OECD Education Working Papers No. 310

Towards more diverse and flexible international large-scale assessments

Tomoya Okubo

https://dx.doi.org/10.1787/0417b5ec-en

OECD

Organisation for Economic Co-operation and Development

EDU/WKP(2024)5

Unclassified **English text only**

13 March 2024

DIRECTORATE FOR EDUCATION AND SKILLS

Towards More Diverse and Flexible International Large-Scale Assessments

OECD Education Working Paper No. 310

Tomoya OKUBO (OECD)

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

Tomoya OKUBO, Tomoya.OKUBO@oecd.org

JT03539387

OECD EDUCATION WORKING PAPERS SERIES

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

This working paper has been authorised by Andreas Schleicher, Director of the Directorate for Education and Skills, OECD.

---

www.oecd.org/edu/workingpapers

Acknowledgement

The author expresses sincere gratitude to Elena Govorova (2E Estudios & Evaluaciones) and Takahiro Terao (The National Center for University Entrance Examinations) for their meticulous reviews and insightful feedback on this paper. Additionally, profound thanks are extended to the following OECD analysts for their significant contributions: Michael Ward, Francesco Avvisati, Claudia Tamassia, Marco Paccagnella, Nate Reinertsen, and Tanja Bastianic.

Abstract

Drawing mainly on OECD data and experience, this paper explores two major enhancements to the utility of international large-scale assessments (ILSAs). The first is the diversification of assessments, focusing on specific groups or individuals to offer more targeted diagnoses. This diversification allows the robust, internationally standardised scales to be applied at both group and individual levels, broadening their impact. The second enhancement is the flexibilisation of assessments. This involves the ongoing refinement of the item bank, increasing the adaptability and relevance of the assessments. Additionally, the paper presents prototypes of new assessment tools derived from existing assessments, employing the methodologies discussed herein. These innovations represent significant strides in the evolution and application of international large-scale assessments.

Table of contents

Tables

Figures

1. Design of international large-scale assessment

1.1. Purposes and outcomes of international large-scale assessments

International large-scale assessments (ILSAs), including the OECD's Programme for International Student Assessment (PISA), the Programme for the International Assessment of Adult Competencies (PIAAC), the IEA´s the Trends in International Mathematics and Science Study (TIMSS), and the Progress in International Reading Literacy Study (PIRLS), primarily focus on assessing proficiency distributions in cognitive domains such as mathematics, reading, and science of each participating country/economy. These assessments also collect background questionnaires from participants, offering a comprehensive view of the learning environment, economic, cultural, and social status (ESCS), and other contextual factors influencing respondents' performance. This approach enables stakeholders to derive meaningful insights, formulate educational policies, and devise targeted interventions to improve learning outcomes across countries/economies.

In ILSAs, a representative sample is selected within each country/economy, and each respondent is assigned a sample weight to account for systematic differences in probability sampling. Utilising statistical analysis with these sample weights helps maintain the representativeness of the estimates at the population level. One of the primary objectives of ILSAs is to obtain representative statistics for the target measures that can be compared across participating countries/economies. On the other hand, it is important to note that diagnosing the proficiency of each sampled individual[1] is not the primary goal of ILSAs.

Therefore, the primary outcome of ILSAs lies in the dataset that includes cognitive test scores, respondents' background information, and the associated sample weights. These published datasets have played a crucial role in various research endeavours and notably impacted the development of education policies (Hopfenbeck et al., 2018[1]). While the media often focuses only on country/economy rankings based on cognitive test scores, it is important to emphasise that the central outcome of ILSAs is the comprehensive dataset containing a wide range of variables that are useful for secondary analysis. Consequently, ILSAs are purposefully designed to produce a dataset that gives unbiased, consistent, and internationally comparable statistics.

Significantly, ILSAs are strategically designed to facilitate comparisons across cycles. The scales utilised in ILSAs remain consistent across survey cycles, ensuring that scores are comparable from one cycle to the next, provided that item banks are appropriately maintained. Additionally, meticulous efforts are made to develop and maintain cognitive items and questionnaires to maximise comparability with previous cycles. This approach empowers participating countries/economies to concentrate on assessing changes in scores over time rather than overly emphasising comparisons with others. This particular feature sets ILSAs apart from educational assessments that may lack this longitudinal aspect.

To maintain horizontal comparability (across countries/economies) and vertical comparability (across test cycles), ILSAs implement stringent test administration regulations. These encompass the utilisation of advanced psychometric modelling and a multifaceted validation process for test instruments to ensure the assessments maintain a dependable and valid scale.

[1] In this paper, "individual" and "group" describe a single test-taker and a collective of test-takers, respectively. "Population" refers to the sampled individuals who embody the targeted demographic. "Sub-population" signifies a segment of the population that maintains the representativeness of the conditioned target, whereas a group may not always possess this representativeness.

Furthermore, specialised committees comprised of experts convene to deliberate on technical matters related to all aspects, including scaling, and these play a pivotal role in shaping the analytical framework for cognitive domains and background questionnaires.

Technical reports that contain technical explanations of the survey and a report on assessment frameworks that presents the conceptual foundation of the assessment are published for transparency (OECD, 2014[2]; OECD, 2017[3]; Martin, von Davier and Mullis, 2020[4]; von Davier et al., 2023[5]). In addition, other materials, such as a data analysis manual (OECD, 2009[6]; OECD, 2009[7]), are also available to support readers and users. Another important role of ILSAs is to publish conceptual frameworks broadly discussed by experts and technical standards that are examined thoroughly. Those concepts and standards have impacted national-level curriculum and educational assessments. Thus, another significant responsibility of ILSAs is disseminating expert-discussed concepts related to target measures and meticulously scrutinised technical standards. These concepts and standards hold a profound influence not only on national-level curriculum development and educational assessments but also on researchers in education-related fields. Further insights are elaborated in the works of Clarke (2012[8]) and Cresswell, Schwantner and Waters (2015[9]).

1.2. Growing demands for multi-purpose utilisation of assessment

As elaborated in the previous subsection, ILSAs adhere to rigorous technical standards and validation procedures (OECD, 2017[3]), encompassing cognitive items and background questionnaires. Likewise, administration protocols and scaling methodologies are meticulously delineated to guarantee the steadfast preservation of international comparability in the results. Consequently, ILSAs are widely recognised as high-quality instruments primarily focusing on producing unbiased statistics at the population level (i.e., country/economy).

Conversely, it is important to note that ILSAs are not designed to produce individual diagnoses. Furthermore, the produced dataset of ILSAs does not ensure unbiased statistics with regard to proficiency at a group level (e.g., schools) because of the mathematical property of the statistical modelling ILSAs employ for producing the output (OECD, 2017[3]; Okubo, 2022[10]). Additionally, individual scores (normally reported as a set of plausible values) are biased estimates of the underlying proficiency of every single respondent, as they are influenced not only by answers to the cognitive test items but also by background characteristics. Therefore, individual-level proficiency statistics output as a form of plausible values should not be employed for diagnostic purposes. These constraints result from the statistical modelling utilised by ILSAs rather than stemming from the cognitive items and administration protocols they implement.

In the context of scoring methodologies for ILSAs, the design of test forms[2] is strategically tailored to address specific challenges. Given the constrained time available to respondents during these assessments, it is impractical for each respondent to respond to a large enough set of questions that comprehensively cover all aspects, subdomains and response formats of the target domain or multiple domains. ILSAs prioritise optimising content coverage and score reliability at the population level rather than focusing on individual or group levels. To achieve this objective, respondents are assigned different sets of items, ensuring a diverse range of content coverage at the population level. This approach facilitates the collection of responses to a substantial number of items collectively, contributing to robust assessments at

[2] In this paper, the term “test form” refers to a collection of units or items, which encompasses single or multiple domains.

the population level. Concurrently, this strategy underscores that ILSAs are not structured to prioritise the reliability and validity of group or individual-level outputs.

Nonetheless, there is a growing demand for utilising ILSA instruments to obtain insights into proficiency levels, both at the group and individual levels, on a demand basis. The anticipation is that users of these assessments will be able to estimate the proficiency levels of specific individuals or groups on the same international scale used for the population level.

Furthermore, there is a growing interest among certain policymakers in identifying and offering support to groups that demonstrate significantly lower proficiencies than the average. Given that ILSA test forms are assembled to encompass a broad spectrum of proficiency levels, their current format is not ideal for assessing extreme-level groups in their current state. Such assessments not aligning with respondents' proficiency can lead to substantial measurement errors, resulting in low-quality datasets.

One potential solution may entail implementing a multi-stage adaptive testing (MSAT) approach (Chang and Ying, 1999[11]; Yamamoto, Shin and Khorramdel, 2019[12]). MSAT presents an effective approach to assessing proficiency by supplying respondents with items tailored to their tentatively assessed proficiency levels. In the PISA framework, a test is segmented into three stages, with the items for the second and third stages being chosen in accordance with the provisionally estimated scores from the preceding stages.

However, it is essential to acknowledge that the current MSAT designs utilised in ILSAs (e.g., three-stage testing) may not fully cater to the specific requirements necessary to effectively accommodate and provide accurate assessments for groups with significantly different proficiency levels. Furthermore, to optimise the efficacy of MSAT or other adaptive testing techniques, it is crucial to have a substantial item bank with a sufficient number of items encompassing a diverse range of difficulties, as well as balanced coverage of subdomains and response formats in order to fully utilise the adaptive approach.

Addressing the evolving needs of educational assessments requires refining the methods underpinning these current surveys. One primary consideration is adapting the scoring methodologies, ensuring they remain effective for the specific reporting needs, from group-level assessment to individual-focused diagnostics. Moreover, as the scoring methodologies are optimised to the assessment's target levels, the test form designs must be congruent with these updated methods. This harmonisation ensures that the instruments used in ILSAs remain accurate, relevant, and capable of appropriately analysing proficiency levels at both the group and individual levels.

1.3. Item bank development in large-scale assessment

The development of items used in ILSAs involves a series of critical steps. It begins with item drafting, where items are meticulously developed to align with the assessment's framework and content domains. These items are designed to measure the knowledge or skills being assessed effectively. Subsequently, the content layout of the items is considered, ensuring that the organisation and presentation of items are clear and coherent. This is vital to enable respondents to understand and respond to the items as designed by the item writers.

The item review phase is essential, encompassing various aspects such as verifying content correctness, adhering to copyrights and legal guidelines, considering political and cultural perspectives to prevent bias or insensitivity, and conducting an overall assessment of item appropriateness. This phase is conducted multiple times with different reviewers.

The translation process is meticulously executed as ILSAs are international and, therefore, multilingual by definition. Translations must faithfully convey the intended content while

accommodating linguistic nuances, cultural differences, and the psychometric properties of words that may impact item performance. This careful approach ensures that assessment items are effectively adapted to diverse linguistic and cultural settings, preserving the validity and reliability of the assessments.

After these preparatory stages, field trials are undertaken to collect data for psychometric analysis. This examination assesses item performance with regard to psychometric properties across a wide spectrum of populations encompassing various subpopulations (OECD, 2017[3]; von Davier et al., 2023[5]). Items that exhibit inadequate psychometric properties are either omitted from the item bank or subjected to revision to enhance their quality.

Despite the comprehensive validation process, including field trials, it is important to note that only the items that successfully pass validation with the main study data are utilised for proficiency estimations. This final selection ensures that high-quality items contribute to the overall assessment, upholding the validity and reliability of the results obtained in ILSAs. It is worth noting that the entire item development process of most ILSAs is conducted concurrently among participating countries to maintain consistent timelines and ensure seamless coordination of efforts. This synchronised approach is essential for the successful execution of ILSAs, allowing for the timely and standardised development of assessment items across countries/economies.

In the context of an ILSA, it is common for some items to be excluded after field trials. This exclusion can occur due to time constraints that prevent the revision of items based on the results of psychometric analysis conducted with the field trial data. It is important to recognise that these excluded items may still have the potential to be revised and transformed into effective assessment items. Furthermore, continuous cycles of item development, data-collection, item analysis, and item revising present not only an opportunity to improve the quality of the items but also a valuable opportunity for item writers to enhance their capacity and expertise. This, in turn, contributes to the overall efficiency of item development within the entire system, ultimately benefiting the quality and effectiveness of future assessments.

A notable consideration in ILSAs is the substantial cost of field trials. These trials can be expensive and significantly burden the participating countries/economies. Nonetheless, if the developed items cannot be revised based on the results of the psychometric analysis based on the field-trial data, field trials become less valuable as it is one of the main purposes of conducting field trials. It is crucial to carefully assess the role and cost of field trials in the planning and executing ILSAs, ensuring they align with the assessment's objectives and goals. The essence of item trialling lies in its ability to verify whether the developed items possess the requisite psychometric properties to measure aspects of the intended construct accurately. This objective can be achieved using data sampled from a subset of the target population, underscoring that representativeness is not a strict prerequisite for item trialling.

Item validation analysis, an essential process in the development of assessment instruments, can be integrated into the main data collection phase, thereby reducing the dependence on separate field trials. This approach, known as in-test item trialling, involves embedding trial items within the test forms used in actual assessments. This strategy is especially effective when there are sufficient validated items in the test forms to construct the measure reliably. Adopting in-test item trialling is particularly advantageous in assessments that occur sporadically or have shorter periodic cycles, such as annual assessments. It offers a cost-effective alternative to traditional field trials, as it eliminates the need for separate testing phases dedicated solely to item validation.

Field trials, therefore, are not always a mandatory requirement in item bank development, especially in the case of sporadic or short-time periodic assessments. However, note that field trials play an important role for field-operation trial purposes. This shift from traditional field

trials to in-test trials offers considerable strategic benefits. It can potentially reduce costs and administrative burdens for those managing and participating in the assessment. This streamlined approach maintains the effectiveness of the validation process while adapting it to suit the demands and constraints of sporadic assessments, potentially leading to a more efficient and cost-effective methodology for achieving the key objectives of item validation.

1.4. Objective and Overview

The previous subsections briefly introduced the growing demands for multi-purpose utilisation of ILSAs and the item development processes in periodic and sporadic assessments. Subsection 1.2 explained that carefully and well-developed large-scale assessments can be used for group-level proficiency estimation and individual diagnosis purposes by developing scoring methodologies and test form designs adopted for the purposes. Subsection 1.3 focused on the challenges of item development and the cost-effectiveness of field trials in ILSAs. Furthermore, the possibility of in-test trials in assessments was introduced.

The objectives of this paper encompass the following:

1. Define the outputs of cognitive tests at the population, group, and individual levels. Concurrently, formulate estimation methodologies tailored to obtain these outputs. In alignment with this, establish test form design principles to optimise the efficient assignment of cognitive items among respondents for the methodologies.
2. Describe the comprehensive process of item bank management in sporadic assessments, particularly those that do not incorporate field trials. Additionally, it provides a detailed formulation of the item parameter estimation and the item validation procedures.
3. Illustrate the practical application of the methodologies introduced in this study by designing new assessment tools.

The subsequent sections of this paper are organised as follows:

Section 2 introduces methodologies and procedures for enhancing the diversity of ILSAs. Specifically, it encompasses defining assessment outcomes, addressing population, group and individual levels, and formulating estimation methodologies for these outcomes. The discussion also delves into the statistical properties of the estimates. Additionally, this section provides principles for designing test forms that optimise assessment outputs.

Section 3 proposes a streamlined approach for test administration in ILSA, focusing on integrating item development and validation processes directly into in-test item trialling. The section covers the essentials of developing test items, their validation, and item parameter estimation, all within the main study framework. This integrated method advocates for a dynamic approach to item management, signifying a shift towards more efficient and adaptable practices in ILSAs.

Sections 4 and 5 detail the development processes of new assessment tools based on the original ILSAs. Section 4 presents the design of a new online assessment tool to evaluate proficiency levels for both groups and individuals contextualised within the PIAAC framework. Section 5 introduces an assessment tool specifically tailored for a condensed version of PISA for Development (OECD, n.d.[13]).

Section 6 summarises the introduced methods that enable more diverse and flexible ILSAs. Additionally, it delves into the limitations of these methodologies and approaches, drawing insights from the assessment development performed in the preceding sections.

2. Towards more diverse assessments

2.1. Outcomes of cognitive assessments at different layers

2.1.1. Population-level output

ILSAs primarily estimate proficiency distributions at the population level. These population-focused assessments ensure that the outputs, which are the proficiency distributions at the population level, exhibit high levels of validity, reliability, and international comparability. Significantly, ILSAs use representative sampling to secure statistically representative outcomes. Subsequently, the individual-level output of a cognitive assessment is generated within the dataset in the form of plausible values (PVs). These PVs are released with the information collected in the background questionnaire in micro-level datasets for public use.

In a typical ILSA, each respondent engages with two to three domains within a 120-minute timeframe at most, responding to around 60-90 items distributed across these domains. However, estimating a point estimate of a respondent's proficiency is often impractical due to the considerable standard error of the estimated proficiency based on 20-30 items per domain. This standard error tends to be disproportionately larger than the population's proficiency distribution. Moreover, the content coverage of the assessment with only 20-30 items does not reach an ideal level.

The focus of ILSAs on population-level statistics makes the rotational test form an ideal choice. However, under this design, respondents do not take items from every domain or subdomain. Consequently, it is not feasible to estimate respondents' proficiencies across all domains based solely on the cognitive items they respond to. Filling the dataset with such domain scores would lead to a sparse dataset, which is not suitable for secondary analysis.

Even if participants were to engage with all the target domains, adequately covering each domain with a sufficient number of items, using the dataset filled with point estimates for secondary analysis could introduce biases. Mathematically, using the point estimate of each respondent to estimate the proficiency distribution of the population underestimates the distribution's variance, which also means that the correlation coefficients between the proficiencies and other variables are overestimated. As such, point estimates of respondents should not be used for estimating population-level proficiency distributions.

In practice, to mitigate the risk of potential misuse, datasets released by ILSAs do not include the point estimates of respondents' proficiencies. Instead, the dataset provides multiple PVs, random values drawn from the probability distribution corresponding to each respondent's predicted proficiency. This methodology for generating PVs is commonly referred to as population modelling within the context of ILSAs. The mathematical properties of PVs are discussed in the following sections.

It is important to note that PVs are assigned for all domains, even if a respondent did not take a specific domain. Mathematically, the dataset is designed to offer unbiased estimates of the population level parameters (e.g., mean and variance of the proficiency distribution), making it ideal for data analysis conducted by researchers. Given this goal, the collection of auxiliary information (i.e., respondents' background data) on each respondent is not only useful for contextualising results, but an important design feature that enables to perform the population

modelling. For further discussion and details of the PVs, refer to OECD (2009[6]; 2017[3]). The theoretical background of the population modelling is given by Okubo (2022[10]).

2.1.2. Group-level output

As a group of respondents (e.g., schools, cities, etc.) takes enough items to estimate the group-level proficiency distributions, the estimated proficiency distributions can still be considered valid and reliable at the group level. The group-level statistics offer a distinct advantage in understanding target group proficiencies in relation to national and international benchmarks. However, the group-level statistics should not be calculated based on the PVs generated through the population modelling using the entire population data since those PVs do not ensure the unbiasedness of a group-level proficiency distribution. The mathematical properties of the PVs are discussed in the following sections.

Performing population modelling with a limited sample size is impractical unless the sample includes a diverse representation of respondents spanning a wide spectrum of proficiency levels and with significant variability in background information. Consequently, generating PVs based on a small-sized dataset is discouraged for mathematical reasons. Hence, it is recommended to estimate group-level outputs for cognitive assessments directly by employing a statistical model that only leverages data from the respondents within the specific group, not through the generated PVs for the population or the groups (See Section 2.2 for details). To ensure validity and reliability, the modelling approach at the group level, such as using marginal likelihood estimation with respect to the parameters of the proficiency distributions, should be founded on a substantial number of responses to cognitive items that comprehensively address the target concept. Note that the estimates are still mapped on the international scale for the population.

Within group-level cognitive assessment output, one typically considers two fundamental categories of statistics to gain insights into the proficiency levels. The first category encompasses estimates of the mean and variance of the normal distribution regarding proficiencies. The second category of statistics focuses on estimating parameters of the multinomial distribution of proficiencies. In contrast to the normal distribution, the multinomial distribution deals with discrete categories or scores. This method offers a more detailed and granular view of the distribution of proficiencies across various skill levels within the group, enabling an understanding of the precise breakdown of proficiency levels. In summary, these two sets of statistics fulfil distinct yet complementary roles. The estimates of mean and variance provide a continuous overview of proficiency levels, while the parameters of the multinomial distribution offer a discrete breakdown of proficiency categories.

2.1.3. Individual-level output

The primary purpose of individual-focused assessment is to obtain diagnostic information on a respondent's cognitive domain proficiencies. PVs exhibit advantageous mathematical properties when utilised for population-level statistics; however, they may not align well with individual diagnosis purposes. This misalignment stems from the incorporation of a prior distribution estimated from responses to the background questionnaire in the generation of PVs. For individual-level assessments, it is essential to base the output solely on the responses to the cognitive items and their parameters, relying on the likelihood function with respect to the proficiency for estimation.

Nevertheless, practical constraints, such as limited testing time and the restricted number of items each respondent can effectively engage with, make it unfeasible to estimate a precise point estimate of proficiency with a small standard error at the individual level.

Consequently, the size of the standard error must be carefully considered when choosing the format of individual-level outputs. This consideration is crucial to prevent any overestimation of score reliability. One potential output format is a band score, wherein a respondent's proficiency is categorised into one of the predefined bands. Typically, each band corresponds to a proficiency level that aligns with specific can-do statements, offering a more categorical representation of an individual's proficiency. While it doesn't directly reduce the standard error, a band score approach helps alleviate the expectation for outputs to provide highly granular information.

Another viable output option is a pass/fail classification. While mathematically, it can be considered a special case of the band score approach with only two bands, it is distinct from a test-design perspective. An assessment employing a pass/fail classification output is tailored to identify participants who meet a specific proficiency threshold, whereas an assessment using a band score approach encompasses a broader range of proficiency levels without a specific focus on one particular level.

While utilising a precise point estimate for an individual's proficiency, this approach substantially burdens the respondent. Furthermore, obtaining such precise point estimates can be costly from the test administrator's standpoint and may not offer cost-effective benefits. Additionally, these estimates may not provide unbiased estimates of population-level statistics, as mentioned above. Therefore, alternative methods like the ones previously mentioned are often preferred due to their practicality and efficiency.

2.1.4. Summary

The following Table 1 summarises the discussions in Subsection 2.1.

Table 1 Outcomes of cognitive assessments at different layers

	Population-focused assessment	**Group-focused assessment**	**Individual-focused assessment**
Purpose	Estimate comparable proficiency distributions across countries/economies in cognitive domains and publish the dataset for secondary analysis.	Gain insights into target group proficiencies against national and international benchmarks.	Obtain diagnostic information on a respondent's cognitive domain proficiencies.
Sampling	Representative sampling	Census / random sampling	Not applicable
Background questionnaire	Primary purpose	Not envisioned (only demographic variables)	Only demographic variables
Data analysis[3] with the dataset	Major use	Limited use	Not applicable
Output (population level)	Parameters of the proficiency distributions	Not applicable	Not applicable
Output (group level)	See "group-focused assessment."	Parameters of the proficiency distributions	Not applicable
Output (individual level)	PVs / See "individual-focused assessment."	See "individual-focused assessment."	Band score or pass/fail classification

[3] For example, analysing relationships between proficiency and background variables or estimating the score distribution of the variables.

2.2. Approaches to estimating outputs

This subsection outlines the estimation procedures for the outputs at population, group, and individual levels, introduced in Subsection 2.1.

For the population-level output, the process of PVs generation is explained, along with an overview of their mathematical properties. It is important to note that while PVs do not directly represent the parameters, the statistics derived from PVs represent the underlying parameters. This distinction underscores the utility of PVs in statistical analysis, as they serve as a valuable tool for representing population-level characteristics and parameters.

For the group-level output, the estimation procedure of the proficiency distribution of a group/population is defined. Furthermore, the procedure for estimating the proportion of categorical proficiency levels is also introduced. This procedure elucidates how insights into a group's collective proficiency levels are derived, providing valuable information for educational assessments.

Two distinct approaches are introduced for the individual outputs: the band score approach and the pass/fail classification. These approaches are designed to assess and categorise individual proficiency levels, offering a more practical perspective on an individual's performance within the assessment framework.

These defined procedures collectively contribute to a comprehensive understanding of how each output type is derived and assessed within educational assessments, catering to various analytical and diagnostic needs.

2.2.1. Plausible values

The PVs for each respondent are drawn from a M-dimensional posterior distribution, where the likelihood functions for each domain are independent. In contrast, the prior distribution follows a multivariate normal distribution in M dimensions. ILSAs employ item response theory (IRT; Lord and Novick (1968[14]), Lord (1980[15])) for the measurement model, in which a probability of responding to category $k\ (0, \dots, K_j - 1)$ of an item $j\ (1, \dots, J)$ is defined with proficiency (θ) and parameters of item j ($\mathbf{\Lambda}_j$). It is called item category response function (ICRF) and is denoted as $p_{jk}(\theta|\mathbf{\Lambda}_j)$. In most ILSAs, generalised partial credit model (GPCM; Muraki (1992[16])) is employed; however, the graded response model (GRM; Samejima (1969[17])) is also a choice (Thissen and Steinberg, 1986[18]).

Considering the implications of fixing slope parameters when employing IRT models is crucial. Models that assume a constant slope parameter across items, like the partial credit model (PCM; Masters (1982[19])) or the Rasch model (Rasch, 1960[20]), can inadvertently introduce issues in the assessment analysis. These issues primarily include the overestimation of residual variances and the consequential problem of scale shrinkage. Therefore, when selecting an IRT model for assessment purposes, it is advisable to use models that allow for variability in item discrimination, such as the GRM or the GPCM. These models do not constrain the slope parameters, thereby providing a more model-data fit and potentially more accurate measure of the respondents' abilities. This is important, especially when the parameters of new items are estimated in every testing cycle and added to the item bank, as it minimises the risk of scale shrinking. Note that ILSAs utilise a unidimensional IRT model.

Under the local independent assumption, the likelihood function with respect to θ given by a binalised response vector of respondent $i\ (= 1, \dots, N)$, $\boldsymbol{u}_i$, is defined as

Equation 1

$$L(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}) = f(\boldsymbol{u}_i|\theta, \boldsymbol{\Lambda}) = \prod_{j=1}^{J} \prod_{k=0}^{K_j-1} p_{jk}\left(\theta|\boldsymbol{\Lambda}_j\right)^{u_{ijk}}$$

In ILSAs, latent regression modelling (LRM) is employed to estimate the parameters of the prior distribution of individual respondents, which is assumed to follow a normal distribution. In the LRM, $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N]$, are regressed on covariates $\boldsymbol{y}$ generated from the responses to the background questionnaires to estimate regression parameters $\boldsymbol{\Gamma}$.

$$\boldsymbol{\theta} = \boldsymbol{\Gamma}\boldsymbol{y} + \boldsymbol{d}$$

where

Equation 2

$$\begin{aligned} \mathrm{E}[\boldsymbol{d}] &= \boldsymbol{0} \\ \mathrm{Cov}[\boldsymbol{\Gamma}\boldsymbol{y}, \boldsymbol{d}] &= \boldsymbol{0} \end{aligned}$$

are assumed in the LRM.

In the population modelling, the respondent's proficiency $\boldsymbol{\theta}_i$ assumed to follow

Equation 3

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\Gamma}\boldsymbol{y}_i, \boldsymbol{\Sigma})$$

Note the residual covariance matrix of the LRM, $\mathrm{V}[\boldsymbol{d}] = \boldsymbol{\Sigma}$, is common to all the respondents within a country/economy. Therefore, the prior distribution of respondent i forms as $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$, where the prior distribution forms a M-dimensional normal distribution.

Thus, the posterior distribution of respondent i is formulated as

$$p(\boldsymbol{\theta}_i|\boldsymbol{u}_i, \boldsymbol{y}_i, \boldsymbol{\Lambda}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}) \propto f(\boldsymbol{u}_i|\boldsymbol{\theta}_i, \boldsymbol{\Lambda})\, h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$$

The PVs are drawn multiple times from the posterior distribution, giving consistent and unbiased statistics of proficiency distribution. Details of the procedure and the mathematical property are explained in OECD (2017[3]) and Okubo (2022[10]).

The above equation indicates that each respondent obtains multiple M-dimensional PVs regardless of the domains they are assigned in the cognitive assessment because of $h(\boldsymbol{\theta}_i|\boldsymbol{y}_i, \boldsymbol{\Gamma}, \boldsymbol{\Sigma})$. This makes test form design flexible. However, to assess the correlation structure of the prior distribution accurately, a sufficient number of respondents must participate in assessments that encompass various combinations of two domains. It is crucial that the participating respondents represent all possible combinations to estimate correlations between any two domains effectively.

In the LRM, each participant must furnish enough responses to background questionnaires, which serve as covariates. This data is essential for robustly estimating the prior parameters. An assessment lacking the collection of background questionnaire data would be unable to implement the LRM effectively; thus, it cannot generate PVs. Furthermore, to obtain stable estimates of $\boldsymbol{\Gamma}$, the form of $L(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda})$ of all respondents should be stable; thus, respondents to be analysed in the LRM dataset need to respond to enough items. In the case of PISA, students who did not respond to six or more items in a target domain are excluded from the dataset to be analysed with LRM. However, those excluded students are included when generating PVs if they are considered eligible students to put in the dataset.

In order to mitigate the potential impact of unforeseen bias stemming from the prior distributions, it is advisable to scrutinise Equation 2 and Equation 3 independently for each

subgroup g within the population that calculates score gaps. For instance, creating scatterplots between $\boldsymbol{\Gamma y}_g$ and $\boldsymbol{\theta}_g$ can be a valuable technique. This subgroup analysis should encompass factors such as gender and ESCS quartiles to ensure a comprehensive evaluation of the score gaps. In many instances, the assumption underlying Equation 2 and Equation 3 may not remain valid for groups characterised by a small number of respondents. For example, when respondents are organised into groups based on schools, this grouping can potentially violate the Equation 2 and Equation 3 assumptions. Such violations may introduce bias into group-level proficiency distributions due to unexpected effects stemming from prior distributions. It becomes particularly crucial to scrutinise Equation 2 and Equation 3 when dealing with smaller groups to ensure validity.

For the effective application of LRM, it is essential to have a sufficient number of covariates and a sample size to maintain key statistical assumptions: linearity, independence, normality, and homoscedasticity in the residuals. These assumptions should remain valid across various sub-populations, such as different gender groups or quartiles of ESCS. Within the framework of ILSAs, it is recommended to have at least 500 respondents in the sample size for LRM, provided each participant responds to enough items to form $h_i(\theta)$. For a detailed exploration of how sample size influences the precision of LRM estimates, refer to Appendix A.

2.2.2. Proficiency distribution of group

If the data does not meet the conditions for conducting population modelling, it becomes necessary to estimate the proficiency distribution solely based on the likelihood function. Here, a methodology for estimating the proficiency distribution of a group is introduced. It is important to note that this modelling itself is incorporated into the process of population modelling; however, in the population-wide context, it is necessary to apply sample weights when estimating parameters to ensure the sample is representative.

Let $\boldsymbol{\Psi}$ be the parameters with respect to the proficiency distribution of a group and assume that $\boldsymbol{\Lambda}$ is given. Under the local independent assumption, the likelihood function is

$$L(\boldsymbol{\Psi}|\boldsymbol{u},\boldsymbol{\Lambda}) = \prod_{i=1}^{N}\int_{-\infty}^{+\infty} f(\boldsymbol{u}_i|\theta,\boldsymbol{\Lambda})h(\theta|\boldsymbol{\Psi})\mathrm{d}\theta$$

Since it cannot be maximised algebraically, a numerical optimisation is employed. In ILSAs, the expectation (E) and maximisation (M) algorithm (Dempster, Laird and Rubin, 1977[21]; Bock and Aitkin, 1981[22]) is employed, where a likelihood function that includes integral, namely marginal likelihood function, is maximised through iterative steps.

The conditional distribution of missing data, in this case θ, is calculated based on the observed data and provisional parameters $\boldsymbol{\Psi}^{(t)}$ as follows:

Equation 4

$$h_i^{(t)}\left(\theta|\boldsymbol{u}_i,\boldsymbol{\Lambda},\boldsymbol{\Psi}^{(t)}\right) = \frac{f(\boldsymbol{u}_i|\theta,\boldsymbol{\Lambda})h\left(\theta|\boldsymbol{\Psi}^{(t)}\right)}{\int_{-\infty}^{+\infty} f(\boldsymbol{u}_i|\theta,\boldsymbol{\Lambda})h(\theta|\boldsymbol{\Psi}^{(t)})\mathrm{d}\theta}$$

With the conditional distribution of missing data, the expected loglikelihood of the complete data is computed (E-step)

Equation 5

$$E\ln L = \sum_{i=1}^{N}\int_{-\infty}^{+\infty} f(\boldsymbol{u}_i,\theta|\boldsymbol{\Psi},\boldsymbol{\Lambda})h_i^{(t)}\left(\theta|\boldsymbol{u}_i,\boldsymbol{\Lambda},\boldsymbol{\Psi}^{(t)}\right)\mathrm{d}\theta$$

where $f(\boldsymbol{u}_i, \theta|\boldsymbol{\Psi}, \boldsymbol{\Lambda})$ is the joint distribution of the observed data $\boldsymbol{u}_i$ and the missing data θ.

The provisional parameters $\boldsymbol{\Psi}^{(t)}$ are updated by maximising Equation 5 (M-step):

Equation 6

$$\frac{\partial E \ln L}{\partial \boldsymbol{\Psi}} = 0$$

The updated parameters, $\boldsymbol{\Psi}^{(t+1)}$, is used in the next E step. Note the marginal likelihood function is evaluated at $\boldsymbol{\Psi}^{(t+1)}$ always follow

$$L\left(\boldsymbol{\Psi}^{(t+1)}|\boldsymbol{u}\right) \geq L\left(\boldsymbol{\Psi}^{(t)}|\boldsymbol{u}\right)$$

The steps E (Equation 5) and M (Equation 6) are repeated until the parameters converge.

If a distribution follows $N(\mu, \sigma^2)$, $\boldsymbol{\Psi}^{(t+1)} = \left[\mu^{(t+1)}, \sigma^{2(t+1)}\right]$, it is updated as follows:

Equation 7

$$\mu^{(t+1)} = \frac{1}{N}\int_{-\infty}^{+\infty} \theta \sum_{i=1}^{N} h_i\left(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right) \mathrm{d}\theta$$

$$\sigma^{2(t+1)} = \frac{1}{N}\int_{-\infty}^{+\infty} \left(\theta - \mu^{(t+1)}\right)^2 \sum_{i=1}^{N} h_i\left(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right) \mathrm{d}\theta$$

In practice, numerical integration is employed; thus, Equation 4 evaluated at $\theta = \theta_q$ is thus described as

$$h_i^{(t)}\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right) \approx \frac{f\left(\boldsymbol{u}_i|\theta_q, \boldsymbol{\Lambda}\right)h\left(\theta_q|\boldsymbol{\Psi}^{(t)}\right)}{\sum_{q=1}^{Q} f\left(\boldsymbol{u}_i|\theta_q, \boldsymbol{\Lambda}\right)h\left(\theta_q|\boldsymbol{\Psi}^{(t)}\right)}$$

where $q(= 1, \ldots, Q)$ is an index of quadrature points for numerical integration. Furthermore, Equation 5 is computed as

$$E \ln L = \sum_{q=1}^{Q}\sum_{j=1}^{J}\sum_{k=0}^{K_j-1} F_{jk}^{(t)}\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right) \ln p_{jk}\left(\theta_q|\boldsymbol{\Lambda}_j\right)$$

where

$$F_{jk}^{(t)}\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right) = \sum_{i=1}^{N} u_{ijk} h_i^{(t)}\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right)$$

Consequently, Equation 7 is approximated as follows:

Equation 8

$$\mu^{(t+1)} \approx \frac{1}{N}\sum_{q=1}^{Q} \theta_q \sum_{i=1}^{N} h_i\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right)$$

$$\sigma^{2(t+1)} \approx \frac{1}{N}\sum_{q=1}^{Q} \left(\theta_q - \mu^{(t+1)}\right)^2 \sum_{i=1}^{N} h_i\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right)$$

If strong assumptions about the group proficiency distribution are not desired, a generic multinomial distribution can be assumed. Under the assumption of multinomial distribution, the parameters $\boldsymbol{\Psi} = \left[\pi_1, \ldots, \pi_Q\right]$ are calculated as

Equation 9

$$\pi_q = \frac{\sum_{i=1}^{N} h_i\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right)}{\sum_{i=1}^{Q}\sum_{i=1}^{N} h_i\left(\theta_q|\boldsymbol{u}_i, \boldsymbol{\Lambda}, \boldsymbol{\Psi}^{(t)}\right)}$$

It is crucial to emphasise that Equation 7 (or Equation 8) and Equation 9 solely rely on participants' responses. Consequently, the proficiency distribution of a group, estimated through marginal maximum likelihood estimation via the EM algorithm (MMLE-EM), is solely derived from the cognitive assessment data.

2.2.3. Band score

In this context, the band score approach is introduced as an individual-level output, which involves categorising the continuous proficiency scale into discrete bands. Typically, ILSAs adopt a range of six to ten bands for this purpose, although some examinations utilise a larger number, often 30 to 40 categories, to maintain precision in the output. The decision regarding the number of bands to use should align with the objectives of the assessments.

In ILSAs, the number of proficiency bands is directly aligned with the summary descriptions of proficiency levels as outlined in the assessment frameworks. These proficiency levels within ILSAs are derived from the cognitive demands required by the assessment tasks. Thresholds are established to demarcate changes in these demands. For instance, both PISA and PIAAC employ seven to eight distinct bands. Each of these bands corresponds with detailed descriptions that articulate the capabilities of respondents at each proficiency level. This structure ensures that the bands are meaningfully connected to the cognitive skills and abilities the assessments aim to measure.

At an individual, the loglikelihood function with respect to θ is defined as follows:

Equation 10

$$\ln L(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}) = \sum_{j=1}^{J}\sum_{k=0}^{K_j-1} u_{ijk} \ln p_{jk}\left(\theta|\boldsymbol{\Lambda}_j\right)$$

Equation 10 can be optimised through Newton-Raphson method (MLE-NR), where θ is updated iteratively as follows.

Equation 11

$$\theta^{(t+1)} = \theta^{(t)} - H\left(\theta^{(t)}\right)^{-1} g\left(\theta^{(t)}\right)$$

Here, $H\left(\theta^{(t)}\right)$ and $g\left(\theta^{(t)}\right)$ are the hessian and the gradient of the loglikelihood functions (Equation 10) with respect to θ evaluated at $\theta = \theta^{(t)}$, respectively. Equation 11 is repeated until $\left|\theta^{(t+1)} - \theta^{(t)}\right|$ reaches to the criterion, and $\theta^{(t)}$ at the last cycle is employed as the maximum likelihood estimate $\hat{\theta}$. The band score will be assigned based on $\hat{\theta}$.

$H\left(\theta^{(t)}\right)$ can be replaced with the expectation of it, $-I(\theta)$, which ensures non-negative value (Hambleton and Swaminathan, 1985[23]).

$$\mathrm{E}[H(\theta)] = -I(\theta)$$

The standard error of $\hat{\theta}$ is approximated as follows.

Equation 12

$$\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$$

It shows the precision of $\hat{\theta}$, $\mathrm{V}[\hat{\theta}|\theta]$, is dependent on θ. This mathematical property holds significant importance for individual diagnosis and test form design.

For mathematical reasons, the weighted maximum likelihood estimator (WLE; Warm (1989[24])) is employed for the estimation of θ in practice. It penalises the likelihood function defined in Equation 1 with the square root of the information function $I(\theta)$ (Firth, 1992[25]).

$$wL(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}) = L(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda})\sqrt{I(\theta)}$$

Thus, the weighted loglikelihood function to be maximised is described as follows:

$$\log wL(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}) = \log L(\theta|\boldsymbol{u}_i, \boldsymbol{\Lambda}) + \frac{1}{2}\log I(\theta)$$

WLE effectively corrects bias in the estimation of the ability parameter $\hat{\theta}$. Additionally, from a practical standpoint, it ensures convergence for $\hat{\theta}$ of respondents who scored either full marks or zero marks.

Technically, the determination of the number of bands and their width should be guided by the size of $\mathrm{SE}(\hat{\theta})$ to ensure that each band's width is not excessively narrow compared to the $\mathrm{SE}(\hat{\theta})$ of the target proficiency. An approach for setting the bandwidth in proficiency assessments involves ensuring that the 95% confidence interval (CI) of a given proficiency estimate falls within the range of the adjacent score bands at all proficiency levels. For instance, if the standard error for proficiency estimates is approximately 10 points, it is advisable to set the bandwidth to more than 20 points. This approach guarantees that the 95% CI of any estimate, even those right on the thresholds between two adjacent bands, will still fall within the range of the adjacent band. This configuration minimises the risk of misclassification due to the inherent uncertainty in the estimate.

2.2.4. Pass/fail classification

Mathematically, the pass/fail classification can be regarded as a specific instance of the band score approach, characterised by only two score bands. In practical terms, during the design of such assessments, items are strategically curated to ensure that the reliability at the threshold, which separates the two categories (i.e., the pass/fail threshold), remains sufficiently high. In this subsection, we delve into the method for pass/fail classification. Similar to the band score approach, pass/fail classification is evaluated solely based on the likelihood function. The point estimate derived in Equation 11 is employed for pass/fail classification in a manner analogous to the band score approach. Specifically, a respondent is deemed to pass the threshold if $\hat{\theta} > \tau$, where τ represents the threshold value.

In many instances, the information function $I(\theta)$ exhibits an asymmetric distribution. This implies that the precision of proficiency estimation above the threshold τ varies from that below the threshold τ. Therefore, it becomes crucial to curate items in such a way that the information function $I(\theta)$ forms a symmetric distribution at the threshold τ, particularly in assessments designed for pass/fail classification. If not accounted for, the proportion of failed respondents in a group $\hat{\rho}$ may exhibit bias compared to the true proportions ρ due to the varying precisions between the two groups. Further details regarding instrument design can be found in Subsection 2.3.

Nevertheless, assembling items to achieve this symmetry is not straightforward. Therefore, this subsection introduces a methodology for correcting the bias of the proportion $\hat{\rho}$, aiming to obtain an unbiased estimator of ρ of a group.

The proportion of participants in a group, whose proficiencies follow a distribution $N(\mu, \sigma^2)$, and are categorised as "fail" can be determined using

$$\rho = \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\theta-\mu}{\sigma}\right)^2\right) \mathrm{d}\theta$$

Furthermore, the probability function of $\hat{\theta}$ being classified as "fail" based on the test form that a respondent took is described by

Equation 13

$$\phi(\hat{\theta}) = \int_{-\infty}^{\tau} \frac{I(\hat{\theta})}{\sqrt{2\pi}} \exp\left(-\frac{(z-\hat{\theta})^2}{2} I(\hat{\theta})\right) \mathrm{d}z$$

Hence, the proportion of "fail" in the group $N(\mu, \sigma^2)$ is computed as follows:

$$\hat{\rho} = \frac{1}{Z(\hat{\theta})} \int_{-\infty}^{\varphi} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{\hat{\theta}-\mu}{\sigma}\right)^2\right) \phi(\hat{\theta})\, \mathrm{d}\hat{\theta}$$

Here, $Z(\hat{\theta})$ represents the normalisation constant of the function, and φ is the adjusted threshold that provides an unbiased estimate of $\hat{\rho}$ at the group level. To find the value of φ that satisfies the condition $\rho = \hat{\rho}$, numerical integration can be employed in practical applications. Moreover, note that the estimated parameters obtained from Equation 7 can be utilised for $N(\mu, \sigma^2)$ within the framework of Equation 13. If there is no available information on $N(\mu, \sigma^2)$, it can be omitted.

2.2.5. Summary

The following Table 2 summarises the discussions of Subsection 2.2

Table 2 Approaches to estimating and reporting test scores

	Plausible values (PVs)	**Parameters of proficiency distributions**	**Band score**	**Pass/fail classification**
Content	PVs are random values drawn from a respondent's posterior distribution, which reflects their predicted proficiency from the prior distribution	Parameters of the proficiency distribution (mean and variance) for a population or group is derived from the likelihood function, using the target samples' responses to cognitive items	Band score is a categorical score, which is determined from the point estimate of the proficiency score based on the likelihood function	Pass/fail classification is a binary decision indicating if a respondent meets the threshold established by the point estimate from the likelihood function regarding the proficiency level
Using the test scores (output) for data analysis	Adequate	Not applicable	Possible	Possible
Population-level statistics	Unbiased statistics based on a proper sampling	Unbiased	Not applicable	Not applicable

Group-level statistics	Influenced by the prior distributions	Unbiased	Not applicable	Not applicable
Individual-level statistics	Influenced by the prior distribution	Although it would be a large standard error, the point estimate related to a respondent's proficiency level can still be estimated	Using the likelihood function derived from responses to cognitive items	Using the likelihood function based on responses to cognitive items that are highly discriminative around the threshold
Estimation method	Population modelling	MMLE-EM	MLE / WLE	MLE / WLE
Sample	A sample size of over 500 from the representative samples for each domain is recommended for a stable estimate of LRM parameters (See Appendix A)	To accurately estimate the proficiency distribution at the group level, a minimum of 15 to 25 respondents is required, with the specific number contingent on the number of items each respondent completes	One	One

2.3. Optimising instrument design

This subsection explores the best instrument design strategies for four distinct outputs: plausible values, group-level proficiency distribution, band score, and pass/fail classification.

2.3.1. Optimal instrument design for plausible values

An ILSA, which generates PVs, offers the most adaptable instrument design, provided enough respondents participate in the assessment. Typically, these assessments use a rotational instrument design, where item sets are rotated across test forms (van der Linden, Veldkamp and Carlson, 2004[26]). This approach arranges various sets of items and domains within test forms, enabling data collection on a wide range of items and domains. Such a design optimises the coverage of the constructs represented by the items.

An assessment designed to employ PVs can accommodate as many domains as necessary as long as it ensures a sufficient number of responses for each item. Crucially, each domain combination must have an adequate number of respondents to determine the correlation coefficients between them. As a result, including more domains in an assessment necessitates a larger pool of respondents. Typically, regardless of the total number of domains in the assessment, a respondent will engage with a maximum of three domains. It is important to note that the number of domains within each test form can differ from one test form to another.

The number of items within a domain may also differ across test forms. However, as elaborated in Subsection 2.2, there must be enough items to shape a distinct likelihood function in relation to θ to ensure stable parameter estimates for the LRM. Hence, in many instances, a test form typically centres around one or two domains, each containing 20 or more items. This ensures a reliable estimation of θ that spans a broad spectrum of the construct. Moreover, the correlations between the domains' prior distributions are determined

based on the items included in the test form. As such, the items must be well-distributed across subdomains and response formats to ensure the validity of the estimated θ.

There can be numerous test forms; however, each item or domain within a test form must be balanced regarding its item positions. Assessments designed for the population level often feature many test form patterns, accounting for the diversity of items and their positioning within the test forms. The designed test forms should be assigned randomly to respondents.

Test forms designed for PVs output are unsuited for multi-stage adaptive testing (MSAT). This is because the main objective in such assessments is not to enhance individual-level reliability but rather to maximise validity and reliability at the population model level. Furthermore, population modelling considerably enhances the reliability of proficiency estimates on an individual level, thereby reducing the benefits of implementing MSAT. However, if the item bank is sufficiently expansive, maintaining the content validity of the assessment while employing MSAT could be achievable. Otherwise, MSAT tends to prioritise reliability over the balanced content of the items, which can compromise the equilibrium of item properties for each respondent.

In this assessment, collecting comprehensive background questionnaires from each respondent is crucial to generating the covariates used in LRM. In PISA, 120 minutes are devoted to cognitive assessments and at least 30 minutes are set aside for the student background questionnaires. These questionnaires primarily consist of two types: a core one that gathers the respondent's background information and an optional one that captures their perspectives on specific areas. The background information also plays a role when categorising respondents for differential item functioning (DIF) analysis (van de Vijver et al., 2019[27]). Acquiring ample background information is vital for ensuring reliable and valid PVs in assessments aimed at producing PVs.

2.3.2. Optimal instrument design for group-level proficiency distribution

An ILSA targeting the estimation of the group-level proficiency distribution, as defined in Equation 8, provides test form designs as flexible as those for PVs. Much like the test form designs tailored for PVs, an adequate number of respondents is essential for properly estimating the proficiency's covariance structure. The distinction between the test form designs for PVs and those for aggregated proficiency distribution lies in the required number of respondents. The approach for aggregated proficiency distribution doesn't necessitate as many respondents as the one for PVs, given that it doesn't employ LRM.

Similar to the design for PVs, there can be numerous domains as long as each has a sufficient sample size to estimate the group proficiency distribution and the correlations between domains. Respondents are not required to take all domains, typically focusing on two to three. While PVs in a dataset can be employed to estimate correlations between domains, this method necessitates the estimation of correlation coefficients during the IRT modelling process. It is crucial to note that this approach is geared towards estimating the distribution of aggregated respondents directly within the modelling, not generating scores for individual respondents. Consequently, it is infeasible to determine correlation coefficients between proficiencies of cognitive domains and other variables using this dataset.

The number of items within a domain should be ample enough to form an unimodal likelihood function at the individual level. However, given that the goal is not to estimate individual proficiency levels, it is unnecessary to cover an extensive range of the construct or emphasise measurement precision. From a mathematical standpoint, capturing at least one incorrect and one correct response from each student is preferred even WLE is employed in the estimation. Consequently, from a practical standpoint, a domain ought to consist of at

least six to eight items covering a broad difficulty spectrum, with the exact minimum number depending on the breadth of the concept.

In most cases, the design and pattern of test forms are diverse, ensuring comprehensive coverage across a broad proficiency spectrum at the group level. The variation in-test form design serves not just to ensure thorough content coverage but also to cancel out any potential position effects of items and domains. Position effects can arise when the placement or sequence of items influences the responses, either due to fatigue, engagement, or recency effects. These effects can be minimised or offset by employing different order patterns for sets of items and domains.

In adopting this approach, the scope of background questionnaires is streamlined, focusing primarily on collecting respondents' demographic details since this approach does not require the LRM. This concise data collection primarily serves the purpose of enabling item invariance analysis among different demographic groups (Meredith, 1993[28]). By narrowing down the information solicited, the process becomes more efficient, and respondents may be more inclined to complete the questionnaire, potentially increasing response rates.

2.3.3. Optimal instrument design for band score

Unlike previous methodologies, determining band scores or a pass/fail classification demands a higher emphasis on ensuring reliability and validity at the individual level. However, this can be challenging given the constraints on test length that a respondent can reasonably be expected to undertake. Typically, there is a balance between test length and the quality of the data collected from the respondents. Extended tests may provide more accurate results but can also lead to respondent fatigue and reduced engagement. Conversely, shorter tests might not capture the full breadth of a participant's capabilities. As a result, in many scenarios, the granularity or resolution of individual-level outputs might be compromised, with scores or classifications possibly being rougher than ideal. This highlights the inherent challenges in designing assessments that are both manageable for participants and yield meaningful results at an individual level.

In the specified methodology, every participant must be assessed across all domains. However, a challenge arises when determining the number of items to include within each domain. The aim is to strike a balance between maintaining the test's feasibility in terms of duration and ensuring the precision of the scores derived. A proposed criterion to guide this balance is assembling items so that the 95% CIs of the estimated θ do not span more than three bands. This criterion essentially means that the true proficiency level of a respondent is likely to fall within a range of just one point above or below the estimated band score.

Techniques like adaptive testing or branching are highly effective in streamlining a domain-specific test. It is crucial, however, to ensure that the content of chosen items spans a broad spectrum, irrespective of the assessment's item selection or assembly method. The response format of these items is another vital consideration in this process since they affect performances. For an effective test design, it is essential to develop a comprehensive item bank, which includes various response formats in a diverse range of content and varying levels of item difficulty. This ensures a balanced representation of different response formats across all levels of item difficulty, contributing to a more valid assessment, especially when employing adaptive testing or an item branching approach.

Given that this approach centres on the individual level, a single test form would typically suffice. However, note that content coverage cannot be sufficient in an individual-level assessment due to the limited test length that a respondent can take. Due to the inherent nature of the assessment type, respondents might retake the test multiple times over a period (i.e., pre- and post-assessments). As a result, several test forms may be designed to

accommodate this. For this very reason, it is essential to maintain a detailed record of the respondents' participation history and the test forms they have been assigned. This ensures that individuals do not receive a test form they have already encountered in the past.

In the band score approach, background questionnaires can be excluded since none of the information contributes to the score estimation, as indicated in Subsection 2.2. Thus, while the background questionnaire is gathered, it serves analytical and reporting purposes or future item validation purposes rather than supplementing the score estimation process.

2.3.4. Optimal instrument design for pass/fail classification

In contrast to other methods, the pass/fail classification focuses on a specific proficiency level. Given that the target proficiency level is predetermined, hovering around the threshold of the pass/fail classification, it is unnecessary for an adaptive item selection strategy. The assembly of items should be geared towards maximising the reliability of measurements around this threshold. While striving for comprehensive content coverage is important, practical constraints limit the number of items a respondent can handle, often resulting in reduced content coverage within this scoring approach.

In the pass/fail classification method, respondents need to take all the domains. Given that this approach centres on a specific proficiency level, it requires fewer items in a domain compared to the assessments for the band score approach. Typically, this method features at least six items per domain, even though the target proficiency level is limited. A critical aspect of this approach is disclosing the probability of misclassification at the assumed proficiency distribution.

In this approach, though using a single test form design is feasible, employing a multiple test form is recommended for better item management. A key advantage of the multiple test form design is its ability to diminish item exposure risks. This management of item exposure is particularly vital for assessments that focus on individual diagnosis. Ensuring a uniform distribution of response formats across the various test forms is critical to the pass/fail classification approach's test form design.

For the pass/fail classification approach, there is no inherent need for a respondent's background information. Similar to the band score approach, such information can be excluded from the survey unless it serves other objectives. However, practically speaking, collecting at least the gender, age, and spoken languages at home of the respondent can be valuable for subsequent item validation procedures.

2.3.5. Summary

Table 3 summarises the discussions in Subsection 2.3.

Table 3 Optimal instrument designs for the defined outcomes

	Plausible values (PVs)	**Proficiency distributions**	**Band score**	**Pass/fail classification**
Plan	Maximise both content coverage and reliability for the population	Maximise both content coverage and reliability for each group	Maximise the reliability of the proficiency level estimate. Yet, content coverage is compromised	Maximise the reliability around the threshold. Yet, content coverage is compromised
Test form pattern	Rotational test forms (+MSAT)	Rotational test forms (+MSAT)	Multiple test forms / Adaptive testing	Multiple test forms / Single test form for each threshold of proficiency levels

Adaptive approach	Multi-stage adaptive testing is well-suited for the purpose	Multi-stage adaptive testing is well-suited for the purpose	It is recommended to employ adaptive testing or item branching	Not employed
Domain	Each respondent receives a specific subset of the available domains	Each respondent receives a specific subset of the available domains	Each respondent covers all target domains	Each respondent covers all target domains
Number of items	Every student should receive as comprehensive content coverage as possible. Additionally, item assignments should be well-balanced concerning subdomains and response formats at the population level	Every student should receive as comprehensive content coverage as possible. Additionally, item assignments should be well-balanced concerning subdomains and response formats at the population level	If an adaptive approach is employed, there should be more than eight items per domain. Otherwise, a test form should be designed to encompass the 95% CI within neighbouring bands	A minimum of six items, with highly discriminative power around the threshold, is required per domain
Background questionnaires	A sufficient number of background questionnaire items are needed for estimating stable LRM parameters	Ensure demographic variables (e.g., language, gender, age, etc.) are available to assess DIF	Not necessarily needed	Not necessarily needed

3. Towards more flexible assessment

Section 3 delves into the technical procedures for enhancing flexibility in ILSAs. Drawing from the preceding discussions, it is evident that the framework, the scale, and the meticulously developed items of ILSAs are of exceptional quality. Such elements hold the potential for versatile applications beyond their original intent, which is to obtain representative statistics of countries/economies at the same period to capture a snapshot of the education systems of the countries/economies at a time point. This section extends beyond foundational considerations to delve deeper into methodological innovations tailored for enhancing assessment adaptability. Subsection 3.1 elucidates the pivotal elements fundamental to the processes of item bank development. Subsequent to this, Subsection 3.2 embarks on a detailed comparative analysis, discerning the procedural distinctions between the two aforementioned distinct assessment modalities. Furthermore, Subsection 3.3 delves into the technicalities of item validation and parameter estimation, detailing how new items are scaled according to an internationally standardised scale. Building upon the foundations from Subsection3.3, Subsection 3.4 shifts focus to in-test trialling. This technique, applied within a primary study, facilitates item trials, particularly within intermittent assessments. Lastly, Subsection 3.5 underscores the significance of maintaining an expansive item bank in ILSAs. It further elucidates a method effective for amplifying the size of an item bank.

3.1. Process of test implementation

3.1.1. Assessment design

The first step in the assessment design phase is developing the assessment framework. This step outlines the intended areas or skills to be measured and how they should be assessed. Essential activities in this stage encompass literature reviews, consultations with domain experts, and preliminary investigative studies. Equally paramount is the identification of the

target age cohort at the same time as developing the assessment framework, ensuring that the assessment is tailored appropriately. Engaging diverse user categories during this framework definition process is indispensable. This inclusive approach ensures the assessment garners wider acceptance among its intended audience and stakeholders.

The next step is defining the target level. During this step, the primary goal is to determine the level at which assessment results will be reported. Typically, there are three primary options: the population level, which focuses on broader entities like countries/economies; the group level, which narrows the focus to entities like schools; and the individual level, which offers a more personalised assessment. Most ILSAs typically concentrate on the population level, aiming to produce statistics that accurately reflect previously determined target populations, requiring probabilistic sampling techniques to ensure the representativeness of the results. On the other hand, some assessments are designed for a group level to gain insights on a level more specific than the population but broader than individuals. Regardless of the chosen level, ensuring that subsequent methodologies are designed with this decision in mind is crucial. Additionally, this stage involves defining and refining deliverables and output formats to align with the chosen target level.

With the foundational definitions in place, the focus transitions to refining the test designs. Key decisions in this stage relate to the mode of data collection, the length of background questionnaires, the selection of testing devices or platforms, and the blueprinting[4] of the item bank. Test form design is also defined in detail in this phase. A fundamental consideration inherent to this phase is the frequency of test administration. While ILSAs typically operate on a cycle spanning three to four years, some on-demand assessments, such as TOEFL[5] and IELTS[6], are more sporadic and contingent on demand. This variation significantly influences test management procedures, with a specific emphasis on the item bank development process.

3.1.2. Instrument development

The first step in the instrument development phase is item development. Upon finalising the assessment framework in the assessment design phase, the next step is to develop items (or questions) that assess the specified construct under the blueprint of the item bank. This process typically starts with drafting items, followed by rigorous reviews for content accuracy and any potential bias (Schedl and Malloy, 2014[29]). Next, a pilot run of the items may be conducted with a selected group. If the assessments are multilingual, an added dimension to the process involves translating and adapting the test materials for each language. In the context of LSAs, this entire procedure might undergo multiple iterations, engaging diverse team members.

The second step is assembling test forms. During this step, items are organised into test forms to maximise the reliability and validity of the output defined in the assessment design phase. For instance, an assessment focusing on population-level statistics employs a rotational test form design, where it assembles various sets of items and domains into distinct test forms, ensuring comprehensive coverage of each domain's constructs and yielding valid outputs with the limited number of items each respondent takes. Conversely, when assessments are intended for individual diagnosis, test forms typically contain more items to guarantee adequate reliability and validity at the individual level.

[4] The systematic process of designing and organising a collection of cognitive items or questionnaires to ensure that they comprehensively cover the content and objectives of the assessment.

[5] Test of English as a Foreign Language

[6] International English Language Testing System

3.1.3. Item validation and parameter estimation

At the beginning of the item validation and the parameter estimation phase, item trialling takes precedence. It validates the functioning of assessment tools. The main focus here is to test how well items function and to confirm the accuracy of scoring guidelines. For ILSAs, it is vital to ensure comparability across different languages, known as item invariance.

Although it is considered that item trialling is always conducted before the main study, this is not always the case. When skipped, the quality of items is determined using data from the main study itself. Items with inadequate quality are then removed from the final or future analysis. However, since many items often don't meet the required standards, it is generally advisable to try them beforehand. Furthermore, field trials are useful to test the overall flow of the assessment and the reliability and functioning of the computer delivery systems.

The data collection stage, also referred to as test administration, is crucial. Here, monitoring the number of participants and their response rates is vital, as these factors significantly shape the study's overall design and accuracy. Additionally, the marking process begins in this stage. Ensuring that scores marked by different human markers are consistent across the board is paramount.

Upon completing the data collection and marking stages, attention transitions to the tasks of scaling or standard setting. If the assessment employs the item response theory (IRT) as its scaling method, estimating or validating item parameters is necessary[7]. With the parameters in place, target proficiencies are estimated. This process in ILSAs culminates in presenting the proficiency of the target on an international scale.

3.2. Periodic assessment and sporadic assessment

ILSAs can be characterised by their unique approach to assessment administration, dividing them primarily into periodic and sporadic assessments. Periodic assessments are consistently scheduled at regular intervals. This approach relies on a sequential item bank management system where items are methodically developed for the next assessment cycle and validated for their item psychometric properties through a field trial. Countries participating in these assessments adhere to a standardised timeline, ensuring synchronicity in administration. In contrast, sporadic assessments offer a more adaptable model. They are conducted when needed, based on specific demands. This approach sees a continuous flow in item development, with items being integrated into assessments once validation process is completed. A distinguishing feature of sporadic assessments is the timeline autonomy they grant participants, allowing for a more versatile administration process.

During the assessment design phase, two significant distinctions emerge between periodic and sporadic assessment types. Firstly, there's the matter of the target level. Periodic assessments primarily aim at the population level, ensuring high-level comparability across participating countries. Sporadic assessments, in contrast, display versatility in their target levels, adapting to user-specific demands, although mainly targeted at the group and individual levels. The second differentiation lies in the flexibility accorded to framework refinement. Since periodic assessments require all countries to align with a shared timeline, their framework must be established several years before the main study. By benefiting from their adaptability, sporadic assessments can incorporate and act upon refined frameworks.

[7] For those interested in a more detailed exploration of this estimation or validation process, reference materials like OECD (OECD, 2017[3]; OECD, n.d.[30]) offer comprehensive insights.

Both assessment types follow distinct procedures during the instrument development and item validation and parameter estimation phases. These phases progress simultaneously across all participating entities for periodic assessments targeting the subsequent main study. In this approach, new items—without estimated item parameters—are compiled into test forms and subjected to a field trial to evaluate their psychometric properties. Only those that meet the psychometric standards are considered for the main study. Trend items[8] may be excluded from these field trial test forms. The primary aim here is to assess the functionality of the items. Due to the tight timeline, items demonstrating poor psychometric properties during the field trials are typically discarded rather than revised. As such, in periodic assessments, the field trial primarily serves as a filter to select items for the main study.

In sporadic assessment, the phases of instrument development and item validation and parameter estimation proceed separately across the participating entities. Following the meticulous drafting and revision process, new assessment items are initially subjected to linguistic translation for a select cohort of languages corresponding to those countries and start testing soon. After this, these new items undergo item trialling within these countries, the primary objective being evaluating their psychometric properties. Should these items align with the stipulated psychometric standards, they are expanded to include translations in other languages. Conversely, when these items demonstrate a deviation from the desired metrics, they are either subjected to further iterative refinement or are unequivocally discarded. Any modifications or refinements to these items must be fundamentally anchored in findings extracted from the psychometric evaluations of the novel items, utilising data amassed from their preliminary deployments. Notably, in many sporadic assessments, the field trial phase is seamlessly integrated into the main study, often referred to as in-test trialling.

Sporadic assessments offer flexible assessment administration and streamlined item bank development. However, they also demand consistent and adaptable test form assembly and item revisions, depending on the outcomes of item trialling within the item bank. To accommodate the unique demands of sporadic assessments, a computer-based testing (CBT) delivery mode becomes essential, given its capability for managing item exposure of the new items through regular test form reconfigurations.

3.3. Technical standard for item validation and parameter estimation

The technical process of item validation and parameter estimation in ILSAs stands as a cornerstone in the scaling procedure. This process primarily encompasses item validation and item parameter estimation. During validation, there is a comparison between the expected item functioning (specifically, ICRF) and the pseudo-observed frequency. This comparison is executed for every country/economy separately. New items, or those exhibiting a mismatch between the ICRF and the observed data, undergo item parameter estimation.

This procedure is applicable to data from various levels of assessment, including both group and individual levels A crucial requirement for the data utilised in the item validation and parameter estimation phase is that it must encompass a sufficient number of item responses to comprehensively represent the concept being measured. Without this breadth in item response, the pseudo-observed frequencies calculated from data may lack validity and reliability. Additionally, data consolidation is feasible, provided that the data are collected under identical conditions. In practice, employing data from individual-level assessments can be challenging, as these assessments are not typically designed to encompass a broad

[8] Items that are repeatedly used across multiple test cycles to maintain comparability of the scale.

spectrum of the target concepts at an aggregated level. Conversely, group-level assessments are more likely to meet these conditions, making their data more suitable for use in this context.

Each parameter is estimated to be consistent across countries/economies, called "international parameters" for new items. Conversely, items from a specific country/economy that don't align well with model-data fit are estimated without the parameter constraint. These are termed "national parameters." This model is recognised as a multi-group model with the partial invariance assumption, and the MMLE-EM technique is employed for its parameter estimation. This subsection examines the crucial technical standards essential for item validation and parameter estimation, emphasising the factors that influence the quality of item parameters. The following discussions will highlight an optimised test form design curated for item validation and parameter estimation in sporadic assessments.

In the scaling framework, $\boldsymbol{\Lambda}^*$ represents the fixed item parameters associated with trend items, while $\boldsymbol{\Lambda}_g$ denotes the parameters for new items corresponding to country/economy $g\ (1, \ldots, G)$. It is pertinent to note that $\boldsymbol{\Lambda}^*$ can be either international, unique, or national parameters. $\boldsymbol{\varphi}_g = [\mu_g, \sigma_g^2]$ denotes parameters related to the proficiency distribution for a given country/economy g, which is assumed to follow normal distribution.

The likelihood function that should be maximised within the model is defined as:

Equation 14

$$L(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}, \boldsymbol{\varphi} \mid \boldsymbol{u}) = \prod_{g=1}^{G} \prod_{i=1}^{N_g} \int_{\theta} f\left(\boldsymbol{u}_{gi} | \theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g\right) h\left(\theta | \boldsymbol{\varphi}_g\right) \mathrm{d}\theta$$

In this context, $f\left(\boldsymbol{u}_{gi} | \theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g\right) h\left(\theta | \boldsymbol{\varphi}_g\right)$ signifies the joint distribution of the observed data $\boldsymbol{u}$ and the missing data θ. Note $\boldsymbol{\Lambda}^*$ is not a parameter to be maximised, it is already given. Unlike the likelihood function defined in the previous section, the parameters to be maximised when scaling new item parameters are both $\boldsymbol{\Lambda}_g$ and $\boldsymbol{\varphi}_g$, for all g.

To maximise Equation 14, MMLE-EM is employed. Within this algorithm, $\boldsymbol{\varphi}_g$ and $\boldsymbol{\Lambda}_g$ are updated iteratively. Specifically, the missing data is replaced with its expectation, based on the provisional parameters $\boldsymbol{\varphi}_g^{(t)}$ and $\boldsymbol{\Lambda}_g^{(t)}$ at the t-th cycle of E and M iteration.

During the E-steps, the conditional distribution of the missing data is computed, taking into account the provisional item parameters $\boldsymbol{\Lambda}_g^{(t)}$ and fixed-parameters $\boldsymbol{\Lambda}^*$:

Equation 15

$$h_{gi}^{(t)}(\theta) = h_{gi}\left(\theta | \boldsymbol{u}_{gi}, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}, \boldsymbol{\varphi}_g^{(t)}\right) = \frac{f\left(\boldsymbol{u}_{gi} | \theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}\right) h\left(\theta | \boldsymbol{\varphi}_g^{(t)}\right)}{\int_{\theta} f\left(\boldsymbol{u}_{gi} | \theta, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}\right) h\left(\theta | \boldsymbol{\varphi}_g^{(t)}\right) \mathrm{d}\theta}$$

Additionally, the expected frequency for each category k of item j is determined using the conditional distribution of θ as highlighted in Equation 15.

Equation 16

$$F_{gjk}^{(t)}(\theta) = \sum_{i=1}^{N_g} u_{gijk} h_{gi}^{(t)}(\theta)$$

In Equation 14, responses are substituted by the expectation $F_{gjk}^{(t)}(\theta)$.

Equation 17

$$\mathrm{E}\ln L_g\left(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}, \boldsymbol{\varphi}_g^{(t)} | \boldsymbol{u}_{gi}\right) = \int_\theta \sum_{j=1}^{J} \sum_{k=0}^{K_j-1} F_{gjk}^{(t)}(\theta)\ \ln p_{gjk}^{(t)}(\theta)\,\mathrm{d}\theta$$

The M-steps involve optimising Equation 17 with respect to $\boldsymbol{\Lambda}_g^{(t)}$ first and then $\boldsymbol{\varphi}_g^{(t)}$ based on $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Lambda}_g^{(t+1)}$ (Equation 7 or Equation 8). In the context of ILSAs, the partial invariance assumption is set for $\boldsymbol{\Lambda}_g$; therefore, item parameters are constrained to be equal among countries/economies unless the model-data fit is poor. The discrepancy between the model and data is called differential item functioning (DIF). Note that the parameters of each item $\boldsymbol{\Lambda}_{gj}^{(t)}$ is maximised independently from other items using Newton-Raphson method. This cycle of E and M steps continues until convergence. The model composes the computation of $\mathrm{E}\ln L_g\left(\boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}_g^{(t)}, \boldsymbol{\varphi}_g^{(t)} | \boldsymbol{u}_{gi}\right)$ of all countries/economies within an EM cycle because of the parameter constraints set in the model (i.e., partial invariance assumption). For details of the estimation, see OECD (2017) and Okubo (2022).

The DIF is checked through RMSD (root mean squared deviation), formulated as follows:

$$\mathrm{RMSD}_{gj} = \frac{1}{K_j}\sum_{k=0}^{K_j-1}\sqrt{\int_\theta \left(o_{gjk}(\theta) - p_{gjk}(\theta)\right)^2 f_g(\theta)\,\mathrm{d}\theta}$$

where $f_g(\theta)$[9] is the proficiency distribution of country/economy g (OECD, n.d.[30]). Here, $o_{gjk}(\theta)$ is the pseudo-observed frequency calculated based on Equation 15; namely,

$$o_{gjk}(\theta) = \frac{\sum_{i=1}^{N_g} u_{gijk} h_{gi}(\theta)}{\sum_{k=0}^{K_j-1}\sum_{i=1}^{N_g} u_{gijk} h_{gi}(\theta)}$$

In PISA cognitive domains, 0.12 is set as the cut-off criterion for RMSD (OECD, 2017[3]), while 0.15 is employed in PIAAC (OECD, 2014[2]).

Equation 17 reveals that during the maximisation of $\boldsymbol{\Lambda}_g^{(t)}$, θ is integrated out from the likelihood function, obviating the need to factor in $\boldsymbol{\varphi}_g$ while estimating parameters for new items. This further suggests that item parameters can either be validated or estimated using convenient samples and do not require the representativeness of the samples. However, in practice, samples should cover a wide range of proficiency, and the item difficulties should be distributed in a way that is widely aligned with the samples.

According to Equation 16 and Equation 17, $\boldsymbol{\Lambda}_g$ is derived from the expectations computed using the conditional distribution of the missing data, $h_{gi}(\theta)$, which, in turn, is dependent on the fixed item parameters $\boldsymbol{\Lambda}^*$. The integrity of $h_{gi}(\theta)$ is therefore crucial for an unbiased estimation of $\boldsymbol{\Lambda}_g$. Ensuring a robust estimate necessitates the inclusion of a substantial number of trend items in a test form that spans the entirety of the construct, facilitating the computation of a valid and reliable $h_{gi}(\theta)$. The ratio of new items to trend items does not impact the quality of measurement; however, the most important factor is the inclusion of an

[9] The function traditionally used in assessments can be effectively substituted with $f_{gj}(\theta)$, which represents the proficiency distribution of respondents who answered item j in country/economy g. This substitution is particularly beneficial in assessments that utilise an adaptive testing approach.

adequate number of trend items with appropriate psychometric properties into a test form to yield a valid and reliable $h_{gi}(\theta)$. To monitor and verify this integral component, the information function of trend items within a test form becomes an indispensable tool.

Conversely, the number of new items in a test form, specifically the number of item parameters ($\boldsymbol{\Lambda}_g$) to be maximised, does not influence the reliability or unbiasedness of the $\boldsymbol{\Lambda}_g$ estimates. This stems from the fact that each $\boldsymbol{\Lambda}_{gj}^{(t)}$ is maximised independently of other items within an EM cycle. Consequently, the number of $\boldsymbol{\Lambda}_{gj}^{(t)}$ to be maximised during an EM cycle does not relate to the estimation's quality. There is flexibility to incorporate as many new items as needed, provided that $h_{gi}(\theta)$ maintains its reliability and validity. However, from a practical standpoint, it is more critical to have a higher number of trend items in a test form than to include more new items to ensure the quality of $h_{gi}(\theta)$.

The quality of $\widehat{\boldsymbol{\Lambda}}_{gj}$ is significantly influenced by the number of respondents for the item. Conventionally, 200-500 respondents are considered the minimum number for accurately estimating $\boldsymbol{\Lambda}_{gj}$ depending on the psychometric property of the target measure (Waller, 1981[31]). This is under the assumption that every respondent engages with an ample assortment of trend items, ensuring a comprehensive coverage of the construct. It is also essential that item difficulties correspond well with the proficiency levels of the respondents. The discussions on the sources of the error in a factor analysis model can be found in MacCallum and Tucker (1991[32]).

In the context of ILSAs, it is critical to ensure that constrained parameters, particularly international ones, are estimated without bias from countries/economies that exhibit DIF. This underscores the importance of a rigorous DIF detection process. Mathematically, as the number of countries/economies in the dataset increases, the likelihood of identifying DIF properly also rises. For reference, the proportion of invariant items (comprising both trend items and new items) of the PISA reading domain ranged from 70% to 95% across 70 countries, with an average of 88% (OECD, n.d.[30]). Given the pivotal role that DIF item detection plays in the scaling of ILSAs, it is strongly recommended that data collection efforts encompass a diverse range of countries/economies, particularly those representing varied language groups.

3.4. In-test trialling

Within sporadic assessments, item trialling is conducted using the in-test trialling approach in most cases. This method embeds test items, specifically those necessitating either item parameter estimation or an item validation process, directly into the test forms designed for the main study. A significant benefit of in-test trialling is its ability to facilitate item validation and parameter estimation without needing an additional, distinct field trial, thus alleviating potential logistical strains on test administrators. This subsection aims to delve into the structure and the procedure of the in-test trialling technique.

When designing in-test trialling, the main factors to be considered in the context of ILSAs are the number of new items and the trend items in-test forms, the number of required respondents per new item, and the number of countries/economies required for the item invariance analysis.

Firstly, the number of new items and the number of trend items in a test form are the essential factors to be considered. The more new items are inserted into a test form, the more effective data collection is. On the other hand, the number of trend items, or the information function constituted by the trend items, ensures the quality of the scale, as indicated in Subsection 3.4.

Ideally, the trend items in a test form should have sufficient content validity and a desirable shape of information function for a respondent. Furthermore, the construct coverage of the trend items presented with the new items should be well-balanced at the group of respondents and an individual level; therefore, new items should be assembled into multiple test forms. However, too many items in a test form jeopardise the respondents' engagement; thus, the number of new items inserted into a test form is compromised as it can be covered by increasing the total number of respondents who participate in an assessment. As mathematically explained in Subsection 3.4, the content coverage of new items within a test form does not influence the validity of the estimated parameters.

Second, the number of responses garnered by new items plays a pivotal role in ensuring the reliability of parameter estimates. As highlighted in Subsection 3.4, for a cognitive item's parameter estimation, there is a baseline requirement of 200 valid responses. Moreover, the proficiency spectrum of the respondents must reflect the broader population, even though the sample's exact representativeness is not necessary. Upon reaching the designated number of responses for a new item (for instance, 200), one can derive provisional parameters using MMLE-EM. Should the standard errors of the item estimates be large, it is prudent to persist in gathering responses. Conversely, if they are not notably large, examine the profile of the pseudo-observed frequency $o_{gjk}(\theta)$. If this $o_{gjk}(\theta)$ exhibits stable functions, conclude the data collection for that item in the country/economy, or continues the data collection if not.

Lastly, the data collection process persists until data has been adequately sourced from a sufficient number of countries/economies. In ILSAs, verifying item invariance across different countries/economies is pivotal, ensuring that scores derived from different regions remain comparable. After the provisional parameters of items are estimated within one country, the cognitive items are translated into another language. Thereon, responses to these translated items are continuously collected until two conditions are met: a low standard error and a stable $o_{gjk}(\theta)$ spanning various proficiency levels. This iterative process persists until the ICRF formulated using the provisional parameters from different countries, aligns into a specific trajectory. Following the estimation of many of these provisional parameters, the item's parameter is estimated via the multi-group IRT model, incorporating the partial invariance assumption, as delineated in Section 3.4. Whether via international or national parameters, the finalised parameters estimated at this juncture signify the item's validation and parameter estimation, rendering it fit for inclusion in the assessment.

3.5. Item cloning

The earlier subsections delved into several key topics: the general process of item bank development, the introduction of an intermittent assessment approach, enabling users to engage with the assessment as needed, the essential technical criteria for initiating newly crafted items, and a methodology for incorporating new items within the framework of sporadic assessment. This subsection will illuminate a methodological approach geared towards the sustainable expansion of an item bank.

Item cloning, a useful technique in item development, facilitates the expansion of an item bank without compromising the consistent evaluation of foundational constructs. This strategy involves developing new items from existing ones by altering specific elements. Commonly, while the stimuli or context remains the same, the stem undergoes modification. Another prevalent adaptation involves shifting the response format, such as transitioning from a multiple-choice to an open-ended response or vice versa. Alternatively, altering the options within a multiple-choice item can also serve the purpose. Notably, these cloned items can exhibit significant variances in their psychometric properties compared to their originals.

To ensure a diverse range of response formats and to cover the target measure comprehensively, it is advisable to create a substantial number of cloned items. Ideally, this involves generating 30 to 50 variations for each original item. This can be achieved by varying the patterns of response formats, offering different options in multiple-choice formats, and altering the stem of the item. Additionally, these variations should be developed with an eye toward creating different levels of item difficulty to enhance the item bank.

Several advantages are inherent to this technique. Primarily, generating cloned items tends to be faster and more streamlined than developing entirely new ones. This method also upholds a strong sense of construct validity, given that the foundational structure of the original item remains the same. Despite the modifications, this approach curbs test-takers from merely memorising items since there are various similar items, thus enhancing item security. The diversity introduced by these slight alterations results in a richer item set. From a financial standpoint, item cloning is often less expensive as it leverages previously validated items, saving resources typically used to validate new ones.

Developing multiple versions of an original item is especially beneficial from a content point of view. In essence, item cloning is not just a methodological strategy but also a tool that aids in discerning deeper layers of the item's structure. As item cloning becomes more ingrained in the item bank, the demand for an item management tool intensifies. No respondent must be presented with multiple cloned items in a single test session (i.e., enemy items). This is crucial for ensuring broad content coverage and for adhering to the local independence assumption of the presented items. Thus, an adept item management tool, seamlessly integrated with the item bank, becomes indispensable. It provides rigorous oversight during the assembly of test forms, ensuring that cloned items are judiciously spread out. Additionally, even when cloned items undergo minor tweaks, it is essential to rigorously evaluate their psychometric properties. Thus, a robust item management tool is paramount when leveraging the item cloning technique.

4. Data illustration: PIAAC Education and Skills Online

4.1. Education and Skills Online

4.1.1. Outline

In the context of the Programme for the International Assessment of Adult Competencies (PIAAC), there is an increasing demand for a PIAAC-linked assessment suitable for research, diagnostic, or follow-up studies. Researchers and educators seek an assessment tool that delivers individual-level results and ensures these outputs align with the PIAAC proficiency scales. Furthermore, there is a call for these results to be benchmarked against the national and international results of the PIAAC participating countries.

Education & Skills Online (E&S Online) is an assessment instrument developed to produce individualised outcomes seamlessly integrated with the proficiency scale of PIAAC based on Cycle 1 data. This encompasses domains of Literacy, Numeracy, and the skill of Problem-Solving in technologically-rich settings. E&S Online ensures that every outcome can be compared to both national and international benchmarks available for all participant countries/economies of PIAAC. In addition to its primary cognitive assessments, it explores a range of non-cognitive dimensions. This comprehensive approach sheds light on areas such as skill use, career interest, health, and well-being (OECD, n.d.[33]).

E&S Online caters to a wide and varied user base. Organisations specialising in adult literacy and numeracy training can leverage the assessment tool to gauge the proficiency levels of

learners, comparing training outcomes to both national and international standards. Research professionals also benefit from the assessment's offerings, as the tool provides tests that align with PIAAC benchmarks, ensuring accuracy and relevance in their studies. On the governmental front, agencies can deploy E&S Online to discern the educational needs of specific demographics, such as the unemployed, those at potential risk, or adults from economically challenging backgrounds. In essence, E&S Online's adaptability makes it an indispensable tool for various stakeholders in the education and employment sectors.

Subsection 4.1 provides a detailed overview of the test form design and the psychometric characteristics of the original version of E&S Online. This foundation is essential for users of the assessment tool, as it allows for a clear understanding of how the assessment functions and the validity of its measurements. By establishing the groundwork with this introduction, Subsection 4.1 sets the stage for Subsection 4.2, where new versions of E&S Online will be presented. The comparison between the original and new versions is crucial for highlighting improvements, changes in functionality, and any potential impacts on methodology and interpretation that may result from these updates.

4.1.2. Test form design

The original E&S Online is designed to offer a comprehensive assessment experience, expected to require around 120 minutes for full completion. This duration includes core domains of Literacy and Numeracy, supplementary reading components, and problem-solving exercises in technology-rich environments, in addition to the assessments of non-cognitive skills. The core part of the test, including a background questionnaire and assessments for Literacy and Numeracy, is estimated to take approximately 65 minutes. Users can choose whether to engage with additional optional modules depending on their individual needs. The assessment is facilitated via a CBT platform.

Table 4 shows the test form design of the core part of the original E&S online. The structure of each test form within the assessment consists of three distinct stages. The original E&S Online utilises an MSAT comprising three stages. In this setup, the test forms are created by combining three item clusters from Stage 1 and three item clusters from Stage 2. It begins with a preliminary stage containing three items each for Literacy and Numeracy, serving as a sorting mechanism. This is followed by the first stage, which presents a more substantial set of nine items for each domain. The assessment concludes with the second stage, comprising 11 items for each of Literacy and Numeracy. Each of the latter stages is allocated 30 minutes. A respondent completes 23 items for each domain, assessing their competencies in these core areas. It is important to note that the assessment employs the MSAT approach.

Table 4 Test form design of the core part of the original E&S Online

	Literacy	Numeracy	Total
Test forms	9	9	9
Number of stages			3
Adaptive / Branching approach			Multi-Stage Adaptive Testing (MSAT)
Number of items (on average)	23	23	46
Number of units (on average)	11.3	16.7	28
Testing time	About 30 minutes	About 30 minutes	About 60 minutes

4.1.3. Psychometric property

Figure 1 and Figure 2 offer a visual representation of the standard errors and the 95% CIs of proficiency estimates on the PIAAC scale for Literacy and Numeracy as derived from the E&S Online assessment. In Figure 1, each coloured line represents the standard error across a range of abilities for different Literacy test forms introduced in Table 4, illustrating the precision of each test form's measurement capabilities. Similarly, Figure 2 shows the precision for Numeracy. These figures are particularly valuable at the individual level, underscoring the reliability of scores that policymakers, educators, and researchers rely on for crafting educational interventions and policy decisions based on the scores.

In the original version of E&S Online, Literacy and Numeracy are scored on the PIAAC international scale which ranges from 0 to 500. These scores are assigned in 10-point increments (i.e., band score), providing a detailed gradation of a respondent's proficiencies within each domain. To facilitate interpretation and application of the results, scores are categorised into one of five proficiency levels. These levels are defined by the complexity and type of cognitive skills required to perform tasks associated with each level (OECD, 2021[34]). In E&S Online, the highest proficiency levels, Levels 4 and 5, are combined into a single category. It is important to note that the scale employed in E&S Online is identical to the one used in the original PIAAC.

The precision of the Literacy test forms in the original E&S Online assessment, as indicated in Figure 1, varies significantly across the proficiency levels, largely due to the MSAT approach used in the assessment. The standard errors range from as low as 12 to as high as 50, depending on the proficiency level being assessed. Notably, the assessment provides the greatest precision at the threshold between Levels 2 and 3, where the standard errors are between 12 and 18, varying with each specific test form. This variation in precision underscores the importance of the MSAT design, which tailors the difficulty of the questions to the respondent's proficiency level, thus affecting the reliability of the score at different points along the proficiency scale. Given that the standard errors primarily fall between 12.5 and 37.5 across proficiency Levels 1 to 4, the 95% CIs for the proficiency estimates vary substantially. The interval may extend approximately from 50 points to 150 points for a given proficiency level.

The standard errors associated with the Numeracy test forms exhibit significant variability across different proficiency levels. They span from a minimum of 15 to a maximum of 50 in correspondence with the cutoff scores for proficiency bands, varying in accordance with the specific proficiency level assessed. The tool yields the most precise results at the threshold between Levels 2 and 3, with standard errors approximately at 15. The data illustrated in Figure 2 indicates that the precision is higher for lower proficiency levels than for higher ones. Considering the standard errors, which predominantly range from 15 to 35 across proficiency Levels 1 to 4, there are variations in the 95% CIs for proficiency scores. These intervals span from 60 points to 140 points.

Figure 1 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the Literacy domain of the original E&S Online

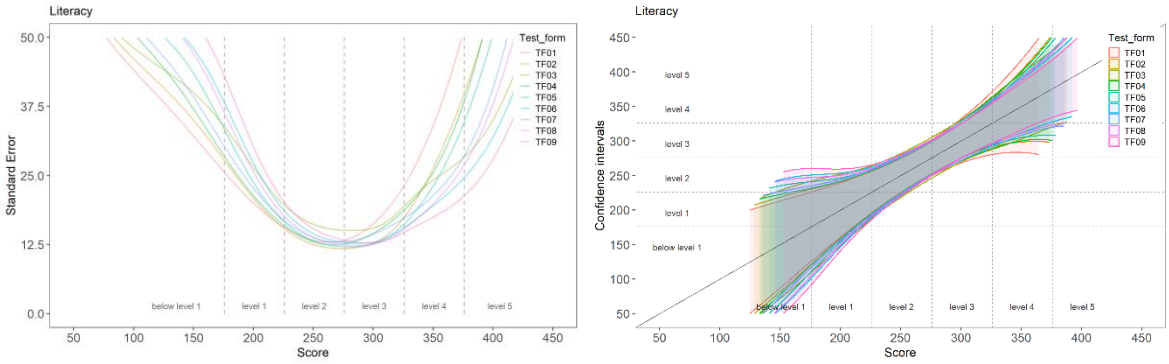
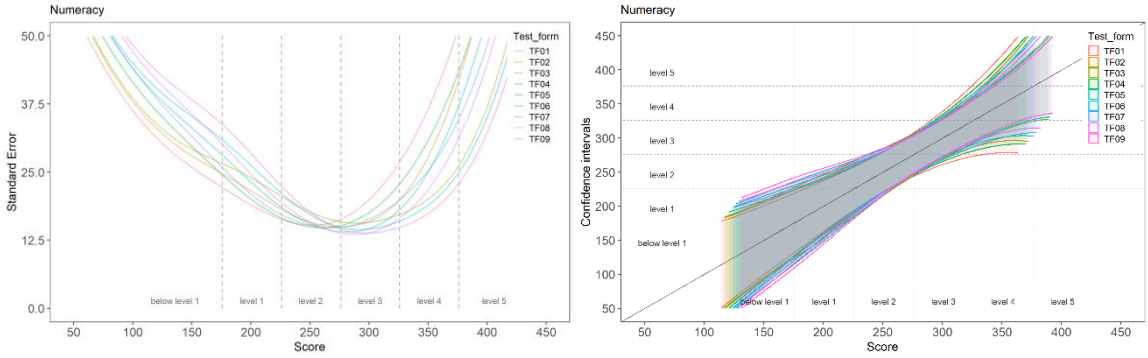


Figure 2 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the Numeracy domain of het original E&S Online



4.2. Prototype of new Education & Skills Online

4.2.1. Outline

The revitalisation of E&S Online is in direct response to its persistent demand and the need for alignment with the latest assessment frameworks for PIAAC Cycle 2. This proposed enhancement is designed to update and synchronise the cognitive assessments with the newest framework, thereby boosting their relevance and comparability. Such an update is not just timely but essential, considering the recent revisions in Literacy and Numeracy frameworks and the incorporation of adaptive problem-solving as an entirely new domain. Beyond updating the content domains, the revision also proposes technical and strategic improvements. This entails the adoption of a modular design for the assessment, which is expected to provide a more personalised assessment for users.

Crucially, the new versions introduce a multi-purpose design that caters to the diverse requirements of individual users and institutions. These proposed versions, outlined in this subsection and summarised in Table 5, are still under development and will be refined during this phase. Marking a departure from the existing model, users and institutions will have the flexibility to select which cognitive domains to assess, choosing from Literacy, Numeracy, and Adaptive Problem-Solving, with each choice varying in assessment duration and the precision level of proficiency estimation for the test-takers.

Subsection 4.2 presents two prototypes of the E&S Online upgrades. These are specifically tailored to deliver results at the individual level, offering a straightforward pass/fail classification and a more detailed band score approach for evaluation. Such enhancements are geared towards amplifying the effectiveness of E&S Online, positioning it as a more powerful instrument for assessing and fostering the growth of educational competencies.

The first prototype, the Certificate version, is developed to provide a short assessment, yielding individuals a binary pass/fail outcome using one of the PIAAC proficiency thresholds. This variant of E&S Online is particularly designed for efficiency, requiring as little as about 12 minutes to evaluate proficiency for each domain for this purpose. This version caters to individuals or organisations in need of rapid diagnostics to determine proficiency levels in cognitive domains.

The second prototype, the Distribution version, utilises a band score method, offering a more granular performance metric across predefined proficiency bands aligned with the PIAAC proficiency levels. This version requires approximately 20 minutes for assessing a single domain, with an additional 12 minutes for each extra domain assessed. The distribution version provides a richer, more nuanced profile of an individual's proficiencies, making it ideal for users who require a detailed assessment of their skills with a short assessment tool.

While the Distribution version is tailored primarily for individual assessment, its versatility allows it to aggregate data for a group, provided there is a sufficiently large sample. This feature grants the distribution version dual functionality, enabling it to serve both individual diagnostic purposes and broader organisational needs. This adaptability significantly broadens the scope and applicability of E&S Online, making it a potent tool for both personal and professional development contexts.

Table 5 New Education & Skills Online: Certification version and Distribution version

	Certification version	**Distribution version**
Output (individual level)	Outcome: A respondent is classified as "Pass" or "Fail" at the specified target classification level. Reporting: An individual report is issued to each respondent, detailing their performance and outcome in the assessment	Outcome: Each respondent is allocated to a proficiency band within the assessment framework that reflects their level of performance. Reporting: A personalised report is generated for each individual, providing a comprehensive overview of their performance metrics
Output (Group level)	Outcome: Pass/fail ratios for each group are reported. The proficiency distribution of each group is not reported. Reporting: Group reports detail collective proficiency and non-cognitive assessment outcomes	Outcome: Reports reflect the distribution of proficiency across bands and overall group proficiency. Reporting: Group reports summarise proficiency and non-cognitive outcomes
Test duration	About 12 minutes for each domain	About 20 minutes for the first domain. Plus 12 minutes for each subsequent domain
Core Background Questionnaire	Demographic variables (e.g., age, gender, language, employment status). Up to 5 minutes	Demographic variables (e.g., age, gender, language, employment status). Up to 5 minutes
Optional modules	BQ Extension, including skills used at work and in everyday life (20 min). Social and Emotional Skills, Subjective Well-Being and Health, Financial Literacy, Green Skills, etc. can also be an option in future	BQ Extension, including skills used at work and in everyday life (20 min). Social and Emotional Skills, Subjective Well-Being and Health, Financial Literacy, Green Skills, etc. can also be an option in future

4.2.2. Test form design

Table 6 outlines the test form design for the Certification version of the new E&S Online assessment. Each domain is assessed with about ten items. The current instance illustrates the case for two domains, demonstrating how the design is targeted to ensure accurate and dependable measurement at critical proficiency thresholds. The design of the test forms is strategic, aiming to maximise the reliability precisely at the proficiency level thresholds. Consequently, the number of test forms corresponds to the number of these thresholds, if not more, ensuring a targeted approach for different proficiency levels. Test forms are specifically chosen based on the desired pass/fail classification level. This meticulous assembly of items within each test form means an adaptive or branching methodology is unnecessary for this assessment tool, as the items are already optimised to distinguish effectively around the threshold levels.

Table 6 Test form design of Certification version of new E&S Online (Prototype)

	Literacy	Numeracy	Total
Test form patterns	One or more test forms for each target threshold		
Number of stages	1	1	2
Adaptive / Branching approach	Not employed		
Number of items (on average)	10	10	20
Number of units (on average)	8	9	17
Testing duration	About 12 minutes	About 12 minutes	About 25 minutes

Additionally, the Certification version of the E&S Online assessment is not designed to offer estimates of group-level proficiency distribution. This is because each test form is concentrated on assessing a specific proficiency level and is not equipped to deliver accurate proficiency estimates beyond its targeted range. More critically, there is a trade-off in the version's content coverage, which inherently restricts the scope of the group-level point estimate. This limitation signifies a potential compromise in the content validity of the assessment tool, as it may not fully represent the breadth of the domain it aims to measure.

Table 7 presents the test form design for the Distribution version of the new E&S Online assessment. The initial domain is evaluated using approximately 16 items, with an expected completion time of 20 minutes. Subsequent domains are assessed with around 10 items estimated to be completed in 12 minutes.

Table 7 Test form design of Distribution version of new E&S Online (Prototype)

	Literacy	Numeracy	Total
Test forms	4 or more	4 or more	4 or more
Number of stages	1 or 2	1 or 2	2-4
Adaptive / Branching approach	Multi-Stage Adaptive Testing (MSAT)		
Number of items (1^{st} / 2^{nd})	16 / 10	16 / 10	26
Number of units (1^{st} / 2^{nd})	7 / 4	6 / 4	10-12
Testing duration (1^{st} / 2^{nd})	20 / 12 minutes	20 / 12 minutes	About 32 minutes

In contrast to the Certification version, the Distribution version requires high reliability across all proficiency levels, particularly for the first stage in the test form. Subsequent stages can be adapted based on the proficiency levels estimated from the initial stage, utilising the MSAT approach. Given the strong correlation between Numeracy and Literacy proficiencies (OECD, 2017[3]), using the proficiency estimate from one domain as a provisional level for another is a practical approach. This makes the assessment of second and subsequent domains concise, requiring fewer items than the first.

Furthermore, the Distribution version is structured to produce reliable estimates of group-level proficiency distribution by covering a broad spectrum of proficiency levels. Notably, a rotational test form design is employed to guarantee extensive content coverage at the group level. This makes the Distribution version a more comprehensive tool, capable of providing individual-level band scores, complemented by the possibility of estimating the entire proficiency distribution (mean, variance and percentiles) for the group.

4.2.3. Psychometric property

Figure 3 graphically represents the standard error functions defined in Equation 12 and the 95% CIs for prototype test forms of the Certification and Distribution versions of the new E&S Online assessment in Literacy. The items of both versions were selected provisionally from the existing PIAAC item bank. In order to scale the scores of the new tools on the PIAAC scale, the item parameters of the PIAAC were used. The Certification version's curve is depicted in red, highlighting its intended objective: to ascertain whether a respondent has attained Level 1 competency. As indicated by Figure 3, the Certification version's standard error in Literacy minimises significantly at the critical threshold between Below Level 1 and Level 1. This illustrates that, despite the test duration being only a third of the original E&S Online, the precision in determining proficiency at this threshold is comparable, if not improved (Figure 1). However, it is important to note that for proficiency Level 2 or higher, the standard error of Literacy in the Certification version increases substantially, indicating less precision in measuring higher proficiency levels than in the orignal version of the assessment. As noted earlier, the prototype for the Certification version establishes a hypothetical target at the threshold between Below Level 1 and Level 1, merely for illustrative purposes.

The standard error function for the Distribution version prototype in Literacy, represented in blue, shows a spread predominantly across the lower proficiency levels, from Level 1 to Level 3. Within these levels, the standard error values oscillate between 15 and 25. In contrast to the Certification version, this version's function spans a broader spectrum, encompassing Levels 1 to 3 comprehensively but not extending to Levels 4 and 5. The 95% CIs for Levels 1 through 3 are observed to fall within a range of 60 to 100 points. This range indicates that the confidence bands maintain a maximum span of three levels at most proficiency levels.

In light of the Distribution version's reliability, there is an opportunity to refine the reporting of proficiency by dividing each level into narrower bands. These would include distinctions such as Lower Level 1, Upper Level 1, Lower Level 2, Upper Level 2, and so forth, which could allow for a more nuanced interpretation of proficiency within the assessed range.

Figure 4 illustrates the standard error functions and the 95% CIs for the prototype test forms of Numeracy within the Certification and Distribution versions of the new assessment, distinguished by red and blue colours, respectively. The red curve of the Certification version focuses on the threshold between Below Level 1 and Level 1 as an example, underscoring the flexibility in setting target proficiency levels according to specific needs. In comparison with the original version of the assessment, the standard error (Figure 2) around the chosen target

level is impressively low considering the number of items a respondent takes, emphasising the significance of aligning test form design with assessment objectives.

The Distribution version's standard error function, depicted by the blue curve, thoroughly encompasses Levels 1 to 3, illustrating its extensive range. The precision of this version is accentuated by its 95% CIs for these levels, which are tightly bound within a 50- to 80-point range. This precision parallels that observed in the Literacy component and indicates that the proficiency bands in the Distribution version could be more finely segmented than those defined in the original PIAAC levels (i.e., six categories).

Figure 3 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the Literacy domain of the Certification and Distribution versions

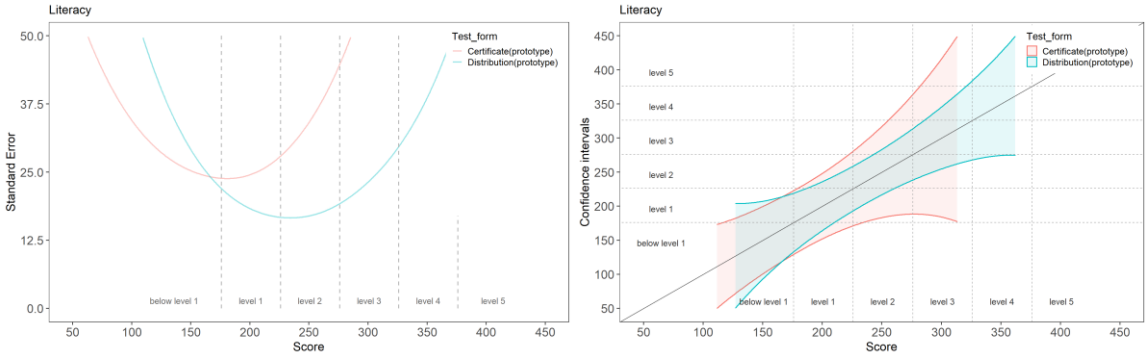
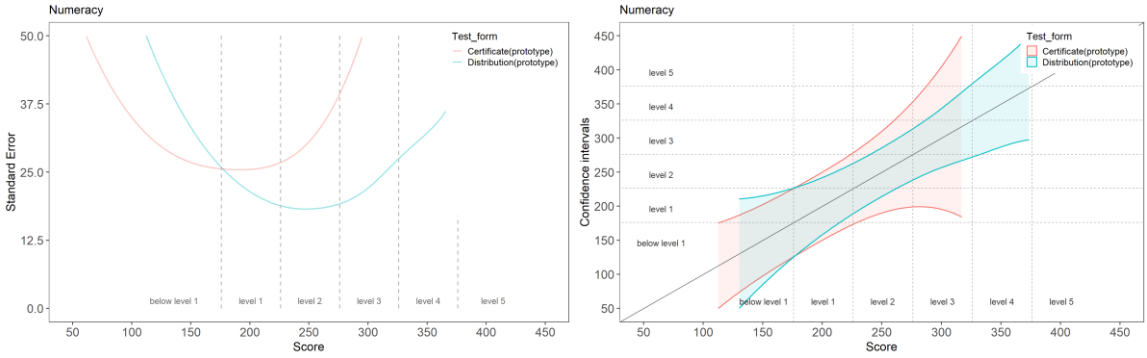


Figure 4 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the Numeracy domain of the Certification and Distribution versions



4.3. Summary

Subsection 4.1 delved into the test form design and the psychometric characteristics of the original E&S Online assessment. This original version requires approximately an hour to assess key domains: Literacy and Numeracy. It generates proficiency scores in 10-point increments on the PIAAC international scale, estimated from the responses to both cognitive items and background questionnaires. Given that the standard errors for the proficiency levels vary from 12 to 25 (Figure 1 and Figure 2) in Levels 1, 2, and 3, depending on both the test form used and the estimated proficiency levels, the 95% CIs can stretch from around 50 to 100 points among these levels. This variance suggests that there may be excessive uncertainty around the reported score for certain proficiency ranges.

Subsection 4.2 introduced two new prototype models for the E&S Online assessment: the Certification and the Distribution versions. The Certification version is tailored to provide a binary pass/fail result for each individual within a specific level after a concise 12-minute test per domain. Typically, it is anticipated that a respondent would complete assessments in two domains in addition to core background questionnaires, with the entire process taking roughly 30 minutes. In contrast, the Distribution version is designed to estimate a more detailed band score and delivers a richer profile of an individual's proficiencies. As a result, this version involves a longer testing duration, requiring 20 minutes for the first domain, while additional domains take about 12 minutes each. This version allows users to estimate a group's proficiency distribution of each domain in addition to the individual-level band score.

The graphic representations in Figure 3 and Figure 4 serve as solid evidence that the designed Certification and Distribution versions maintain their respective assessment integrity despite the reduced test durations. This demonstrates that the assessments are robust enough to sustain their diagnostic capabilities if the test is designed properly. These assessments' reliability and efficiency can be enhanced by developing a more diverse and extensive item bank. A larger pool would provide a richer database from which items can be drawn, allowing for a finer calibration of test difficulty to the abilities of the respondents. Moreover, an adaptive testing approach holds promise for the Distribution version. Adaptive testing tailoring the difficulty of items to the respondent's demonstrated abilities in real-time can deliver an efficient and individualised assessment that maintains or even enhances precision.

However, the brevity of these assessments, a feature of their design, does bring about challenges, particularly in content coverage. The reduced number of items means that each assessment might only provide a snapshot of proficiency, which might not capture the full spectrum of a respondent's abilities or knowledge. This limitation is intrinsic to the format and is a trade-off for the gains in efficiency and convenience. Developing a balanced assessment that provides both a comprehensive evaluation of abilities and efficient administration is a complex task and an ongoing challenge in the current assessment schemes.

5. Data illustration: PISA Household Survey Module

5.1. PISA for Development

5.1.1. Outline

As PISA expanded its reach, the need for an evolved assessment model became evident to address the diverse educational challenges of an increasing number of participating middle-income and low-income countries. In 2013, the OECD and several partners initiated the PISA for Development (PISA-D) to adapt the PISA survey tools for these varied contexts, aiming to aid policymaking in these countries/economies.

PISA-D's goal is to enhance the assessment capacity of participating nations, facilitating the execution of extensive learning evaluations and the interpretation and application of their outcomes to inform policy and decision-making. It employs modified PISA instruments, adjusted for relevance in middle- and low-income countries, while ensuring the results remain comparable to the main PISA scores. The initiative measures student competencies in reading (READ), mathematics (MATH), and science (SCIE) and includes questionnaires that explore student backgrounds, educational settings, and pedagogical approaches.

In 2017, the PISA-D assessment was administered to 34 605 students from seven countries, representing approximately 1.3 million 15-year-olds. This two-hour paper-based assessment focused on essential cognitive skills, with questions mainly drawn from previous PISA tests.

Besides the cognitive assessment, students and school personnel completed detailed questionnaires. Students spent 35 minutes providing data on their socio-educational context, while school staff and teachers dedicated 20 minutes to describe the educational system and their instructional environment.

5.1.2. Test form design

Within the PISA-D assessment, cognitive tests cover three domains: MATH, READ, and SCIE, with the latter not discussed in this paper. Each domain's items are organised into four clusters to be completed within 30 minutes. The MATH domain consists of 16 items per cluster, which include 10 to 13 units sourced from various OECD assessments, including PISA, PISA for Schools, and PIAAC. The READ domain is made up of 16 to 17 items per cluster, equivalent to 5 to 6 units, drawn from PISA, PISA for Schools, the Literacy Assessment and Monitoring Programme (LAMP) by UNESCO (UNESCO Institute for Statistics, 2009[35]), and PIAAC.

Students take two clusters from each of the two domains during the assessment, culminating in a 120-minute test comprising four 30-minute item clusters. PISA-D encompasses two main response formats. The first is multiple-choice, which is further divided into single selection, where participants choose one correct answer, and complex multiple-choice, requiring multiple true/false judgements. The second format is constructed response, involving numeric and text entries. These responses are processed either automatically or through manual coding, ensuring a comprehensive assessment of the participants' proficiencies.

Table 8 presents the PISA-D test form design and specifies the total number of items (or units) per domain, excluding SCIE. In each domain, the total item count is between 32 and 34 for the two combined clusters, leading to approximately 65 items to be completed within the 120-minute testing period.

Table 8 Test form design of PISA-D

Test form #	1st stage	2nd stage	3rd stage	4th stage	MATH (60 mins)	READ (60 mins)
TF01	R1 / 17(5)	R2 / 16(5)	S1	S2	0	33(10)
TF02	S2	S3	R2	R3 / 16(6)	0	32(11)
TF03	R3	R4 / 17(6)	S3	S4	0	33(12)
TF04	S4	S1	R4	R1	0	34(11)
TF05	S1	S2	M1 / 16(13)	M2 / 16(11)	32(24)	0
TF06	M2	M3 / 16(10)	S2	S3	32(22)	0
TF07	S3	S4	M3	M4 / 16(10)	32(20)	0
TF08	M4	M1	S4	S1	32(23)	0
TF09	M1	M2	R1	R2	32(24)	33(10)
TF10	R2	R3	M2	M3	32(21)	32(11)
TF11	M3	M4	R3	R4	32(20)	33(12)
TF12	R4	R1	M4	M1	32(23)	34(11)

5.1.1. Psychometric property

Figure 5 outlines the standard error functions (Equation 12) and 95% CIs for MATH and READ of PISA-D, illustrating four distinct curves for each. There are two patterns of the

item-cluster positions in each curve, which have the completely same standard error function. In PISA-D, the scale is set on the normal distribution, $N(500, 100)$, consistent with the PISA international scale; hence, PISA-D adopts the same nine proficiency levels defined in PISA.

As presented in Figure 5, the READ test forms cater to a broad proficiency range, from Level 1 through Level 5, with standard errors ranging from 30 to 50 points. These proficiencies and their standard errors are estimated based on about 33 cognitive items. The reliability of these READ test forms is maintained at an equivalent level across different versions.

Figure 6 shows the MATH test forms in PISA-D, which is predominantly on Levels 1a to 4, carrying standard errors approximately within the 25 to 35 points range. It is noted that the standard errors escalate to around 50 points at the lower and upper parts of the proficiency spectrum, particularly at Levels 1b and 5.

Figure 5 and Figure 6 highlight the measurement capabilities of the PISA-D test forms for the READ and MATH domains, respectively. The figures illustrate that the PISA-D test forms can measure 95% of the PISA-assessed population within a standard error margin of 50 points. This level of standard error is deemed acceptable for population-level statistics, assuming that the sample size is sufficiently large to provide reliable estimates.

Despite concerns regarding content validity, the assessment is considered to have an acceptable degree of validity. This acceptability is partly because the design involves four sets of 15-17 items distributed randomly among the students. Such a distribution of item sets helps to ensure broad coverage of the test content across different respondents, which supports the validity of the assessment in terms of its representativeness of the skills and knowledge areas being measured. The random assignment of item sets can help to mitigate any biases that might arise from a more fixed item distribution and contribute to a more equitable assessment across the diverse PISA-D population.

Figure 5 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the READ domain of the PISA-D

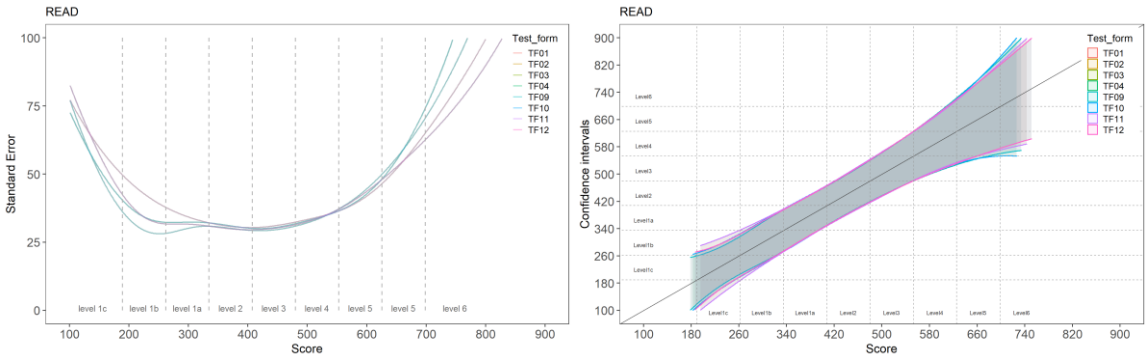
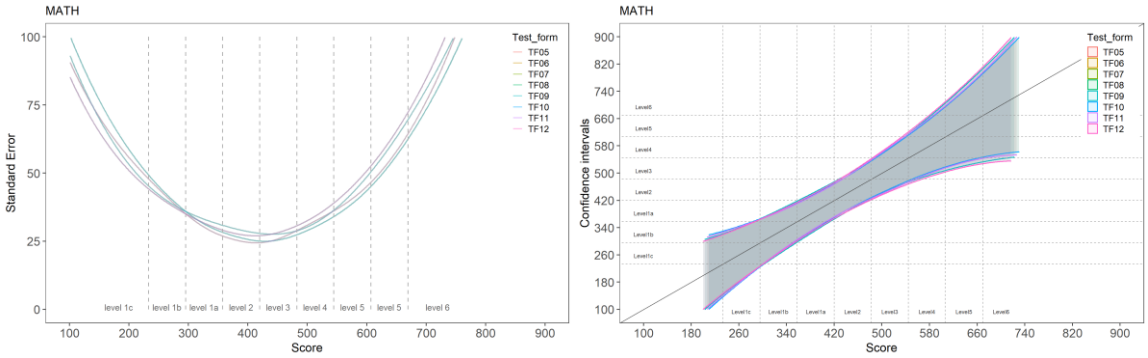


Figure 6 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the MATH domain of the PISA-D



5.2. Prototype of PISA Household Survey Module

5.2.1. Outline

The OECD has introduced the PISA household survey module (PISA-HSM), a tool specifically crafted to evaluate if 14-to-16-year-olds meet the "minimum levels of proficiency in reading and mathematics" as per the PISA standards, which align with the United Nations' Sustainable Development Goal (SDG-4) monitoring framework. The development of PISA-HSM has been meticulously adapted to fulfil the precise measurement objectives and to fit the logistical demands of conducting household assessments. It can be administered either on paper or via a tablet computer.

A purpose-driven module has been constructed utilising items from the PISA-D assessment. This module uses the standard PISA-D test as a reference point to ensure the new test's validity and reliability. Two distinct versions of the PISA-HSM are available – one with a duration of 30 minutes and an extended version of 45 minutes. Both versions contain READ and MATH domains.

The primary aim of PISA-HSM is to deliver a concise test suitable for household surveys, focusing exclusively on determining if youths have achieved the minimum proficiency levels in reading and mathematics as expected by the end of lower secondary education. This aligns with the targets and indicators set out in SDG 4.1 and 4.1.1.c. The test is specifically engineered to be as brief as feasible while still yielding robust and valid results, essential for a binary pass/fail type assessment, determining whether respondents are at or above the lower boundary of Level 2 proficiency on the PISA scale.

5.2.2. Test form design

For the PISA-HSM, the selection of items for both the 30-minute and 45-minute variants is strategically focused on optimising the accuracy of measurement at the critical juncture between Level 1a and Level 2. The goal is to ensure that the items represent all key content categories and are amenable to automatic scoring.

The PISA-HSM is structured with three stages in 30 and 45-minute formats. Each stage is designed to be completed in 10 or 15 minutes respectively. For both PISA-HSM versions, MATH is contained within a single stage, while READ spans two stages. This design decision reflects the availability of more discriminative items for MATH than READ within the PISA-D item bank. All the test forms try to avoid the risk of not completing the MATH items in the cognitive test session.

The selection of items was governed by specific criteria to maintain the focus on the pivotal proficiency levels:

1. The response formats permitted are keyword input, multiple-choice, or X-type (true/false type).
2. Items are chosen to cover all subdomains adequately.
3. The items must be highly discriminative at the targeted proficiency levels.
4. The total number of items and units should be capped to prevent the MATH domain from exceeding 10 minutes (with no more than 7 units) and the READ domain from going beyond 20 minutes (with a maximum of 14 items or 7 units).
5. Any items not meeting the above conditions can be omitted from the unit.

In the 30-minute version of the PISA-HSM, the items are meticulously arranged as indicated in Table 9. This table categorises the item clusters by stages, with the corresponding number of units presented in parentheses. To mitigate the potential effects of item positioning, a multiple test form design approach is employed in PISA-HSM. This ensures a robust assessment by accounting for any variation in responses that might arise from the order in which items are presented to the participants.

Table 9 Test form design of the 30-minute version of PISA-HSM

Test form #	1st stage	2nd stage	3rd stage	MATH (10 mins)	READ (20 mins)
TF01	M1 / 6(5)	R1 / 6(3)	R2 /5(3)	6(5)	11(6)
TF02	M1	R2	R1	6(5)	11(6)
TF03	R1	M1	R2	6(5)	11(6)
TF04	R2	M1	R1	6(5)	11(6)

Table 10 details the layout of the 45-minute PISA-HSM. Mirroring the structure of the 30-minute version, this longer assessment tool consists of three stages: one stage dedicated to MATH and two to READ. Within the MATH stage, there are eight items encompassing seven units. For READ, any combination of two stages comprises 17 items from seven units. Each of these stages is designed to be completed within 15 minutes, ensuring that the time allotted per item remains consistent with that of the 30-minute version.

Table 10 Test form design of the 45-minute version of PISA-HSM

Test form #	1st stage	2nd stage	3rd stage	MATH (15 mins)	READ (30 mins)
TF01	M2 / 8(7)	R3 / 8(4)	R4 /9(5)	8(7)	17(9)
TF02	M1	R2	R1	8(7)	17(9)
TF03	R1	M1	R2	8(7)	17(9)
TF04	R2	M1	R1	8(7)	17(9)

5.2.3. Psychometric property

Figure 7 illustrates the standard error functions and 95% CIs for READ in both the 30-minute (TF30) and the 45-minute (TF45) PISA-HSM. The graph features two curves: the red curve

corresponds to the 30-minute test, and the blue curve to the 45-minute test. For both versions, the lowest standard errors occur around the crucial proficiency threshold that separates Level 1a from Level 2.

In the 30-minute test, the standard error reaches its minimum of approximately 50 points at this threshold, suggesting that while the test is less precise overall compared to the PISA-D original test forms, it is most accurate where it matters for policy and educational interventions in the PISA-D target countries/economies. The 45-minute version shows an improved minimum standard error of about 40 points at the same threshold, indicating a higher precision level for the longer assessment.

Although the standard errors in these abbreviated tests are generally higher than those found in the full PISA-D test forms, they are closest to PISA-D levels at the proficiency level of interest. Specifically, the 30-minute test has a standard error difference of around 20 points, and the 45-minute test has a 10-point difference when benchmarked against the standard errors of the PISA-D assessments. This demonstrates that the PISA-HSM has been particularly optimised to assess proficiency around the Level 2 benchmark, which is essential for assessing minimum competency levels in the surveyed age group.

Figure 7 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the READ domain of the PISA-HSM

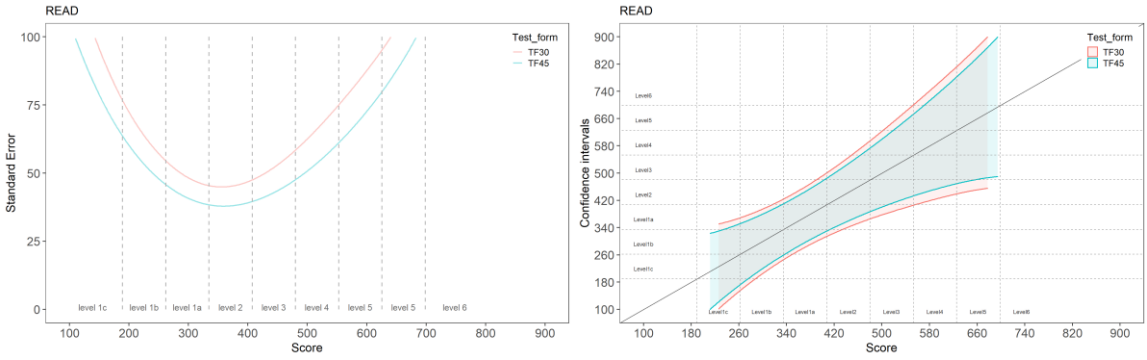
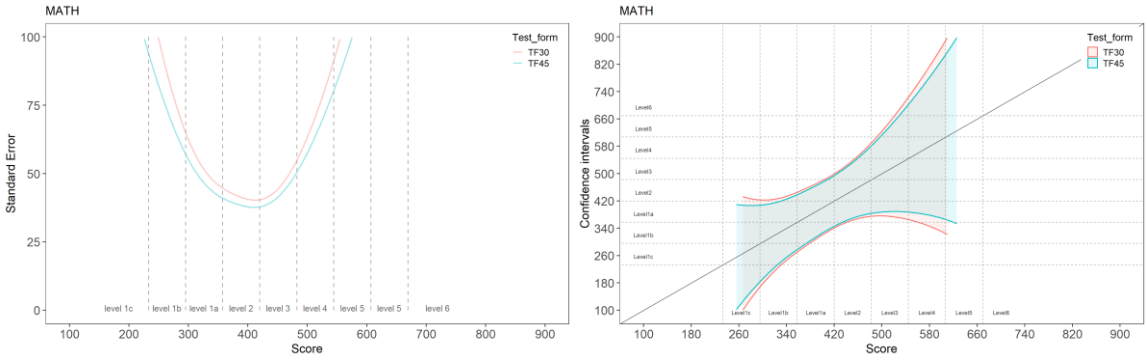


Figure 8 presents the standard error curves and 95% CIs for the MATH domain of the PISA-HSM for both the 30-minute (TF30) and 45-minute (TF45) assessments. The graph distinguishes between the two durations: a red curve for the 30-minute version and a blue curve for the 45-minute one. The most precise measurements for both tests, indicated by the lowest standard error values, are seen around Level 2 of proficiency.

For the shorter 30-minute evaluation, the precision is at its peak with a standard error of 45 points around Level 2, which, while indicative of lesser overall precision compared to more extensive assessments, suggests that the test's accuracy is most focused where it is crucial for policy insights. The longer 45-minute test improves on this with a finer minimum standard error of about 40 points at the same proficiency level.

Despite the inherently higher standard errors in these more concise PISA-HSM versions relative to the PISA-D test forms, the proximity of these errors to those of the PISA-D is narrowest at Level 2. The 30-minute test deviates by approximately 15 points in standard error and the 45-minute test by about 10 points from the PISA-D test forms, underscoring the PISA-HSM's effective calibration for pinpointing proficiency around the Level 2 benchmark.

Figure 8 Standard errors (left figure) and 95% confidence intervals (right figure) of proficiency estimates in the MATH domain of the PISA-HSM



5.3. Summary

Subsection 5.1 of the paper presents an in-depth look at the PISA for Development (PISA-D) assessment's test form design and its psychometric properties. PISA-D, which takes two hours to administer, focuses on core domains such as reading and mathematics. It is designed to generate ten plausible values (PVs) on the PISA scale for each student. The assessment's standard error functions indicate that the test forms encompass a broad proficiency range, with some scores exceeding 600 on the PISA scale. However, the use of PVs for individual diagnosis is not recommended, as discussed in Section 2 of the paper. This is due to the underlying statistical models designed for group-level inference rather than individual-level assessment. Hence, determinations regarding pass or fail outcomes should be solely based on a student's responses to cognitive items, without factoring in prior distribution estimated from background questionnaire data.

Subsection 5.2, the paper introduces two prototypes for PISA-HSM: a shorter 30-minute version and a more detailed 45-minute version. Both are developed to swiftly provide a binary pass/fail classification. While the 30-minute version offers a rapid assessment, the 45-minute version yields a more reliable judgement of an individual's proficiency diagnosis. The longer format requires 15 minutes for MATH and 30 minutes for READ. Importantly, the 45-minute version not only facilitates individual pass/fail outcomes but also allows for the estimation of a group's proficiency distribution within each domain. This feature adds a layer of utility for users interested in broader educational insights at the group level.

The graphical data presented in Figure 7 and Figure 8 affirm the effectiveness of the PISA-HSM in its 30-minute and 45-minute formats, which, despite the brevity of their design, uphold the integrity of the assessment. These figures underscore the importance of thoughtful test construction: even with shortened durations, the assessments can yield precise diagnostics when items are carefully chosen, and the structure of the test is meticulously planned. While primarily aimed at lower and middle-income countries, the PISA-D initiative encompasses a range of competencies extending into the upper echelons of student performance. This broad coverage ensures that the assessment remains relevant even for higher-achieving students within these regions.

The PISA-HSM adaptation strategically departs from this approach by focusing exclusively on the crucial juncture between Level 1a and Level 2 proficiency. By concentrating on this critical threshold, the PISA-HSM tailors its assessment to discern whether students meet a defined minimum proficiency standard, essential for tracking progress toward educational goals like those outlined in the SDGs. The exclusion of items targeting higher performance levels in the PISA-HSM does not significantly impair the assessment's precision of PISA-

HSM; it simply recalibrates the focus to align with the intended measurement objectives of determining minimum proficiency levels in a household survey context. This ensures that the assessment is more streamlined and retains the necessary accuracy where it is most needed.

6. Discussions

6.1. Towards more diverse assessments

Subsection 2.1 provided a detailed exploration of the varying outputs of ILSAs across three specific target layers: population, group, and individual levels. Each layer targets a different aspect of proficiency measurement, necessitating distinct approaches in data interpretation and analysis. Building on this foundation, Subsection 2.2 introduced the scoring methodologies applicable to these outputs, complete with an in-depth look at their mathematical properties. Section 2 is crucial as it outlines the methods used to calculate scores at different assessment levels and delves into the underlying mathematical properties of the models. Such an understanding is vital for accurately interpreting the results of ILSAs and ensuring that the assessments are both reliable and relevant to the specific needs of different target group.

At a population level, the focus is on estimating proficiency across a country/economy. The key priorities here are validity, reliability, and international comparability. Therefore, representative sampling is employed to estimate statistically representative outcomes at the population level. Furthermore, the rotational test form design is employed to ensure the content validity of the statistics over proficiencies. As an output, PVs are generated for each respondent for all domains regardless of the participation in the respective domains. These PVs provide unbiased population-level statistics and are suited for secondary analysis. However, the aggregated PVs for a group of respondents do not ensure unbiased statistics over proficiencies because the residuals of group g, $\boldsymbol{d}_g$, may not follow $N(\mathbf{0}, \boldsymbol{\Sigma})$. Aggregating PVs of the subpopulations, such as gender, likely shows unbiased statistics; however, the groups with a limited number of individuals, such as schools, may not.

Therefore, for group-level outputs, the parameters with regard to the proficiency distribution of the group should be estimated directly based on the likelihood function instead of aggregating the PVs of the group. Reliability and validity are contingent on a group of individuals answering a sufficient range of items that comprehensively cover the measure's content. Therefore, enough items need to be taken by a sufficient number of respondents to ensure both reliability and validity at the group level. The minimum number of items, as well as the minimum number of respondents, depends on the width of the concept to be measured and the precision of the measure expected to have.

In individual-focused assessments, the goal is to diagnose proficiencies in specific cognitive domains. As mentioned above, using PVs for diagnosis is inappropriate. Practical limitations, such as restricted testing time, often make precise estimations challenging. Consequently, these assessments typically use band scores or pass/fail classifications instead of point estimates due to the relatively large standard errors of these estimates when compared to the population proficiency distributions. Moreover, Subsection 2.2 proposed a new methodology to obtain an unbiased proportion of the pass/fail classification at a group level.

Consequently, the assessment outputs must be carefully defined based on their intended target level to ensure accuracy and prevent any potential bias or misinterpretation. Furthermore, it underscores the importance of adopting a scoring approach that is specifically tailored to the objectives and requirements of each target level. An inappropriate scoring

method leads to biased results, making it imperative to customise the scoring to align with each assessment's specific goals and contexts. This customisation ensures that the results are both meaningful and relevant to the intended audience, thereby enhancing the overall effectiveness and usefulness of the assessment.

Subsection 2.3 discussed the optimal test form designs for each scoring approach, PVs, group-level proficiency distribution, band scores, and pass/fail classification, introduced in Subsection 2.2. For the production of PVs, ILSAs typically employ a flexible rotational test form design, particularly effective when there is a large pool of respondents. This design strategy allows for a broad range of data collection across various sets of items and domains, ensuring comprehensive coverage of the constructs. In this setup, it is not necessary for all respondents to take every domain; they can focus on a selected few. However, having enough respondents per domain combination is crucial to determine accurate correlation coefficients. Thus, including more domains requires a larger number of respondents. Test forms generally centre around one or two domains with 20 or more items each, ensuring reliable estimation of proficiency. Furthermore, comprehensive background questionnaires with sufficient variables are vital for generating covariates for the LRM, a critical step in producing reliable PVs.

The test form design remains as flexible as for PVs for group-level proficiency distribution. However, the required number of respondents is typically lower than for PVs, as this approach does not employ the LRM. Similar to PVs, the design allows for the inclusion of numerous domains as long as each has a sufficient sample size. The goal is to estimate the group proficiency distribution directly within the modelling process, not generate individual scores. The number of items in a domain should be enough to form an unimodal likelihood function at the individual level, although extensive content coverage is not a priority.

In contrast, when determining band scores or pass/fail classifications, there is a heightened emphasis on reliability and validity at the individual level. Achieving this balance can be challenging due to constraints on test length. To address this, every participant is assessed across all domains, but the number of items included in each domain is carefully determined to maintain the test's feasibility and ensure the precision of scores. Adaptive testing methods can be beneficial here for tailoring domain-specific tests. For the pass/fail classification method, the focus is on a specific proficiency level, making it unnecessary to employ an adaptive item selection strategy. The aim is to maximise measurement reliability around the pass/fail threshold, typically requiring respondents to cover all domains with a minimum number of items.

Consequently, these test form designs are carefully crafted to meet the unique requirements of each assessment type, balancing the need for comprehensive content coverage, respondent engagement, and reliability in measurement. Each design caters to the specific objectives of the assessments, whether at the individual, group, or population level, ensuring accuracy and efficiency in scoring and interpretation.

Section 2 emphasises the importance of expanding the item bank in assessment management. This expansion is identified as one of the top priority tasks, essential for delivering assessments that are not only reliable and valid but also efficient. The enlargement of the item bank directly contributes to the enhancement of the overall quality and effectiveness of the assessments. Additionally, the section highlights the necessity of implementing a flexible test form design, which is integral to the successful integration of assessments into a CBT system. The shift towards CBT offers significant advantages. It allows for greater flexibility in various aspects of assessment management, including item development, test form assembly, and test administration.

The use of CBT systems facilitates a more dynamic approach to managing assessments. It provides the ability to quickly adapt and update assessment materials, tailor test forms to

specific needs, and efficiently administer tests. This flexibility is particularly beneficial in responding to evolving educational standards and diverse testing requirements. By embracing CBT, assessment processes become more streamlined, responsive, and tailored to the current educational landscape, ensuring they remain relevant and effective.

6.2. Towards more flexible assessments

Section 3 reviewed technical procedures for enhancing the flexibility of large-scale assessment management. Subsection 3.1 focused on item bank development and item banking processes. Subsection 3.2 compared different test management strategies from a testing cycle viewpoint, specifically highlighted item validation and parameter estimation processes. Subsection 3.3 detailed the technical standards for item validation and parameter estimation, ensuring the validity and reliability of the assessments. Subsections 3.4 and 3.5 introduced in-test item trialling and item cloning techniques, instrumental in adding flexibility and enriching the item bank for these assessments.

Periodic assessments occur at regular, predefined intervals and follow a systematic item bank management process. All participating entities in periodic assessments adhere to a standardised timeline, keeping comparability as high as possible at a certain time. In-test trialling can be employed in periodic assessments; however, the test forms must be designed carefully so as not to lose comparability with the previous cycles. On the other hand, sporadic assessments are more flexible and are conducted as per specific requirements. These assessments allow participating entities more autonomy in scheduling, offering varied administration options. Furthermore, it is easier to implement newly developed items into main studies.

In most ILSAs, the assessments are designed as a periodic assessment, that usually focus on the population-level statistics and aims for strict comparison with regard to proficiency distributions of the participating entities at a certain time. In contrast, sporadic assessments are more adaptable, often targeting group and individual levels to meet specific user needs. The instrument development and item validation and parameter estimation phases differ significantly between these two types. In periodic assessments, these phases are synchronised across entities for the main study, with new items trialled for psychometric properties before selection. In sporadic assessments, instrument development and item validation and parameter estimation are more independent, with items being drafted, revised, translated, and trialled separately. Often, sporadic assessments combine field trials with the main study in a process.

Considering the costs and administration burdens, in-test trialling is recommended over independent field trials unless the intervals of the assessments are periodic and very long. In-test trialling gives frequent opportunity to revise and validate developed items, which makes item bank management efficient. In order to manage in-test trialling successfully, it is important to develop a test management system that allows designing and implementing items and test forms to be flexible to control item exposure properly.

The technical standard for item validation and parameter estimation in ILSAs plays a crucial role in the scaling procedure, primarily focusing on the validation and parameter estimation of assessment items. This process is complex and involves several critical steps. It begins with item validation and parameter estimation, where the expected ICRF is compared against the pseudo-observed frequency for each participating country/economy. Items that are new or do not align with expected functioning undergo parameter estimation. The methodology employs a multi-group model with a partial invariance assumption, estimating item parameters using the MMLE-EM.

DIF is a critical aspect of this process, assessed through RMSD, which measures the discrepancy between the model and the data. This assessment ensures that parameters, especially international ones, are estimated without bias from countries/economies that exhibit DIF. As the diversity of the participating countries/economies increases, the accuracy of identifying DIF becomes increasingly essential. To effectively detect DIF, data collection must encompass a broad range of countries/economies, particularly those representing different language groups.

The test forms for item validation and parameter estimation should include a substantial number of trend items with appropriate psychometric properties, ensuring the computation of a valid and reliable conditional distribution of the missing data (i.e., θ in Equation 5). When estimating new parameters on the original scale in assessments, the ratio of new items to trend items is of secondary importance compared to ensuring an adequate quantity of trend items. Essentially, the key factor is having enough trend items in a test form, regardless of the number of new items it contains. It is essential to have enough respondents for each item, typically set at a minimum of 200-500 from a wide range of proficiency levels, depending on the reliability of the measure. Data collection for a trial item continues until the standard errors of its estimates are within acceptable ranges, and the pseudo-observed frequency profile shows stability.

Overall, ILSAs in-test trialling provides an efficient means of item validation and parameter estimation and validation within the main study, eliminating the need for separate field trials. This method requires careful management of items, adequate response collection, and meticulous data collection across countries/economies to ensure the high-quality and comparability of assessment outcomes.

Subsection 3.5 delved into a strategic approach for sustainably expanding an item bank, a method particularly relevant in the context of item development for ILSAs. Item cloning offers several benefits. It allows for a more efficient and quicker item generation process than developing completely new items. It maintains strong construct validity, as the basic structure of the original item is preserved. Item cloning also prevents test-takers from memorising items by introducing similar yet varied items, thus enhancing item security. Economically, it is more cost-effective, utilising already validated items and reducing the need for additional validation resources.

From an analytical perspective, creating multiple versions of an original item is invaluable. It not only facilitates a more flexible assessment but also offers insights into the characteristics of the item by analysing variations in the ICRFs of the cloned items. Hence, item cloning is a technique for enlarging an item bank and a means to gain a deeper understanding of item structures.

The necessity for an efficient item management tool should be noted. Such a tool is crucial to ensure that respondents do not encounter multiple cloned items in one test session, which is important for maintaining extensive content coverage and upholding the principle of local independence assumption of the measurement. This tool is integral in managing the assembly of test forms, ensuring a strategic distribution of cloned items.

In summary, item cloning is a pivotal strategy for efficiently expanding an item bank, offering advantages in terms of development speed, cost-effectiveness, and construct validity, while also enriching the assessment content. The effectiveness of this strategy relies heavily on the use of a robust item management tool to guarantee the appropriate distribution and evaluation of cloned items.

6.3. New tools and their limitations

Sections 4 and 5 outlined the development of new assessment prototypes that are derivatives of PIAAC and PISA-D, respectively. These prototypes are designed with the assumption of sporadic assessment, primarily targeting group- and individual-level diagnoses. Both sections represent efforts to create new tools for assessment, leveraging the framework and methodologies of existing ILSAs but redirecting their focus towards specific objectives. This approach signifies an evolution in the use of ILSAs, adapting and refining them to meet particular diagnostic purposes and address various educational stakeholders' diverse needs.

Subsection 4.1 examined the test form design and psychometric properties of the original E&S Online. This version, taking about an hour, assesses key domains like Literacy and Numeracy and generates proficiency scores in 10-point increments on the PIAAC proficiency scale. The test forms of the original E&S Online are appropriately designed. However, dependency on the LRM estimates based on PIAAC Cycle 1 data to obtain prior distributions limits its use to only countries involved in that cycle. The new version proposes moving to a likelihood approach instead of the previous population modelling, expanding its applicability beyond countries that participated in Cycle 1 and group and individual levels.

Subsection 4.2 introduced two new E&S Online versions: the Certification version, offering a binary pass/fail result after a 12-minute test per domain, and the Distribution version, estimating band scores with longer testing periods (20 minutes for the first domain and 12 minutes for additional ones). The standard error functions of the new tools demonstrate these versions' assessment reliability, even with shorter test durations.

A larger, more diverse item bank would enhance these assessments' reliability and efficiency. Adaptive testing, especially for the Distribution version of the new E&S Online, will improve assessment efficiency and precision. However, the brevity of these tests, a key feature of their design, poses challenges in content coverage. Reducing the number of items in an assessment could potentially compromise its content validity. Balancing comprehensive assessment and efficient administration remains a complex and ongoing challenge in these assessment schemes.

Section 5 introduced two prototypes for PISA-HSM, designed for pass/fail classification: a concise 30-minute version and a more comprehensive 45-minute version. The 45-minute version, allocating 15 minutes for MATH and 30 minutes for READ, offers a more detailed individual proficiency diagnosis and enables estimation of group proficiency distribution. The standard error functions of the new tools validate the effectiveness of both versions, demonstrating that, despite their shortened formats, they maintain the reliability of the assessment around the target levels. This effectiveness is attributed to careful test construction and item selection.

The PISA-D initiative caters to a broad range of competencies, making it relevant for students in lower and middle-income countries. In contrast, the PISA-HSM specifically focuses on the crucial threshold between Level 1a and Level 2 proficiency. This narrowed focus is strategic, concentrating on whether students meet a defined minimum proficiency standard, essential for tracking educational goals like those in the SDGs. By excluding higher-level items, PISA-HSM remains precise in its targeted measurement objectives within a household survey context, ensuring a streamlined and accurate assessment of minimum proficiency levels.

6.4. Future development

Sections 4 and 5 demonstrated how existing ILSAs can be adapted for both group and individual assessments and how these assessments can be flexible in terms of assessment

design. The sections proposed that these new assessments, tailored for groups and individuals, should be available intermittently and on demand. From a practical point of view, a critical factor for these assessment tools is the immediacy of reporting results. Individual-level band scores and pass/fail determinations can be quickly given once scoring for open-ended items and proficiency estimation of a respondent are completed. In the context of reporting, the major delay in producing these results lies in the scoring.

Automated scoring, therefore, emerges as a key element for rapid diagnosis. Okubo, et al. (2023[36]) demonstrated that AI scoring could effectively evaluate most PISA-type open-ended items, with the quality of this automated scoring closely mirroring that of human scoring. The effectiveness of AI scoring in a multilingual assessment context is noteworthy, particularly as it leverages a multilingual large-scale language model. Given the primary goals of group and individual targeted assessments, providing test-takers with immediate feedback is beneficial and important, even if the scoring quality slightly differs from human scoring. Therefore, the adoption of automated scoring systems is a pivotal advancement in large-scale assessments, particularly for those conducted sporadically, like the illustrated assessments in Sections 4 and 5.

The significance of expanding item banks has been a key focus of this paper. Subsection 3.5 introduced the item cloning technique as a practical method. In addition to the item cloning technique, automatic item generation (AIG), also known as automatic question generation (AQG), is another vital method for enhancing item banks. AIG/AQG, a field of research dedicated to the automated creation of test items, varies in its approach (Embretson, 1999[37]; Brown, Frishkoff and Eskenazi, 2005[38]; Lin, Sung and Chen, 2007[39]; Susanti, Tokunaga and Nishikawa, 2020[40]). Some methods emphasise the automatic generation of multiple-choice options, while others focus on crafting the questions themselves. With the advancement of large language models (LLMs), these techniques have evolved into viable solutions for augmenting item banks. AIG/AQG goes beyond mere expansion; it paves the way for creating more interactive and innovative item formats designed to assess complex skills beyond traditional domains. The progression of AIG/AQG research is essential for the development of cutting-edge assessment techniques.

The expansion of item banks directly impacts the crucial task of test form assembly in-test management, particularly for sporadic assessments. These assessments inherently incorporate in-test trialling as part of their design structure. Consequently, it becomes imperative to assemble test forms that include both trial and trend items, adhering to the criteria outlined in Section 3. Automated test assembly (ATA) emerges as a practical and efficient solution to facilitate this process. ATA is an algorithm designed to address the complexities of linear programming (LP) in the framework of large-scale assessment, specifically tailored to the nuances of test form design (Van der Linden, 2005[41]; Fuchimoto, Minato and Ueno, 2023[42]). This algorithm not only streamlines the process of assembling test forms but also significantly enhances the efficiency of data collection during in-test trialling. Moreover, ATA plays a pivotal role in ensuring the reliability and validity of the test forms, which are key aspects that are essential for the effectiveness of any assessment tool.

Expanding the item bank and the automation of assessment management tasks will expand the scope and improve the usefulness of educational assessments for practitioners. More significantly, technological advancements in-test administration and management are set to transform the role of assessments in education. As these tools become more integrated into daily learning scenarios, providing frequent, instant, and detailed feedback to test-takers, their importance and impact on the educational process will be greatly amplified. This shift towards more immediate and comprehensive diagnostic assessments promises to make educational evaluations a more central and effective component of the learning experience.

References

Antony, J. (ed.) (2014), *Assessment Development*, Wiley-Blackwell. [29]

Bock, R. and M. Aitkin (1981), "Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm", *Psychometrika*, Vol. 46, pp. 443–459. [22]

Brown, J., G. Frishkoff and M. Eskenazi (2005), *Automatic question generation for vocabulary assessment*. [38]

Chang, H. and Z. Ying (1999), "A-stratified multistage computerized adaptive testing", *Applied Psychological Measurement*, Vol. 23, pp. 211–222. [11]

Clarke, M. (2012), "What Matters Most for Student Assessment Systems", *Systems Approach for Better Education Results (SABER) student assessment working paper*, pp. 1-56. [8]

Cresswell, J., U. Schwantner and C. Waters (2015), *A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data*, The World Bank / OECD Publishing. [9]

Dempster, A., N. Laird and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society, Series B.*, Vol. 39, pp. 1-38. [21]

Embretson, S. (1999), "Generating items during testing: Psychometric issues and models.", *Psychometrika*, Vol. 64, pp. 407-433. [37]

Firth, D. (1992), *Bias reduction, the Jeffrey's prior and GLIM*, Springer. [25]

Fuchimoto, K., S. Minato and M. Ueno (2023), "Automated parallel test forms assembly using zero-suppressed binary decision diagrams", *IEEE Access*, pp. 1-11. [42]

Hambleton, R. and H. Swaminathan (1985), *Item response theory: principles and applications*, Kluwer-Nijhoff. [23]

Hopfenbeck, T. et al. (2018), "Lessons Learned from PISA: A Systematic Review of Peer-Reviewed Articles on the Programme for International Student Assessment", *Scandinavian Journal of Educational Research*, Vol. 62, pp. 333-353. [1]

Lin, Y., L. Sung and M. Chen (2007), *An automatic multiple-choice question generation scheme for English adjective understanding*. [39]

Lord, F. (1980), *Applications of item response theory to practical testing problems*, Erlbaum. [15]

Lord, F. and M. Novick (1968), *Statistical Theories of Mental Test Scores*, Addison-Wesley. [14]

MacCallum, R. and L. Tucker (1991), "Representing sources of error in the common-factor model: Implications for theory and practice.", *Psychological Bulletin*, Vol. 109, pp. 502-511. [32]

Martin, M., M. von Davier and I. Mullis (2020), *Methods and Procedures: TIMSS 2019 Technical Report*. [4]

Masters, G. (1982), "A Rasch model for partial credit scoring", *Psychometrika*, Vol. 47, pp. 149–174. [19]

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 58, pp. 525–543. [28]

Muraki, E. (1992), "A generalized partial credit model: Application of an EM algorithm", *Applied Psychological Measurement*, Vol. 16, pp. 159-176. [16]

OECD (2021), *The Assessment Frameworks for Cycle 2 of the Programme for the International Assessment of Adult Competencies*, OECD Publishing. [34]

OECD (2017), *PISA 2015 Technical Report*, OECD Publishing. [3]

OECD (2014), *PIAAC Technical Standards*, https://www.oecd.org/skills/piaac/PIAAC-NPM(2014_06)PIAAC_Technical_Standards_and_Guidelines.pdf (accessed on 1 October 2023). [2]

OECD (2009), *PISA Data Analysis Manual: SAS, Second Edition*. [6]

OECD (2009), *PISA Data Analysis Manual: SPSS, Second Edition*. [7]

OECD (n.d.), *EDUCATION & SKILLS ONLINE ASSESSMENT*, https://www.oecd.org/skills/ESonline-assessment/abouteducationskillsonline/ (accessed on 1 October 2023). [33]

OECD (n.d.), *PISA 2018 Technical Report*, OECD publishing, https://www.oecd.org/pisa/data/pisa2018technicalreport/ (accessed on 1 October 2022). [30]

OECD (n.d.), *PISA for Development*, https://www.oecd.org/pisa/pisa-for-development/ (accessed on 1 October 2023). [13]

Okubo, T. (2022), "Theoretical considerations on scaling methodology in PISA", *OECD EDU working paper* 282, pp. 1-31. [10]

Okubo, T. et al. (2023), "AI scoring for international large-scale assessments using a deep learning model and multilingual data", *OECD Education Working Papers*, Vol. 287, pp. 1-34. [36]

Rasch, G. (1960), *Probabilistic models for some intelligence and attainment tests*, Nielsen and Lydiche. [20]

Samejima, F. (1969), "Estimation of latent ability using a response pattern of graded scores", *Psychometric Monograph*, Vol. 17, pp. 1-100. [17]

Susanti, Y., T. Tokunaga and H. Nishikawa (2020), "Integrating automatic question generation with computerised adaptive test.", *RPTEL 15*, Vol. 9. [40]

Thissen, D. and L. Steinberg (1986), "A taxonomy of item response models", *Psychometrika*, Vol. 51, pp. 567–577. [18]

UNESCO Institute for Statistics (2009), *The Literacy Assessment and Monitoring Programme (LAMP)*, https://unesdoc.unesco.org/ark:/48223/pf0000217138. [35]

van de Vijver, F. et al. (2019), "Invariance analyses in large-scale studies", *OECD Education Working Papers*, Vol. 201, pp. 1-110. [27]

Van der Linden, W. (2005), *Linear Models for Optimal Test Assembly*, Springer. [41]

van der Linden, W., B. Veldkamp and J. Carlson (2004), "Optimizing balanced incomplete block designs for educational assessments", *Applied Psychological Measurement*, Vol. 28, pp. 317–331. [26]

von Davier, M. et al. (2023), *Methods and Procedures: PIRLS 2021 Technical Report*. [5]

Waller, M. (1981), "A procedure for comparing logistic latent trait models", *Journal of Educational Measurement*, Vol. 18, pp. 119-125. [31]

Warm, T. (1989), "Weighted likelihood estimation of ability in the item response theory", *Psychometrika*, Vol. 54, pp. 427-450. [24]

Yamamoto, K., H. Shin and L. Khorramdel (2019), "Introduction of multistage adaptive testing design in PISA 2018", *OECD Education Working Papers*, Vol. 209, pp. 1-29. [12]

A. Appendix

A.1. Evaluating the effects of sample size on LRM through numerical simulation

In this section, the minimum sample size to obtain the stable latent regression modelling (LRM) estimates is investigated. In LRM, the proficiency θ is measured by the responses to the cognitive items and their parameters and is regressed on the principal components derived from background questionnaire variables. Note that the item parameters that define θ, are fixed in the LRM step. ILSAs typically gather a substantial array of background variables. In LRM, rather than using the observed variables directly, principal components that account for a significant portion of the variation in these background questionnaire variables are utilised for some mathematical reasons. In PISA, the model incorporates those components that either explain 80% of the total variance or correspond to up to 5% of the sample size (OECD, 2017[3]). This approach is adopted to prevent numerical instability that could arise from overfitting of the model.

In this simulation, the dependent variable θ specifically denotes one of the main domains, either READ, MATH, or SCIE. The simulation data is generated from an ILSA dataset, which includes 19 items for MATH, 27 items for READ, and 22 items for SCIE for each respondent on average. For the original datasets, the data of two different languages (i.e., populations) are prepared to see the difference of the residuals by the populations. During the simulation process, samples are randomly drawn from the original dataset, with sample sizes ranging from 200 to 1000, increasing in increments of 10. For each sample size condition, 150 separate datasets are generated. It is important to highlight that the number of principal components used as covariates is adjusted based on the sample size, adhering to the previously mentioned technical procedure to avoid overparameterisation and ensure model stability. LRM is then applied to each of the generated datasets. The focus of the analysis is on the residual variances of the models to see the extent to which the covariates account for variations in the proficiency variable θ.

Figure 9, Figure 10, and Figure 11 illustrate the residual standard deviations of the estimated LRM parameters across different conditions, with each language depicted in a unique colour. In these figures, the solid lines signify the mean of the residual standard deviations for each condition, while the shaded bands around these lines represent the 95% confidence intervals of these estimated means. The simulation setting for these figures uses a scale of $N(500, 100)$. These figures indicate that the residual standard deviation is typically smaller for sample sizes below 500 compared to those above 500 in all domains. The deviation increases gradually up to a sample size of around 500, beyond which it stabilises, regardless of further increases in sample size. This pattern suggests that LRM may be overfitting when the sample size is less than 500, which is universal for both populations. Therefore, in the context of ILSAs, it is advisable to have more than 500 respondents for LRM to achieve optimal results. However, it is important to note that this simulation is based on the dependent variable, θ, which is calculated from approximately 20 items or more. Consequently, for modelling in subdomains, a larger sample size might be necessary to ensure the robustness and reliability of the LRM estimates.

Figure 9 Residual standard deviations of the estimated latent regression models in MATH

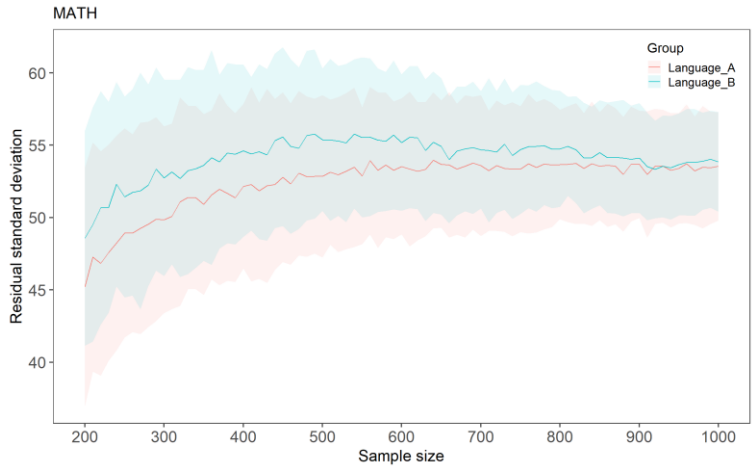


Figure 10 Residual standard deviations of the estimated latent regression models in READ

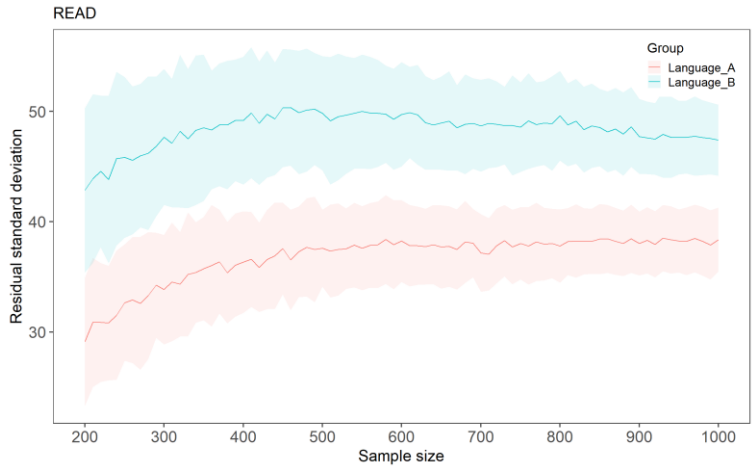


Figure 11 Residual standard deviations of the estimated latent regression models in SCIE