard 8.5.1. All items shall be classified according to the specifications of the framework for each domain.

418. Note: Newly developed items (e.g. for the major domain and the innovative domain) should be classified by test developers based on the framework specifications. For trend items in the major domain, item classifications should be reviewed and revised as needed to correctly reflect the updated framework. All classifications are to be reviewed by the relevant subject matter expert groups. For minor domains consisting of trend items and an unchanged framework compared to the previous cycle, a review of the item classifications is not necessary, unless issues are noticed.

Standard 8.5.2. The proficiency scales shall be defined by subject matter experts using descriptors of the factors that drive the difficulty of items for each domain.

419. Note: Relevant subject matter expert groups play a key role in creating these descriptors and should work with the PISA international contractors and the Secretariat to review and revise the sets of descriptors for the scales and subscales in each domain.

The factors that drive the difficulty of items for each domain may come from theoretical considerations (e.g. those guiding item developers) but may also be validated empirically, after scaling Field Trial and/or Main Survey data. The latter kind of analysis may also end up uncovering unintended drivers of difficulty (e.g. response format, test design), which may not be wanted in defining the proficiency scales.

Standard 8.5.3. Relevant dimensions of the framework shall be used for defining meaningful subscales.

420. Note: The assessment framework may support subscales based on the various dimensions of the framework. Where subscales are included, they should be based on the domain framework, be meaningful for reporting purposes and be defensible with respect to their measurement properties. Measurement properties of subscales can be evaluated, for instance, using established methods for evaluating the added value of subscores beyond overall proficiency scores ((Haberman, 2008[145]); see also the standards on constructing scales from the cognitive tests). Relevant subject matter expert groups play a key role in identifying subscales from framework dimensions and should work with the PISA international contractor to define the subscales in each domain.

Standard 8.5.4. Proficiency levels for newly assessed domains as well as changes and/or additions to existing proficiency levels should be set using an established procedure approved by the Technical Advisory Group (TAG). Changes or additions to existing proficiency levels shall be approved by the PGB.

421. Note: Proficiency levels are defined by choosing values for three variables. First, the success probability of students at a particular proficiency level on a test consisting of items at that level. For example, this can be chosen by computing the success probability of a sample of students from a uniform distribution of proficiencies from a particular level on a testing consisting of items with a uniform distribution of difficulties from that level. Second, the width of the proficiency level is to be set. However, since there is a non-linear relationship between the proficiency scale and the success probability, the interpretation of the width can vary across the scale. For example, a change of 80 PISA points in the middle of the scale can relate to a different change in success probability than a change of 80 points in the lower and upper tails. Third, the probability that a student in the middle of a level would correctly answer an item of average difficulty for that level is to be chosen. In addition, a lower limit is to be determined for the lowest described level, below which no meaningful description of proficiency is possible. In PISA, the floor of the lowest described level is set so that it has the same range as the other described proficiency levels. Depending on the domain, different approaches to defining proficiency levels may better align. However, new procedures for setting proficiency levels would have to be approved by the TAG. In addition, changes and/or additions to existing proficiency levels such as adding new levels or splitting existing levels need to be empirically supported by analysis of Field Trial and/or Main Survey. Such updates of proficiency levels should be developed and supported by the subject matter expert group and the associated methodology should be approved by the TAG. Such changes or additions should also be approved by the PGB.

Standard 8.5.5. An item map shall be created for the major domain consisting of a description of a selected number of released items, the associated scale values of their difficulty, and the associated proficiency levels.

Standard 8.5.6. Summary descriptions, including task characteristics, for each of the proficiency levels shall be provided for the major domain and the innovative domain.

422. Note: To do so, items that discriminate between proficiency levels should be selected using psychometric properties. Subject-matter experts should then use the selected items to provide descriptions of what students at the different levels know and can do. This should be done every cycle, since it concerns the major domain (updated item pool) and innovative domain (new item pool).

8.5.3. Quality assurance

- The Technical Report should contain a chapter describing the procedures followed to set proficiency levels. Furthermore, it should contain the item map for the main domain, the PISA scale intervals associated with each proficiency level, and summary descriptions of the proficiency levels on the overall scale as well as identified subscales (see Standards 9.3).

Chapter 9. Reporting of PISA data

9.1. Standards on producing and reviewing primary international reports of findings

9.1.1. Rationale

423. Along with the PISA database, the primary international reports produced by the OECD on the findings of the PISA study constitute the main output of the PISA programme. The primary reports are intended to provide sufficiently timely and broad access to the results of the PISA study and discuss them while putting them into context. As such, national and international institutions and stakeholders, policy makers, and media use the primary reports as the principal, and often the only, entry point to the PISA database. The primary international reports provide an analysis of the PISA data in order to describe the landscape of educational systems of all participating countries and economies, presenting and comparing country statistics in light of policy questions focusing on quality of education and its role in promoting and supporting lifelong learning of 15-year-olds.

424. The reports function as a bridge between the information captured by the test and the contextual questionnaires and the decisions or actions of the users or readers of the reports (Zapata-Rivera, 2011[157]; Zapata-Rivera and Katz, 2014[158]). Readers of the reports, including education stakeholders and decision makers, interpret text, tables, and figures as indicators of skills, abilities, attitudes and characteristics of students and schools, and they use them for orienting their decisions.

425. PISA primary reports therefore shoulder the responsibility for supporting accurate user interpretation and use of results (O'Leary, Hattie and Griffin, 2017[159]). They contribute to the validity of PISA by guiding readers to understand the results and develop appropriate interpretations of students and systems' performance (see Chapter 2 on Validity). To this end, the PISA reports should provide a scientifically rigorous presentation of the results, and help readers interpret the results to develop valid insights for education policy. Where policy recommendations are provided, the evidence in support of the causal claims underlying these recommendations, and its limitations, should be presented transparently. Such evidence cannot be limited to associations identified in PISA results, but may include, for example, evidence from prior experimental or quasi-experimental studies, as well as theoretical arguments.

426. Establishing comparability of the results across the many national settings of PISA and across languages is a key concern of the PISA study (see Chapter 4). This concern needs to be reflected in the primary international reports. Thus, international comparisons in the initial reports should be limited to measures with demonstrated international comparability. Other measures that do not meet strict comparability criteria can be included in the reports while avoiding international comparisons (for example, showing gender differences across aggregated data for all countries). Comparability considerations should also influence the reporting of differences within countries. For example, if there are reasons to believe that a questionnaire item might function differently for boys and girls then gender comparisons for the items should not be reported. The reports should therefore present meaningful comparisons between countries and within countries, by main socio-demographic characteristics of students and schools.

427. Another key concern in developing the reports is ensuring fairness (see Chapter 5). Fairness in primary international reports manifests in two ways: in terms of coverage

of content in the reports, and in terms of presentation of the results and accessibility of the reports. As developed in Chapter 5, ensuring fairness in PISA should involve avoiding analyses and formulations which may facilitate unfair interpretations and carefully considering what results are emphasised.

428. The international reports should go beyond international rankings of average scores, providing a description of the quality of education systems that reflects what PISA measures as described in the PISA frameworks. In particular, it is important that the reports provide an accessible description of variations in achievement within each country, for example explaining how students are distributed across the different proficiency levels and how much of the variation in scores is explained by school characteristics. The reports should investigate differences between groups or types of students in the access to the resources and processes they need to succeed. The reports should also describe differences between and within countries in the school environment and school processes that promote and support quality of teaching, students' engagement, and learning. Finally, they should also cover social and emotional outcomes, including students' well-being, and their interplay with performance in cognitive tests as well as their relationship with the teaching and learning environment.

429. This section presents the standards that should guide the development and publication of the PISA international reports, covering what processes should be followed, what information and analyses should be included in the reports, and how this content should be presented.

9.1.2. Standards on the initial reports' development process

Standard 9.1.1. The international reports should be drafted and verified on a short timeline to ensure timeliness.

430. Note: International initial reports should be made public within 2 years from the data collection of the study (taking into account one year dedicated to data collection and 6 months for data processing, this means that the report should be developed in less than 6 months). The reports are the main source of information on the cycle's results, and they shall be ready to accompany the data release and guide a diverse audience to understand the main findings of the study. In-depth, thematic analyses can be published afterwards to complement the initial reports.

Standard 9.1.2. The outline of the reports and plan for main analyses shall be determined several months in advance of publication and be reviewed by experts from participating countries and economies.

431. Note: The full reporting plan for a PISA cycle should be coherent and communicated in advance. The plan should be consistent with the assessment goals set by the PISA Governing Board. The proposed outline of the reports and main analyses should be presented in an analysis plan. National experts from participating countries and economies should be provided with opportunities to review the analysis plan.

Standard 9.1.3. Experts from all participating countries and economies shall have an opportunity to review drafts of the reports under appropriate confidentiality conditions.

432. Note: National experts from participating countries and economies shall be provided with multiple opportunities to review the drafts of the reports at different stages.

Countries' comments should be incorporated into the reports by the Secretariat, and justifications should be provided when comments cannot be addressed. Country notes accompanying the reports should also be reviewed by experts from participating countries and economies. This should be done while guaranteeing the confidentiality of the draft versions of report chapters, of the data used for the analyses, and of the full dataset (see also Standards 9.2).

Standard 9.1.4. International reports shall be made available on the OECD PISA website to be downloaded, printed or reproduced in ways that increase ease of use and that minimise the costs of reproduction.

433. Note: Tables and graphs should be downloadable in widely available file formats (e.g. PDF for the reports, Excel sheets for the tables) and be reproducible using widely available photocopying techniques without losing quality in terms of accuracy, interpretability, and legibility.

Standard 9.1.5. Online versions of the international reports should include infographics and tools for interactive data visualisation.

434. Note: The interactive tool should enable readers to access a wider range of analyses while keeping the international primary reports trimmed and focused on fully comparable content. For instance, it can give readers an opportunity to produce trend analyses for specific countries on selected indicators. The interactive tool could also allow for basic sub-group comparisons on time trends.

9.1.3. Standards on the initial reports' content

Standard 9.1.6. The reports shall contain all the necessary information readers need to understand the results and draw appropriate inferences.

435. Note: The reports should contain an introductory chapter summarising the key characteristics and purposes of the PISA study. The presentation and the discussion of the results should accurately reflect what the assessment measures. In order to support understanding of technical aspects of the study, the reports should include information on:

- The purpose of the study and the target age of the student population;
- Examples of items used (released items) and simple presentation of scoring procedures;
- Basic information on procedures for data collection (e.g. mode of administration);
- User friendly, non-technical explanation of scaling methods;
- Definitions of measurement scales, ranges, and sufficient details on their construction;
- Notes and simple description of analysis procedures;
- How to interpret differences in PISA scores, including discussions of statistical and practical significance and a description of the sources of uncertainty in PISA.

Standard 9.1.7. The report shall always indicate the uncertainty associated with statistical estimates.

436. Note: Standard errors or confidence intervals for the statistics should always be presented with the results (e.g. in the same table). Row and column totals should be reported in tables where appropriate.

Standard 9.1.8. Country-level statistics shall be disaggregated into subgroups when the sampling design supports it and sampling standards are met. Primary reports should present country-level results, followed by a disaggregation of country statistics by key student and school characteristics.

437. Note: All analysis of test scores and questionnaire results should be disaggregated by gender and socio-economic status of the students. Additional sub-groups may be relevant for specific analyses, for example: students' immigrant status, language, or school private-public status or location.

Standard 9.1.9. Data quality indicators shall be reported along with the results.

438. Note: Quality indicators showing the extent to which sampling (including population coverage), translation, scaling or test-administration issues affect the validity of comparisons should be developed. Both the quality of country data and quality of measures which do not meet comparability and reliability requirements should be acknowledged and made visible in an appropriate way, for example in the notes to the figures or tables, or in a reader's guide at the beginning of each report. When data quality issues are identified, the OECD Secretariat should decide whether the data should be reported based on the data adjudication group's recommendation.

Standard 9.1.10. PISA results should be contextualised by presenting the factors that might explain the observed differences across countries and economies and across time.

439. Note: Whenever possible, PISA results should be contextualised by providing information about concurrent changes or differences which are relevant to the results. For instance, PISA results should be interpreted considering differences in how education is organised across grade levels, differences in the share of students who do not speak the language of instruction at home, differences in per-capita GDP (avoiding any causal claim, see Standard 9.1.11).

Standard 9.1.11. The description and interpretation of results shall be scientifically rigorous, avoiding claims that are not supported by the data.

440. Note: Statistical significance should be reported, and both statistical and practical significance of the results should be considered in the reporting and discussion of results. The text of the reports should emphasise the distinction between statistical and practical significance of the results. Causal claims which are not supported by appropriate identification strategies based on experimental designs, or on quasi-experimental identification strategies supported by credible assumptions, shall be avoided. Instead, the reports should only refer to the observed relationships as associations, co-relationships, or similar word. discussion should reflect on the potential competing explanations of the results. Reporting and discussion of results should accurately reflect what the assessment measures - as presented in the frameworks and supplemented by validation studies -, avoiding extrapolations.

Standard 9.1.12. Communication material accompanying the primary international reports shall be consistent with their content.

441. Note: Additional material released at the same time as the primary international reports for communication purposes, such as country notes and presentations, should make the results more understandable and accessible to a wider audience, while preserving the integrity of the interpretation. Simplification of the messages should not be at the cost of altering meaning.

9.1.4. Standards on the initial reports' format

Standard 9.1.13. Findings from international reports shall be made available to different audiences in the formats most useful for those audiences.

442. Note: The findings from the international reports should be represented and shared in different formats to amplify access and visibility. They should be available as summaries, extracts, downloadable tables and graphs, and short thematic reports with country profiles.

Standard 9.1.14. The international reports shall be written simply and clearly. Whenever possible, non-technical terminology should be used.

443. Note: The international reports should be designed and drafted with a general audience in mind. The information should be presented in simple, non-technical language, and should refer to technical reports for details. The reports should include only few and simple explanations of the main technical features of the study.

Standard 9.1.15. Tables, graphs and figures should be used to present qualitative information clearly.

444. Note: Whenever possible, complex information should be organised and presented visually in order to enhance understandability. For instance, tables, graphs or figures can be used to present the steps in program implementation, aspects of sample attrition, and domain coverage of the tests and contextual questionnaires.

Standard 9.1.16. Tables, graphs and figures shall be clearly labelled and annotated.

445. Note: Tables, graphs and figures should include a short title that concisely states their subject. Labels should be provided for the names of variables and for their categories. The data source for each table and graph should be identified (produced based on OECD PISA data or from other studies, providing their name and date). When the data presented come from multiple sources or from a source that is not the direct subject of the report, the source note shall clearly match the data or objects with the source.

Standard 9.1.17. The presentation of graphics and figures shall facilitate the correct interpretation of data.

446. Note: Cluttered graphs should be avoided. The usage of acronyms should be minimised, and clear labels should be used for the horizontal and vertical axes. Deviations from ordinary representations of vertical and horizontal axes should be explained.

Standard 9.1.18. The key technical procedures should be presented in boxes or in annexes.

447. Note: For instance, principles of scaling and analytical techniques that are important for understanding the results should be explained in boxes or annexes, rather than in the main text, referring the reader to the technical report for further details.

Standard 9.1.19. The use of footnotes should be minimised. Endnotes should be used sparingly.

448. Note: The usage of footnotes should be limited to explaining exceptional instances that may appear as inconsistent within or between tables. Endnotes should never include essential information for the understanding of the main argument in the body of the chapter.

Standard 9.1.20. Assessment material should be presented with the primary reports. When the assessment itself is delivered digitally, it is important to be as faithful as possible to the student experience when material is presented.

449. Note: Released test items and the original contextual questionnaires should be in an online appendix or made available as supplementary materials of the primary reports. Readers shall be able to make direct reference to them and always consult them while reading the findings of the study. The digital testing experience should be described accurately for readers to understand.

9.1.5. Quality assurance

- The responsible team produces a publication and reporting plan which should be presented to and reviewed by the PISA Governing Board (PGB). The publication and reporting plan establishes the targeted audiences and the different formats of publications that will accompany the international reports. The plan also schedules the analytical and writing activities involved in the drafting and publishing of the reports. The schedule should:
  - Include information on the time that will be dedicated to the drafting and revisions of text, tables, and graphs, including when the draft will be sent for review to the PGB;
  - Allocate sufficient time for consultation with countries and for subsequent revisions.
- The responsible team reports to the PGB indicating how countries' reviews on the publication and reporting plan were addressed.
- A first draft and a final draft of the initial reports are submitted for comments to all participating countries and economies.
- The responsible team reports to the PGB indicating how countries' reviews on the international report drafts were addressed. The responsible analysis/reporting team maintains internal documentation of the following elements. This documentation should be archived for future reference by future analysts:
  - Innovative approaches that were experimented during the analytical and reporting phases;
  - Main challenges experienced and suggested remediation for the next cycle of reporting;

- A record of main comments received by the countries and general recommendations to consider for reporting in the next cycle.

9.2. Standards for releasing data and assessment material

9.2.1. Rationale

450. PISA is committed to providing trustworthy, accessible and transparent information. Making the PISA data and assessment material accessible to the public is fully part of PISA's mission to inform educational policy practices. The rich amount of data collected in PISA can tremendously benefit the education and research community by enabling secondary analyses that go beyond the results of the initial reports. Moreover, making non-confidential information available and ensuring transparency is a way to ensure the public accountability of PISA. At the same time, the dissemination of data and material should be done in compliance with the highest data protection and confidentiality standards (covering the confidentiality of trend items), and without compromising the comparability of PISA.

451. Ensuring accessible and transparent information implies that the raw data should be shared along with clear and exhaustive metadata, explaining which data was collected, how it was collected, and how it was processed to produce the resulting data files and results. This enhances the validity and trustworthiness of PISA by i) ensuring that the data is appropriately and validly used and interpreted by external users, and ii) enabling the reproduction of the scales and results displayed in the PISA reports.

452. The released data should go through extensive quality checks in order to ensure that all the data files are complete and not corrupted, that there is full consistency between labels in the data and the codebooks, and that no confidential information is released.

453. Because of PISA's nature as a cyclical programme which establishes time trends, not all material can be disclosed. In particular, cognitive test items which are intended to be reused in future cycles should be kept confidential in order to prevent training to the test. At the same time, it is important that some exemplary items are made available so as to provide concrete examples of how the target cognitive processes are assessed to PISA data users and PISA reports readers. Thus, only a selection of items, which will not be reused, should be released.

454. Finally, in line with PISA's core principle of fairness as presented in Chapter 5, it is crucial that PISA data and material are shared in accordance with the highest ethical standards. This implies ensuring a full compliance with data protection regulations. Following the AERA, APA and NCME standards (AERA, APA and NCME, 2014[1]) and countries' data security and confidentiality laws (e.g. the European Union's General Data Protection Regulation – GDPR), the OECD should maintain data security protocols to protect confidentiality. PISA should ensure that the data made public does not contain any information which could be used to identify students or schools, or provide sensitive information on them. Data from cognitive laboratories using video-recording or eye-tracking devices should similarly be treated very carefully.

455. The standards in this section aim to establish best practices in ensuring full transparency and accessibility of PISA products while safeguarding the confidentiality of cognitive test items and of participants' personal information. They cover what content should be made available or should be kept confidential, in which formats it should be made available and with what supporting information.

9.2.2. Standards

Standard 9.2.1. **The public use data files shall contain all responses and a selection of process data, with the exclusion of variables that were not collected or derived for all countries and of variables which might be used to identify or disclose sensitive information about schools or participants.**

456. Note: The public use data files containing the raw data from the student cognitive assessments, and core and optional questionnaire assessments, should be made available for download on the OECD website. Separate files should be provided for the student questionnaire, the cognitive test, the school questionnaire, and the teacher questionnaire. These files should not include data that were not collected or derived for all countries, thus excluding national adaptations and extensions. In addition, variables which could be used to identify students, schools or teachers or reveal sensitive information (e.g. school location and name) should not be released in the public datasets. This concerns responses and identifiers, responses to open-ended questions, but also some type of process data which could be used as biometric data. However, such excluded variables could be made available to interested researchers upon request and submission of a research project, under a data confidentiality agreement and following the approval of the PISA Governing Board. Details on which variables are included or excluded should be provided in the accompanying documentation.

Standard 9.2.2. **Participating countries and economies shall be given the opportunity to suppress variables from their national data in the public use files when it significantly increases the risk of reidentification for students, schools or teachers.**

457. Note: Suppressed data should be represented in the database by means of missing codes. Countries shall make available the largest permissible set of information at the highest level of disaggregation possible.

Standard 9.2.3. **Any issue regarding data quality that has been identified during data adjudication shall be acknowledged in the document, including recommendations on how to use data that are flagged for quality issues.**

458. Note: The data adjudication is done on the basis of the Technical Standards.

Standard 9.2.4. **The majority of cognitive test items of trend domains shall be kept confidential for re-use in future cycles. A small subset of items which will not be reused shall be released and made available on the OECD website as sample items.**

459. Note: Released items can be items which were administered in the Main Survey but not carried forward to future cycles, items which were administered during the Field Trial but omitted from the Main Survey, or items which were omitted from the Field Trial. The selection of items as released items needs to be approved by the subject-matter expert group. Released items need to be made available along with an explanation of why these items were omitted from the Field Trial/Main Survey/future PISA cycles, description of unit and of the items, mention of the cognitive process that was targeted (as described in the corresponding framework), scoring scheme and instructions, as well as difficulty information if available.

Standard 9.2.5. All questionnaires shall be made available for download on the OECD website. The questionnaires documents shall include metadata information enabling to associate the questions to the framework.

460. Note: Metadata information include the construct targeted by each scale/question, whether the question is new to the cycle or trend, the source if the question is based on published material, as well as the question identifier/variable name to enable external researchers to identify the variable in the database.

Standard 9.2.6. All datafiles, data products and assessment material should be released at the same time as the release of the international initial reports.

461. Note: Historically, the initial reports and data have been released in December on the year after the data collection. Some data, such as the cognitive and questionnaire data for the innovative domain and optional assessments (e.g. financial literacy, foreign language assessment), might be kept confidential at that date and released later to allow for more analysis and quality checks.

Standard 9.2.7. The variable names in the public datafiles shall be consistent and enable the identification of trend items and new items.

Standard 9.2.8. Codebooks explaining the content of the datafiles shall be made available along with the datafiles.

462. Note: External users should have access to clear and detailed data codebooks to facilitate the navigation of the database. The data codebooks should present the variable names, variable labels, values and values label, range of values (minimum and maximum), format, missing observation scheme and a short description for each variable in the corresponding datafile. In addition, they should report descriptive statistics (frequencies and percentages) for all variables that employ a value scheme as well as those that have been derived or added during data cleaning. The codebooks should be available on the OECD website.

Standard 9.2.9. Compendia providing descriptive statistics for each item, by country, should be made available along with the datafiles.

Standard 9.2.10. The coding scheme in the released data shall enable to identify the different reasons for missing data.

463. Note: The missing data coding conventions should be detailed in the codebooks. The different reasons for missing data can be one of the following:

- Missing/blank: In the cognitive data, it is used to indicate the respondent was not presented the question according to the survey design or ended the assessment early and did not see the question. In the questionnaire data, it is only used to indicate that the respondent ended the assessment early or despite the opportunity, did not take the questionnaire.
- No response/omit: The respondent had an opportunity to answer the question but did not respond. For derived variables, this can also indicate that data were incomplete for a component variable.

- Invalid: Used to indicate a questionnaire item was suppressed by country request or that an answer was not conforming to the expected valid response options. For a paper-based questionnaire, it is used when the respondent indicated more than one choice for an exclusive-choice question. For a computer-based questionnaire, it is used when the response was not in an acceptable range of responses, e.g. the response to a question asking for a percentage was greater than 100.
- Not applicable: A response was provided even though the response to an earlier question directed the respondent to skip that question, or the response could not be determined due to a printing problem or torn booklet. In the questionnaire data, it is also used to indicate a response missing by design (i.e. the respondent was never given the opportunity to answer this question).
- Valid skip: The question was not answered because a response to an earlier question directed the respondent to skip the question.

Standard 9.2.11. Code scripts facilitating the use of the datasets and enabling external researchers to reproduce results shall be made available.

464. Note: Statistical routines and syntaxes used for scoring, creating variables and merging and appending data from the different files should be made available on the OECD website. The scripts should be shared in widely used formats for statistical analysis (e.g. in R, SAS, SPSS or STATA language/format).

Standard 9.2.12. The release of process data from computer logfiles shall be accompanied with a clear documentation on which logfile data were collected and how process indicators were constructed.

465. Note: The procedure used for parsing and scoring process data variables should be documented and made available on the OECD website. The documentation should detail the data verification, merging and cleaning steps undertaken to consolidate the data extracts and create indicator scores that align to the evidence model defined in the framework. If possible, the documentation should be accompanied by programs that allow the users to replicate the indicators from the raw process data. The statistical properties of process data indicators and their relationships with other variables providing validity evidence should be also documented.

Standard 9.2.13. Released datasets shall be made available in formats directly usable or easily convertible for use in common statistical software.

466. Note: Common statistical software currently include HLM, Mplus, SAS, SciPy, SPSS and STATA.

Standard 9.2.14. A digital tool enabling users to visualise and interact with the data should be made available on the OECD website.

467. Note: The tool should be intuitive, easy to use, attractive and accessible. It should enable users who have no experience with the use of statistical software to explore the data.

9.2.3. Quality assurance

- The responsible contractor or research team produces and makes available for download on the OECD website:
  - A set of public use datafiles complying with the above standards, including selected log-files;
  - A set of released cognitive test items, along with a document presenting the items;
  - The set of background questionnaires;
  - A codebook for each datafile;
  - An interactive digital tool for data visualisation.
- The responsible contractor or research team documents in the technical report which data have been suppressed from the public use datafiles and for what reasons, including details of the data integrity and quality checks carried.

9.3. Standards on technical reports

9.3.1. Rationale

468. Technical reports accompany the PISA primary international reports of finding. They are one of the main guarantors of PISA's scientific integrity and transparency. Technical reports provide interested readers and researchers with clear and exhaustive information on the actual design, instrument development, implementation, and data analysis of each PISA cycle. In contrast with the primary international reports, which should be accessible to a wide audience, avoiding technical language whenever possible and keeping technical details to the strict minimum necessary to understand the results, the technical reports should serve as the place to document all technical procedures and aspects of the study.

469. Their objective is twofold: first, they act as a tool for quality control, enabling all stakeholders and external actors to evaluate the overall quality of the data with respect to the four principles of validity, reliability, comparability and fairness, and to assess adherence to the PISA Quality Standards. Their second objective is to enable data analysts and researchers to understand the assessment and database in order to facilitate the use of PISA for further research and the replication of results. Technical reports should ensure to the extent possible that PISA data are used and interpreted appropriately.

470. The main contractor for the design, development and implementation of PISA for a given cycle has the responsibility of drafting the technical report, drawing on contributions of all contractors for the respective parts of the survey. The OECD Secretariat is responsible for verifying its content and finalising its production.

471. The standards in this section define the quality criteria that should guide the development of the technical reports as well as their content.

9.3.2. Standards on the technical reports' development process

Standard 9.3.1. The technical reports should be published at the same time as the public data files are released.

472. Note: This alignment is crucial since the primary reports should refer to the technical reports for technical details that are beyond their scope; it is also important to ensure that external users have all the elements to understand the assessment and use the data appropriately.

Standard 9.3.2. Drafts of the technical reports shall undergo review by the OECD Secretariat and invite the input of National Centres.

473. Note: The timeline for drafting the technical reports should allow sufficient time for review and revision and take into account the expertise of those involved in the technical aspects of implementing PISA at the international and national level.

Standard 9.3.3. The technical reports shall be publicly available on the OECD PISA website.

474. Note: The reports and accompanying table and graphs should be accessible and readable online, and downloadable in widely available file formats (e.g. PDF for texts, Excel for data tables). The published reports should link to and be complemented with additional web-based material such as data analysis scripts or supplementary tables as necessary.

9.3.3. Standards on the technical reports' content

Standard 9.3.4. The technical reports shall provide an overview of PISA, including its general features, features specific to the given cycle, and implementation process.

475. Note: In particular, this should include a summary of the technical innovations implemented in the given cycle compared to the previous cycles, as well as the full list of countries and economies participating in the given cycle, indicating their selected option, questionnaire, and administration mode (such as paper- or computer-based).

Standard 9.3.5. The technical reports shall describe the test design of the given cycle.

476. Note: The reports should present the design goals, the use and specific choices of adaptive testing designs (e.g. Multi-stage Adaptive Tests (MSAT)), and domain coverage. It should cover both the Field Trial and Main Survey design, the computer-based and paper-based versions, as well as the inclusion of specific test forms (e.g. the UH form for students with special education needs). See also Standards 6.6 on test and questionnaire design.

Standard 9.3.6. The technical reports shall present how special education needs were taken into account and supported in the design and administration of the assessment.

477. Note: This should include details of the universal design elements incorporated in the test design and assessment platform, as well as accommodations authorised and students who could benefit from them during survey administration. Data on the use

of accommodations should be made available whenever allowed by data privacy regulations.

Standard 9.3.7. The technical reports shall provide information on the population PISA aims to cover with its sample design, said sample design and procedures to draw a representative student sample, as well as sampling outcomes.

478. Note: The section on sample design should include the stratification variables for each country/economy, population coverage and participation rate standards, a presentation of the target population and sampling frame, how sampled schools were identified and tracked, special school sampling situations (e.g. small schools) and additional sampling options (e.g.oversampling, grade-based sampling), overlap control procedures for other national/international studies, school sampling activities, as well as students and teachers selection procedures. The reports should also provide sampling outcomes such as quality indicators for population coverage, school and student response rates (by country and adjudicated region), teacher response rates, and design effects. See also Standards 6.3 on target population, sampling design and sampling operations.

Standard 9.3.8. The technical reports shall provide information on survey weighting and calculation of sampling variance.

479. Note: This should include the procedures used to derive survey weights (school base weight, trimming factor and non-response adjustment, within-school base weight, trimming and non-response adjustment), and the replication methodology used to estimate sampling variance of parameter estimates. Special sampling cases, e.g. of countries/economies where all students are selected for PISA, should also be presented. See also Standards 8.2 on computing survey weights and preparing datasets for estimating sampling variance.

Standard 9.3.9. The technical reports shall describe the test and questionnaire development.

480. Note: The reports should present the new assessment frameworks and their development process, provide the list of context questionnaires administered in the given cycle, and provide an overview of the item development process (including countries/economies' and experts' involvement and the item review process, and validation studies that have been carried to support the valid interpretation of results, e.g. cognitive laboratories and pilots), coding, item selection and construct coverage. See also Standards 7.1, 7.2 and 7.3 on developing test and questionnaire frameworks and items.

Standard 9.3.10. The technical reports shall present the assessment platform (questionnaire and cognitive test).

481. Note: This should include a presentation of the platform's authoring tools (question templates available in the platform, routing rules), the creation of national adaptations and translations, and data collection procedure including the list of logged events. See also Standards 6.4 on the functionalities of the PISA digital tools.

Standard 9.3.11. The technical reports shall document the translation, adaptation and verification process.

482. Note: The reports should describe the development of source versions (including translatability assessments), PISA translation and adaptation guidelines, training sessions for translators, and translation and quality assurance and control procedures. It should provide a list of all the languages used for each country/economy in the given cycle, and collaborations between countries/economies to produce national versions. See also Standards 7.4 on translations and adaptations.

Standard 9.3.12. The technical reports shall document survey field operations.

483. Note: This should include an overview of the roles and responsibilities of the different stakeholders (e.g. National Project Managers, school co-ordinators or associates, test administrators); the selection process of the school and students sample; the preparation of test booklets, questionnaires and manuals; field operations for both paper- and computer-based assessments; test administration procedures including a detailed overview of the sessions schedule/timing; and material submission process. This documentation should describe how operations aligned with the survey field operations described in the Technical Standards of the given cycle and clear in the cases in which they did not.

Standard 9.3.13. The technical reports shall present quality monitoring procedures for data collection activities.

484. Note: The report should present how countries/economies were able to provide review and feedback following the Field Trial and Main Survey operations through the review questionnaires, and present the role of PISA Quality Monitors and the collection of information during their school visits. This should be in line with the quality management plan described in the Technical Standards of the given cycle, and document cases when said Standards were not met.

Standard 9.3.14. The technical reports shall describe PISA data management procedures.

485. Note: This should include the roles of the different stakeholders at the national and international level, the process and workflow for data management, and quality control procedures. The section should also cover how national response categories were mapped into the international response categories (harmonisation), validation checks, scoring and derivation of new variables from questionnaires, and the preparation of files for public use and analysis (list of data files to be delivered, missing data categorisation, range restriction rules for inconsistent and extreme values and list of variables suppressed). See also the Technical Standards of the given cycle and Quality Standards 9.2 on releasing data and assessment material.

Standard 9.3.15. The technical reports shall provide detailed information on the coding design and procedures, coding reliability studies, and machine-supported coding.

486. Note: This section should provide a description of the items, including the number of cognitive items by domain, administration mode (computer- or paper-based), item format (simple or complex multiple choice, constructed response), coding method (human or computer scored). It should also present the coding preparation (recruitment and training

of national coder teams) and the coding design. Finally, it should present within- and across-country coder reliability by domain, including country level score agreement, item-level scoring reliability, and coding category distributions across coders, as well as the development and outcomes of the machine-supported coding system. See also Standards 7.2 on the development of test items and scoring rubrics.

Standard 9.3.16. The technical reports shall provide information on the scaling procedures and scaling outcomes of cognitive test data.

487. Note: This should include analyses of data yield and quality (such as response time analyses, position effects and engagement). This section of the reports should also provide a description of the psychometric scaling models used, calibration and scaling, as well as the population modelling and multiple imputation procedures employed to obtain plausible values for student performance. This theoretical description should be followed by a presentation of how these models were applied to the given cycle data and their outcomes: this should include the estimation of common international and group-specific item parameters (for different language versions – country-by-language groups model fit), investigations of differential item functioning, dimensionality analyses of new instruments, and the linking to previous cycles. The standards should also describe results of the psychometric scaling and population modelling, such as the invariance of the item parameters across countries/economies, indices of reliability of the PISA scales (by country/economy), measurement error, how the plausible values from the population model were transformed to PISA scales (including transformation coefficients), linking error between the given cycle and previous cycles, test targeting (proficiency levels and classification of items and students for each domain). Finally, it should also present the correlations between domains (by country/economy) and domain sub-scales where applicable, as well as the percentage of respondents at each proficiency level (by country/economy). See also Standards 8.3 on constructing and validating scales from cognitive tests.

Standard 9.3.17. The technical reports shall detail the scaling procedures and construct validation of context questionnaire data.

488. Note: This section should provide a description of all variables derived from questionnaire responses. It should also provide information on the scaling procedures, including the models used, criteria for omitting responses, the estimated model parameters and how to interpret them. It should report on construct validation, including internal consistency, cross-country comparability, and invariance of item parameters. See also Standards 8.4 on constructing and validating scales from questionnaire.

Standard 9.3.18. The technical reports shall provide information on the proficiency scale construction.

489. Note: This should include the development of the scales (classification of items, definition of overall proficiency scale, identification of subscales, development of item maps), the definition of proficiency levels. For the main domain, it should present the outcome proficiency scale and definition of proficiency levels, and include an item map for released items. It should also present the cutpoints for each level and each domain. See also Standards 8.5 on constructing proficiency levels and descriptors.

Standard 9.3.19. The technical reports shall describe the data adjudication process and outcomes.

490. Note: This section should recall and refer to the PISA Technical Standards which describe the standards that guide data adjudication. It should present the information available for adjudication and the steps in the data adjudication process. Finally, it should present data adjudication outcomes, including response rate issues, departures from standard procedures in the national data collection plan or issues arising from implementation.

Standard 9.3.20. The technical reports shall present the international data products delivered.

491. Note: This section should present the list of the public use files produced, as well as the variables used for sampling, weighting and merging files, and missing code conventions. It should also present the products accompanying the public datasets (codebooks, compendia), the data analysis procedures and software tools. See also Standards 9.2 on releasing data and assessment material.

Standard 9.3.21. The technical reports should clearly present the validity argument supporting the purposes and uses of PISA.

492. Note: This section should synthesise the studies and analyses that were conducted to support the valid interpretation and uses of PISA data (directing to the relevant sections of the technical reports for further details) in order to present a validity argument that clearly and succinctly summarises the importance of these studies and analyses for justifying the use of PISA scores for their intended purposes.

9.3.4. Quality assurance

- The responsible contractor produces a schedule indicating the time that will be dedicated to drafting and revising the different sections of the technical report and accompanying material. It should allocate sufficient time for review and revision by the OECD Secretariat and National Centres.
- The responsible contractor produces the technical report in accordance with the quality standards above.

References

Abedi, J. and N. Ewers (2013), *Smarter Balanced Assessment Consortium: Accommodations for English language learners and students with disabilities: A research-based decision algorithm*, Smarter Balanced. [85]

Adams, R. (2005), "Reliability as a measurement design effect", *Studies in Educational Evaluation*, Vol. 31/2-3, pp. 162-172, https://doi.org/10.1016/j.stueduc.2005.05.008. [163]

Adams, R., M. Wilson and W. Wang (1997), "The multidimensional random coefficients multinomial logit model", *Applied Psychological Measurement*, Vol. 21/1, pp. 1-23. [131]

AERA, APA and NCME (2014), *Standards for Educational and Psychological Testing, 2014 Edition*. [1]

Ainley, J. and W. Schulz (2020), "Framework Development in International Large-Scale Assessment Studies", https://doi.org/10.1007/978-3-030-53081-5_3. [109]

Araneda, S. et al. (2022), "Exploring Relationships among Test Takers' Behaviors and Performance Using Response Process Data", *Education Sciences*, Vol. 12/2, p. 104, https://doi.org/10.3390/educsci12020104. [17]

Avvisati, F. (2020), "The measure of socio-economic status in PISA: a review and some suggested improvements", *Large-scale Assessments in Education*, Vol. 8/1, p. 8, https://doi.org/10.1186/s40536-020-00086-x. [78]

Avvisati, F. and P. Givord (2021), "How much do 15-year-olds learn over one year of schooling? An international comparison based on PISA", *OECD Education Working Papers*, No. 257, OECD Publishing, Paris, https://doi.org/10.1787/a28ed097-en. [28]

Beaton, A. and N. Allen (1992), "Interpreting scales through scale anchoring", *Journal of Educational Statistics*, Vol. 17/2, pp. 191-204. [156]

Beaton, A. and R. Zwick (1990), *The effect of changes in the National Assessment: Disentangling the NAEP 1985-86 reading anomaly (Rep. No. ETS-17-TR-21)*, National Center for Education Statistics, Washington, D.C., https://files.eric.ed.gov/fulltext/ED322216.pdf (accessed on 1 August 2018). [58]

Beatty, P. and G. Willis (2007), "Research Synthesis: The Practice of Cognitive Interviewing", *Public Opinion Quarterly*, Vol. 71/2, pp. 287-311, https://doi.org/10.1093/poq/nfm006. [120]

Bechger, T. and G. Maris (2014), "A Statistical Test for Differential Item Pair Functioning", *Psychometrika*, Vol. 80/2, pp. 317-340, https://doi.org/10.1007/s11336-014-9408-y. [144]

Benítez, I., F. Van de Vijver and J. Padilla (2022), "A Mixed Methods Approach to the Analysis of Bias in Cross-cultural Studies", *Sociological Methods & Research*, Vol. 51/1, pp. 237-270, https://doi.org/10.1177/0049124119852390. [161]

Benítez, I., F. Van de Vijver and J. Padilla (2019), "A Mixed Methods Approach to the Analysis of Bias in Cross-cultural Studies", *Sociological Methods & Research*, Vol. 51/1, pp. 237-270, https://doi.org/10.1177/0049124119852390. [73]

Berman, A., E. Haertel and J. Pellegrino (2020), *Comparability Issues in Large-Scale Assessment: Issues and recommendations*, National Academy of Education Press, Washington, D.C. [53]

Berman, A., E. Haertel and J. Pellegrino (eds.) (2020), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, National Academy of Education, https://doi.org/10.31094/2020/1. [63]

Birnbaum, A. (1968), "Some latent traitmodels and their use in inferringanexaminee's ability", in Lord, F. and M. Novick (eds.), *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, M.A. [152]

Blömeke, S. and J. Gustafsson (2017), *Standard setting in education*, Springer. [154]

Brancato, G. et al. (2006), *Handbook of recommended practices for questionnaire development and testing in the European Statistical System*, http://chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.istat.it/it/files/2013/12/Handbook_questionnaire_development_2006.pdf (accessed on 20 November 2023). [121]

Breakspear, S. (2012), "The Policy Impact of PISA: An Exploration of the Normative Effects of International Benchmarking in School System Performance", *OECD Education Working Papers*, No. 71, OECD Publishing, Paris, https://doi.org/10.1787/5k9fdfqffr28-en. [31]

Brown, A. and A. Maydeu-Olivares (2011), "Item Response Modeling of Forced-Choice Questionnaires", *Educational and Psychological Measurement*, Vol. 71/3, pp. 460-502, https://doi.org/10.1177/0013164410375112. [67]

Buchholz, J. (2022), "Mixed-worded scales and acquiescence in educational large-scale assessments", *OECD Education Working Papers*, No. 269, OECD Publishing, Paris, https://doi.org/10.1787/8dd310c0-en. [74]

Buchholz, J. and J. Hartig (2019), "Comparing Attitudes Across Groups: An IRT-Based Item-Fit Statistic for the Analysis of Measurement Invariance", *Applied Psychological Measurement*, Vol. 43/3, pp. 241-250, https://doi.org/10.1177/0146621617748323. [65]

Chung, J. (2016), "The (mis)use of the Finnish teacher education model: 'policy-based evidence-making'?", *Educational Research*, Vol. 58/2, pp. 207-219, https://doi.org/10.1080/00131881.2016.1167485. [32]

Cizek, G. (2012), *Setting performance standards: Foundations, methods, and innovations.*, Routledge. [155]

Cochran, W. (1977), *Sampling Techniques*, John Wiley and Sons, New York, NY. [133]

Costa, D. (2021), "LOGANShiny: An app for illustrating process data analysis from international large-scale assessments", http://ceur-ws.org/Vol-3051/PA_2.pdf (accessed on 20 February 2024). [20]

Cronbach, L. (1971), "Test Validation", in Thorndike, R. (ed.), *Educational Measurement*, American Countil on Education, Washington, D.C. [26]

Cronbach, L. and P. Meehl (1955), "Construct validity in psychological tests", *Psychological Bulletin*, Vol. 52/4, https://doi.org/10.1037/h0040957. [4]

Crotts-Roohr, K. and S. Sireci (2017), "Evaluating Computer-Based Test Accommodations for English Learners", *Educational Assessment*, Vol. 22/1, pp. 35-53, https://doi.org/10.1080/10627197.2016.1271704. [89]

Davey, T., M. Pitoniak and C. Slater (2016), "Designing computerized adaptive tests", in Lane, S., M. Raymond and T. Haladyna (eds.), *Handbook of test development*, Routledge, New York. [103]

Dept, S., A. Ferrari and B. Halleux (2017), "Translation and Cultural Appropriateness of Survey Material in Large-Scale Assessments", in *Implementation of Large-Scale Education Assessments*, Wiley, https://doi.org/10.1002/9781118762462.ch6. [124]

Dept, S., A. Ferrari and L. Wäyrynen (2010), "Developments in Translation Verification Procedures in Three Multilingual Assessments: A Plea for an Integrated Translation and Adaptation Monitoring Tool", in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Wiley, https://doi.org/10.1002/9780470609927.ch9. [125]

Diamantopoulos, A. et al. (2012), "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective", *Journal of the Academy of Marketing Science*, Vol. 40/3, pp. 434-449, https://doi.org/10.1007/s11747-011-0300-3. [122]

Dumais, J. and H. Gough (2012), "School sampling methodology", in Greaney, V. and T. Kellaghan (eds.), *Implementing a national assessment of educational achievement*, The World Bank. [95]

Dumais, J. and H. Gough (2012), "Weighting, estimation and sampling error", in Greaney, V. and T. Kellaghan (eds.), *Implementing a national assessment of educational achievement*, The World Bank. [135]

Educational Testing Service (2016), *ETS International Principles for the Fairness of Assessments: A Manual for Developing Locally Appropriate Fairness Guidelines for Various Countries*. [115]

Ercikan, K., H. Guo and Q. He (2020), "Use of Response Process Data to Inform Group Comparisons and Fairness Research", *Educational Assessment*, Vol. 25/3, https://doi.org/10.1080/10627197.2020.1804353. [66]

Faulkner-Bond, M. and Soland (2020), "Comparability when assessing English learner students", in Berman, A., E. Haertel and J. Pellegrino (eds.), *Comparability issues in large-scale assessment*, National Academy of Education Press, Washington, D.C. [83]

Ferrando, P. and D. Navarro-González (2021), "A Multidimensional Item Response Theory Model for Continuous and Graded Responses With Error in Persons and Items", *Educational and Psychological Measurement*, Vol. 81/6, pp. 1029-1053, https://doi.org/10.1177/0013164421998412. [43]

Glas, C. and K. Jehangir (2014), "Modelling country specific differential item functioning.", in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of international large-scale assessment*, CRC Press. [24]

Goldhammer, F. et al. (2021), "From byproduct to design factor: on validating the interpretation of process indicators based on log data", *Large-scale Assessments in Education*, Vol. 9/1, p. 20, https://doi.org/10.1186/s40536-021-00113-5. [113]

Goodnight, J. (1979), "A tutorial on the SWEEP operator", *The American Statistician*, Vol. 33/3, pp. 149-158. [108]

Grisay, A. (2003), "Translation procedures in OECD/PISA 2000 international assessment", *Language Testing*, Vol. 20/2, pp. 225-240. [128]

Grisay, A. (1999), *Report on the development of the French source version of the PISA test material (OECD/PISA Rep.)*, Australian Council for Educational Research, Melbourne. [127]

Groves, R. and L. Lyberg (2010), "Total Survey Error: Past, Present, and Future", *Public Opinion Quarterly*, Vol. 74/5, pp. 849-879. [41]

Haberman, S. (2008), "When Can Subscores Have Value?", *Journal of Educational and Behavioral Statistics*, Vol. 33/2, pp. 204-229, https://doi.org/10.3102/1076998607302636. [145]

Haberman, S. and S. Sinharay (2010), "Reporting of Subscores Using Multidimensional Item Response Theory", *Psychometrika*, Vol. 75/2, pp. 209-227, https://doi.org/10.1007/s11336-010-9158-4. [146]

Hair, J. (2010), *Multivariate data analysis: a global perspective*, Pearson Education, Upper Saddle River, NJ. [153]

Hambleton, R., P. Merenda and C. Spielberger (2005), *Adapting educational and psychological tests for cross-cultural assessment*, Lawrence Erlbaum, Hillsdale, N.J. [92]

Han, Z., Q. He and M. von Davier (2019), "Predictive Feature Generation and Selection Using Process Data From PISA Interactive Problem-Solving Items: An Application of Random Forests", *Frontiers in Psychology*, Vol. 10, https://doi.org/10.3389/fpsyg.2019.02461. [148]

Helms, J. (2006), "Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective", *American Psychologist*, Vol. 61/8, pp. 845-859. [80]

He, Q., F. Borgonovi and J. Suárez-Álvarez (2023), "Clustering sequential navigation patterns in multiple-source reading tasks with dynamic time warping method", *Journal of Computer Assisted Learning*, Vol. 39/3, pp. 719-736, https://doi.org/10.1111/jcal.12748. [19]

Iancheva, D. et al. (2018), "Translational validity of PASAT and the effect of fatigue and mood in patients with relapsing remitting MS: A functional MRI study", *Journal of Evaluation in Clinical Practice*, Vol. 24/4, pp. 832-838, https://doi.org/10.1111/jep.12913. [13]

Immekus, J., T. Jeong and J. Yoo (2022), "Machine learning procedures for predictor variable selection for schoolwork-related anxiety: evidence from PISA 2015 mathematics, reading, and science assessments", *Large-scale Assessments in Education*, Vol. 10/1, p. 30, https://doi.org/10.1186/s40536-022-00150-8. [147]

International Test Commission and Association of Test Publishers (2022), *Guidelines for technology-based assessment*, https://www.intestcom.org/page/16 (accessed on 20 February 2024). [81]

ITC (2019), "ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations", *International Journal of Testing*, Vol. 19/4, pp. 301-336, https://doi.org/10.1080/15305058.2019.1631024. [49]

ITC (2017), *The ITC Guidelines for Translating and Adapting Tests (2nd edition)*, http://www.InTestCom.org (accessed on 20 February 2024). [54]

ITC (2013), *ITC guidelines on test use (version 1.2)*, https://www.intestcom.org/files/guideline_test_use.pdf (accessed on 20 February 2024). [60]

Joo, S. et al. (2021), "Evaluating Item Fit Statistic Thresholds in PISA: Analysis of Cross-Country Comparability of Cognitive Items", *Educational Measurement: Issues and Practice*, Vol. 40/2, pp. 37-48, https://doi.org/10.1111/emip.12404. [25]

Jude, N. et al. (2021), "Lost in Translation?", in *International Perspectives on School Settings, Education Policy and Digital Strategies*, Verlag Barbara Budrich, https://doi.org/10.2307/j.ctv1gbrzf4.13. [33]

Judkins, D. (1990), "Fay's Method for Variance Estimation", *Journal of Statistics*, Vol. 6, pp. 223-229. [140]

Kane, M. (2013), "Validating the Interpretations and Uses of Test Scores", *Journal of Educational Measurement*, Vol. 50/1, pp. 1-73, https://doi.org/10.1111/jedm.12000. [2]

Kane, M. (2006), "Validation", in Brennan, R. (ed.), *Educational measurement*, American Council on Education/Praeger, Washington, DC. [7]

Kankaraš, M. and J. Suarez-Alvarez (2019), "Assessment framework of the OECD Study on Social and Emotional Skills", *OECD Education Working Papers*, No. 207, OECD Publishing, Paris, https://doi.org/10.1787/5007adef-en. [40]

Kish, L. (1992), "Weighting for unequal Pi", *Journal of Official Statistics*, Vol. 8/2, pp. 183-200. [139]

Kish, L. (1965), *Survey sampling*, John Wiley and Sons. [97]

Kish, L. and M. Frankel (1970), "Balanced Repeated Replications for Standard Errors", *Journal of the American Statistical Association*, Vol. 65/331, pp. 1071-1094, https://doi.org/10.1080/01621459.1970.10481145. [138]

Klieme, E. (2020), "Policies and Practices of Assessment: A Showcase for the Use (and Misuse) of International Large Scale Assessments in Educational Effectiveness Research", in *International Perspectives in Educational Effectiveness Research*, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-030-44810-3_7. [34]

Kolen, M. and R. Brennan (2004), *Test equating, scaling, and linking*, Springer. [142]

Krosnick, J. and S. Presser (2009), "Question and Questionnaire Design", in Wright, J. and P. Marsden (eds.), *Handbook of Survey Research*, Elsevier, San Diego, CA. [123]

Kuger, S. and E. Klieme (2016), "Dimensions of Context Assessment", https://doi.org/10.1007/978-3-319-45357-6_1. [112]

Lane, S. (2014), "Validity evidence based on testing consequences", *Psicothema*. [30]

Lee, J. (2020), "Non-cognitive characteristics and academic achievement in Southeast Asian countries based on PISA 2009, 2012, and 2015", *OECD Education Working Papers*, No. 233, OECD Publishing, Paris, https://doi.org/10.1787/c3626e2f-en. [75]

Lee, J. and F. Borgonovi (2022), "Relationships between Family Socioeconomic Status and Mathematics Achievement in OECD and Non-OECD Countries", *Comparative Education Review*, Vol. 66/2, pp. 199-227, https://doi.org/10.1086/718930. [79]

Lohr, S. (1999), *Sampling: Design and analysis*, Duxbury Press. [98]

Lord, F. (1980), *Applications of item response theory to practical testing problems*, Routledge. [141]

Lu, Y. and S. Sireci (2007), "Validity issues in test speededness", *Educational Measurement: Issues and Practice*, Vol. 26/4, pp. 29-37. [105]

Lyons, S., M. Johnson and B. Hinds (2021), "A Call to Action: Confronting inequity in assessment", https://www.lyonsassessmentconsulting.com/assets/files/Lyons-JohnsonHinds_CalltoAction.pdf. (accessed on 20 February 2024). [90]

Maddox, B. (2023), "The uses of process data in large-scale educational assessments", *OECD Education Working Papers*, No. 286, OECD Publishing, Paris, https://doi.org/10.1787/5d9009ff-en. [114]

Mamedova, S. et al. (2021), *2012-2016 Program for International Student Assessment young adult follow-up study: How reading and mathematics performance at age 15 relate to literacy and numeracy skills and education, workforce, and life outcomes at age 19 (NCES 2021-029).*,). U.S. Department of Education, National Center for Education Statistics, https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021029 (accessed on 10 October 2021). [29]

Martin, M., K. Rust and R. Adams (1999), *Technical standards for IEA studies*, International Association for the Evaluation of Educational Achievement. [51]

Martone, A. and S. Sireci (2009), "Evaluating alignment between curriculum, assessments, and instruction", *Review of Educational Research*, Vol. 4, pp. 1332-1361. [12]

Matsumoto, D. and F. van de Vijver (2011), *Cross-cultural research methods in psychology*, Oxford University Press, Oxford, UK. [91]

McDaniel, M. et al. (2007), "Situational judgment tests, response instructions, and validity: a meta-analysis", *Personnel Psychology*, Vol. 60/1, pp. 63-91, https://doi.org/10.1111/j.1744-6570.2007.00065.x. [68]

Meinck, S. (2020), "Sampling, weighting, and variance estimation", in Wagemaker, H. (ed.), *Reliability and validity of international large-scale assessment*, Springer International Publishing. [96]

Meinck, S. (2015), "Computing Sampling Weights in Large-scale Assessments in Education. Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach.", http://surveyinsights.org/?p=5353 (accessed on 20 February 2024). [136]

Meinck, S. and C. Vandenplas (2021), "Sampling Design in ILSA", https://doi.org/10.1007/978-3-030-38298-8_25-1. [99]

Messick, S. (1989), "Validity", in Linn, R. (ed.), *Educational Measurement*, Macmillan Publishing Co. [6]

Messick, S., A. Beaton and F. Lord (1983), *National assessment of educational progress reconsidered: A new design for a new era*. [100]

Mislevy, R., R. Almond and J. Lukas (2003), "A BRIEF INTRODUCTION TO EVIDENCE-CENTERED DESIGN", *ETS Research Report Series*, Vol. 2003/1, https://doi.org/10.1002/j.2333-8504.2003.tb01908.x. [111]

Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16, pp. 159-177. [151]

Muthén, B. and T. Asparouhov (2014), "IRT studies of many groups: the alignment method", *Frontiers in Psychology*, Vol. 5, https://doi.org/10.3389/fpsyg.2014.00978. [150]

NCES (2002), "The Measurement of Instructional Background Indicators: Cognitive Laboratory Investigations of the Responses of Fourth and Eighth Grade Students and Teachers to Questionnaire Items. Working Paper No. 2002 -0 6", https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200206 (accessed on 20 February 2024). [118]

OECD (2022), *PISA 2022 Mathematics Framework*, https://pisa2022-maths.oecd.org/ (accessed on 20 February 2024). [110]

OECD (2021), *21st-Century Readers: Developing Literacy Skills in a Digital World*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a83d84cb-en. [36]

OECD (2021), *Beyond Academic Learning: First Results from the Survey of Social and Emotional Skills*, OECD Publishing, Paris, https://doi.org/10.1787/92a11084-en. [45]

OECD (2020), *PISA 2018 Technical Report*, OECD Publishing, Paris. [21]

OECD (2020), *PISA 2022 Technical Standards*, OECD Publishing, Paris. [50]

OECD (2019), *PISA 2018 Assessment and Analytical Framework*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b25efab8-en. [11]

OECD (2019), *PISA 2018 Results (Volume I): What Students Know and Can Do*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/5f07c754-en. [3]

OECD (2019), *PISA 2018 Results (Volume II): Where All Students Can Succeed*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en. [72]

OECD (2019), *PISA 2021 Translation and Adaptation Guidelines*, OECD Publishing, Paris. [59]

OECD (2018), *PISA for Development Assessment and Analytical Framework: Reading, Mathematics and Science*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264305274-en. [70]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing, Paris. [46]

OECD (2017), "PISA for Development Assessment and Analytical Framework (READING, MATHEMATICS AND SCIENCE)", *OECD Publishing*, Vol. 1/1. [71]

OECD (2016), *Equations and Inequalities: Making Mathematics Accessible to All*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/9789264258495-en. [27]

O'Leary, T., J. Hattie and P. Griffin (2017), "Actual Interpretations and Use of Scores as Aspects of Validity", *Educational Measurement: Issues and Practice*, Vol. 36/2, pp. 16-23, https://doi.org/10.1111/emip.12141. [159]

Padilla, J. and I. Benitez (2017), "A rationale for and demonstration of the use of DIF and mixed methods", in Zumbo, B. and A. Hubley (eds.), *Understanding and investigating response processes in validation research*, Springer Cham. [160]

Padilla, J. and I. Benítez (2014), "Validity evidence based on response processes", *Psicothema*, pp. 136-144. [15]

Pedrosa, I., J. Suárez-Álvarez and E. García-Cueto (2014), "Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación [Content Validity Evidences: Theoretical Advances and Estimation Methods]", *Acción Psicológica*, Vol. 10/2, p. 3, https://doi.org/10.5944/ap.10.2.11820. [8]

Pepper, D. et al. (2018), "Think aloud: using cognitive interviewing to validate the PISA assessment of student self-efficacy in mathematics", *International Journal of Research and Method in Education*, Vol. 41/1, https://doi.org/10.1080/1743727X.2016.1238891. [16]

Randall, J. (2021), ""Color-Neutral" Is Not a Thing: Redefining Construct Definition and Representation through a Justice-Oriented Critical Antiracist Lens", *Educational Measurement: Issues and Practice*, Vol. 40/4, pp. 82-90, https://doi.org/10.1111/emip.12429. [55]

Randall, J. et al. (2022), "Disrupting White Supremacy in Assessment: Toward a Justice-Oriented, Antiracist Validity Framework", *Educational Assessment*, Vol. 27/2, pp. 170-178, https://doi.org/10.1080/10627197.2022.2042682. [56]

Robitzsch, A. and O. Lüdtke (2023), "Why Full, Partial, or Approximate Measurement Invariance Are Not a Prerequisite for Meaningful and Valid Group Comparisons", *Structural Equation Modeling: A Multidisciplinary Journal*, Vol. 30/6, pp. 859-870, https://doi.org/10.1080/10705511.2023.2191292. [164]

Robitzsch, A. and O. Lüdtke (2022), "Some thoughts on analytical choices in the scaling model for test scores in international large-scale assessment studies", *Measurement Instruments for the Social Sciences*, Vol. 4/1, p. 9, https://doi.org/10.1186/s42409-022-00039-w. [143]

Rutkowski, L. et al. (2010), "International Large-Scale Assessment Data", *Educational Researcher*, Vol. 39/2, pp. 142-151, https://doi.org/10.3102/0013189X10363170. [37]

Rutkowski, L. et al. (2014), "Assessment design for international large-scale assessments", in Rutkowski, L., M. von Davier and D. Rutkowski (eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, CRC Press, Boca Rton, FL. [101]

Rutkowski, L. and D. Rutkowski (2018), "Improving the Comparability and Local Usefulness of International Assessments: A Look Back and A Way Forward", *Scandinavian Journal of Educational Research*, Vol. 62/3, pp. 354-367, https://doi.org/10.1080/00313831.2016.1261044. [116]

Rutkowski, L. and D. Rutkowski (2016), "A Call for a More Measured Approach to Reporting and Interpreting PISA Results", *Educational Researcher*, Vol. 45/4, pp. 252-257, https://doi.org/10.3102/0013189X16649961. [38]

Rutkowski, L., M. von Davier and D. Rutkowski (2013), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*, CRC Press. [130]

Särndal, C., B. Swensson and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, New York, NJ. [134]

Schleicher, A. (2019), *PISA 2018: Insights and Interpretations*, OECD Publishing, Paris. [39]

Shin, H., P. Jewsbury and P. van Rijn (2022), "Generating group-level scores under response accuracy-time conditional dependence", *Large-scale Assessments in Education*, Vol. 10/1, p. 4, https://doi.org/10.1186/s40536-022-00122-y. [149]

Sijtsma, K. (2009), "On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha", *Psychometrika*, Vol. 74/1, pp. 107-120, https://doi.org/10.1007/s11336-008-9101-0. [42]

Sireci, S. (2020), "De-"Constructing" Test Validation", *Chinese/English Journal of Educational Measurement and Evaluation | 教育测量与评估双语季刊*, Vol. 1/1. [5]

Sireci, S. (2020), "Standardization and UNDERSTANDardization in Educational Assessment", *Educational Measurement: Issues and Practice*, Vol. 39/3, pp. 100-105, https://doi.org/10.1111/emip.12377. [57]

Sireci, S. (2013), "Agreeing on validity arguments", *Journal of Educational Measurement*, Vol. 50, pp. 99-104. [35]

Sireci, S. (2005), "Unlabeling the disabled: A perspective on flagging scores from accommodated test administrations", *Educational Researcher*, Vol. 34/1, pp. 3-12. [87]

Sireci, S. (1998), "Gathering and Analyzing Content Validity Data", *Educational Assessment*, Vol. 5/4, pp. 299-321, https://doi.org/10.1207/s15326977ea0504_2. [10]

Sireci, S. (1997), "Problems and issues in linking tests across languages", *Educational Measurement: Issues and Practice*, Vol. 16/1, pp. 12-19. [82]

Sireci, S. and Faulkner-Bond (2014), "Validity evidence based on test content", *Psicothema*, Vol. 26, pp. 100-107. [9]

Sireci, S., K. Han and C. Wells (2008), "Methods for evaluating the validity of test scores for English language learners", *Educational Assessment*, Vol. 13, pp. 108-131. [93]

Sireci, S. and M. O'Riordan (2020), "Comparability issues in assessing individuals with disabilities", in Berman, A., E. Haertel and J. Pellegrino (eds.), *Comparability Issues in Large-Scale Assessment: Issues and recommendations*, National Academy of Education Press, Washington, D.C. [84]

Sireci, S., S. Scarpati and S. Li (2005), "Test Accommodations for Students With Disabilities: An Analysis of the Interaction Hypothesis", *Review of Educational Research*, Vol. 75/4, pp. 457-490, https://doi.org/10.3102/00346543075004457. [86]

Sireci, S. and J. Suarez-Alvarez (2022), "Deriving Decisions from Disrupted Data", *Educational Measurement: Issues and Practice*, Vol. 41/1, pp. 23-27, https://doi.org/10.1111/emip.12499. [48]

Stoeffler, K. et al. (2020), "Gamified performance assessment of collaborative problem solving skills", *Computers in Human Behavior*, Vol. 104, https://doi.org/10.1016/j.chb.2019.05.033. [165]

Suárez-Álvarez, J. et al. (2018), "Using reversed items in Likert scales: A questionable practice", *Psicothema*, Vol. 30/2, pp. 149-158. [76]

Survey Research Center (2016), *Guidelines for Best Practice in Cross-Cultural Surveys*, Survey Research Center, Institute for Social Research, University of Michigan, http://ccsg.isr.umich.edu/ (accessed on 20 February 2024). [126]

Swineford, F. (1974), *The test analysis manual (ETS SR-74-06)*, Educational Testing Service, Princeton, NJ. [106]

Teig, N., R. Scherer and M. Kjærnsli (2020), "Identifying patterns of students' performance on simulated inquiry tasks using <scp>PISA</scp> 2015 log-file data", *Journal of Research in Science Teaching*, Vol. 57/9, pp. 1400-1429, https://doi.org/10.1002/tea.21657. [18]

Teltemann, J. and N. Jude (2019), "Assessments and accountability in secondary education: International trends.", *Research in Comparative and International Education*, Vol. 14/2, pp. 249-271. [117]

Testo, A. et al. (2020), "Neural correlates of the ADHD self-report scale", *Journal of Affective Disorders*, Vol. 263, pp. 141-146, https://doi.org/10.1016/j.jad.2019.10.009. [14]

Thurlow, M. et al. (2016), *Principles and characteristics of inclusive assessment systems in a changing assessment landscape (NCEO Report No. 400)*, University of Minnesota, National Center on Educational Outcomes. [88]

van de Vijver, F., N. Jude and S. Kuger (2019), "Challenges in International Large-Scale Educational Surveys", in Denman, B., L. Suter and E. Smith (eds.), *Sage Handbook of International Comparative Research*, SAGE. [22]

van der Linden, W. (2018), *Handbook of item response theory: Three volume set*, CRC Press. [129]

van der Linden, W. (2005), *Linear models for optimal test design*, Springer, New York. [102]

Vigil-Colet, A., D. Navarro-González and F. Morales-Vives (2020), "To reverse or to not reverse Likert-type items: That is the question", *Psicothema*, Vol. 32/1, pp. 108-114. [77]

von Davier, M., E. Gonzalez and R. Mislevy (2009), "What are plausible values and why are they useful?", *ERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, Vol. 2, pp. 9-36. [62]

von Davier, M. et al. (2019), "Developments in Psychometric Population Models for Technology-Based Large-Scale Assessments: An Overview of Challenges and Opportunities", *Journal of Educational and Behavioral Statistics*, Vol. 44/6, pp. 671-705, https://doi.org/10.3102/1076998619881789. [132]

von Davier, M. et al. (2019), "Evaluating item response theory linking and model fit for data from PISA 2000–2012", *Assessment in Education: Principles, Policy & Practice*, Vol. 26/4, pp. 466-488, https://doi.org/10.1080/0969594X.2019.1586642. [64]

Walton, K. et al. (2022), "A Big Five-Based Multimethod Social and Emotional Skills Assessment: The Mosaic™ by ACT® Social Emotional Learning Assessment", *Journal of Intelligence*, Vol. 10/4, p. 72, https://doi.org/10.3390/jintelligence10040072. [69]

Wells, C. (2021), *Assessing measurement invariance for applied research*, Cambridge University Press, Cambridge. [23]

Willis, G. (2005), *Cognitive Interviewing: A Tool for Improving Questionnaire Design*, Sage Publications, Thousand Oaks, CA. [119]

Winter, P. (2010), *Evaluating the comparability of scores from achievement test variations*, Council of Chief State School Officers, Washington, D.C. [94]

Wolter, K. (2007), *Introduction to Variance Estimation*, Springer, New York, NJ. [137]

Yamamoto, K. et al. (2017), "Developing a Machine-Supported Coding System for Constructed-Response Items in PISA", *ETS Research Report Series*, Vol. 2017/1, pp. 1-15, https://doi.org/10.1002/ets2.12169. [44]

Yamamoto, K. et al. (2017), "Developing a Machine-Supported Coding System for Constructed-Response Items in PISA", *ETS Research Report Series*, Vol. 2017/1, pp. 1-15, https://doi.org/10.1002/ets2.12169. [52]

Yamamoto, K. and M. Lennon (2018), "Understanding and detecting data fabrication in large-scale assessments", *Quality Assurance in Education*, Vol. 26/2, https://doi.org/10.1108/QAE-07-2017-0038. [162]

Yamamoto, K., H. Shin and L. Khorramdel (2019), "Introduction of multistage adaptive testing design in PISA 2018", *OECD Education Working Papers*, No. 209, OECD Publishing, Paris, https://doi.org/10.1787/b9435d4b-en. [47]

Zapata-Rivera, D. (2011), "Designing and evaluating score reports for particular audiences", in Zapata-Rivera, D. and R. Zwick (eds.), *Test score reporting: Perspectives from the ETS score reporting conference (Research Report 11–45, pp. 32–61)*, Educational Testing Service, Princeton, NJ. [157]

Zapata-Rivera, J. and I. Katz (2014), "Keeping your audience in mind: applying audience analysis to the design of interactive score reports", *Assessment in Education: Principles, Policy & Practice*, Vol. 21/4, pp. 442-463, https://doi.org/10.1080/0969594X.2014.936357. [158]

Zeng, G. and E. Zeng (2021), "On the relationship between multicollinearity and separation in logistic regression", *Communications in Statistics-Simulation and Computation*, Vol. 50/7, pp. 1989-1997. [107]

Zenisky, A., R. Hambleton and R. Luecht (2009), "Multistage Testing: Issues, Designs, and Research", in *Elements of Adaptive Testing*, Springer New York, New York, NY, https://doi.org/10.1007/978-0-387-85461-8_18. [61]

Zwick, R. (2012), *A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement (ETS RR-12-08)*, Educational Testing Service, Princeton, NJ. [104]