Ministry of Education
Government of India

सत्यमेव जयते

विद्यया ऽमृतमश्नुते

एनसीईआरटी
NCERT

PARAKH
Performance Assessment, Review, and
Analysis of Knowledge for Holistic Development

PARAKH Rashtriya Sarvekshan 2024

ANALYSIS FRAMEWORK

Shri Sanjay Kumar
Secretary

Department of School
Education & Literacy,
Ministry of Education
Government of India

MESSAGE

Education is not only a fundamental right but also the foundation of a progressive and empowered society. In this context, the PARAKH Rashtriya Sarvekshan 2024 marks a significant advancement in how we understand and improve the quality of education in India. This framework is a testament to our commitment to align educational practices and policies with the objectives outlined in the National Education Policy (NEP) 2020.

This survey transcends traditional methods by adopting cutting-edge statistical models and robust methodologies, ensuring comprehensive coverage of student learning outcomes and the factors influencing them. It evaluates performance across diverse grades, subjects, and regions, providing a true reflection of the state of school education in India. The integration of contextual data—spanning student demographics, teacher qualifications, and school infrastructure—offers a nuanced understanding of the complexities that shape educational achievements.

The outcomes of this survey are not merely statistical indicators; they are a call to action. They provide critical insights that will inform strategic decisions, enabling policymakers and educators to address learning gaps, strengthen the education ecosystem, and ensure equitable learning opportunities for all.

The PARAKH Rashtriya Sarvekshan 2024 is a pivotal step toward achieving the broader vision of educational excellence, equity, and inclusivity as articulated in the NEP 2020. It serves as a robust framework for continuous improvement and innovation in education, paving the way for a brighter and more informed future for our young learners.

Shri Anandrao V Patil
Additional Secretary

Department of School
Education & Literacy,
Ministry of Education
Government of India

MESSAGE

The PARAKH Rashtriya Sarvekshan 2024 represents an ambitious and transformative step in our approach to large-scale educational assessments in India. This initiative is a response to the need for robust, reliable, and actionable data to guide educational reforms and policies across the nation.

At its core, the PARAKH Rashtriya Sarvekshan 2024 framework underscores the principles of inclusivity and precision. Through a carefully designed assessment structure and methodology, it provides a comprehensive evaluation of students' proficiency levels while considering the contextual factors that influence learning outcomes. By adopting a balanced incomplete block design and advanced scaling techniques, the framework ensures that data is not only accurate but also reflective of the diversity inherent in our educational landscape.

This initiative also highlights the importance of an evidence-based approach to policymaking. The insights drawn from PARAKH Rashtriya Sarvekshan 2024 will serve as a critical resource for identifying disparities, allocating resources effectively, and prioritizing interventions that address the most pressing challenges in our education system. It is a framework designed to facilitate informed decisions that can reshape the future of education in India.

As we leverage the outcomes of this survey, we must remain steadfast in our vision to create an education system that is equitable, accessible, and reflective of the aspirations of every child in this country.

FOREWORD

The PARAKH Rashtriya Sarvekshan 2024, under the aegis of NCERT, embodies a progressive and scientific approach to educational assessment in India. It represents a deliberate effort to ensure that national surveys are not only aligned with the vision of the National Education Policy (NEP) 2020 but are also capable of addressing the contemporary challenges and aspirations of our education system.

This framework leverages advanced statistical models, including Item Response Theory (IRT) and latent regression techniques, to enhance the accuracy, reliability, and interpretability of assessment results. The methodological rigor ensures that findings are robust and offer a detailed understanding of student performance across grades, subjects, and regions. Furthermore, the integration of contextual data provides a multidimensional view of the factors influencing educational outcomes.

The insights generated through this framework are crucial for identifying strengths and gaps within the education system. They provide an empirical foundation for designing targeted policies and interventions aimed at improving the quality of education across diverse contexts. PARAKH Rashtriya Sarvekshan 2024 also establishes a precedent for transparency and accountability in educational assessments, ensuring that the outcomes are meaningful and actionable for a wide range of stakeholders.

This initiative is not just about measuring learning outcomes but also about fostering a culture of research and innovation in educational assessment. It reaffirms NCERT's role as a leader in shaping the future of education in India by providing tools and methodologies that are both scientifically sound and practically relevant.

Prof. Dinesh Prasad Saklani
Director
National Council of Educational Research and Training
New Delhi, India

PREFACE

The PARAKH Rashtriya Sarvekshan 2024 is a landmark initiative in the evolution of educational assessments in India. Conceptualized to align with the objectives of the National Education Policy (NEP) 2020, this framework reflects a forward-looking vision for how we evaluate and enhance learning outcomes on a national scale.

The rationale behind PARAKH Rashtriya Sarvekshan 2024 stems from the need to bridge the gap between policy aspirations and ground realities. This framework is designed to provide a comprehensive evaluation of student performance while addressing the diverse contexts in which education is delivered. By employing advanced methodologies such as Item Response Theory (IRT), balanced incomplete block designs, and plausible values, the survey ensures that its findings are both precise and reflective of the varied educational settings across India.

Unlike traditional assessments, PARAKH Rashtriya Sarvekshan 2024 places equal emphasis on contextual variables such as student demographics, teacher attributes, and school infrastructure. This multidimensional approach enables a deeper understanding of the factors influencing student achievement and provides actionable insights for systemic improvement. The development of proficiency levels, coupled with robust reporting mechanisms, ensures that the results are accessible and meaningful to a broad audience, including educators, policymakers, and researchers.

This framework also represents a significant shift toward evidence-based decision-making in education. The data and insights derived from PARAKH Rashtriya Sarvekshan 2024 will serve as a foundation for targeted interventions, resource allocation, and policy formulation, ultimately contributing to a more equitable and inclusive education system. The alignment with international best practices further underscores India's commitment to setting global benchmarks in educational assessment.

Looking ahead, the PARAKH Rashtriya Sarvekshan 2024 is not merely a technical exercise but a roadmap for transformation. It provides a framework that is not only relevant to the present but also adaptable to future educational challenges and opportunities. This document is both a guide and a call to action for all stakeholders committed to improving the state of education in our country.

Let this framework serve as a beacon of our collective aspirations and a testament to our dedication to nurturing the potential of every learner.

Prof. Indrani Bhaduri
CEO & Head, PARAKH and Head, ESD
NCERT, New Delhi

Contents

1 Introduction

PARAKH Rashtriya Sarvekshan (PARAKH RS) 2024, formerly known as the National Achievement Survey (NAS), is designed to evaluate and monitor the performance of India's school education system across the country and over time. Managed by the National Assessment Center, PARAKH (Performance Assessment, Review, and Analysis of Knowledge for Holistic Development), a constituent unit of the National Council of Educational Research and Training (NCERT), PARAKH RS, scheduled for 4th December 2024 will be uniquely tailored to align with the objectives outlined in the National Education Policy (NEP) 2020 and other pertinent global frameworks. This large-scale assessment will be conducted in a single phase across the country, ensuring nationwide participation and standardization. PARAKH RS 2024 is incorporating modifications compared to previous national surveys to lay a foundation for a new trendline of educational achievement data suited to address contemporary educational needs and goals of India.

In this document, analysis frameworks for the cognitive tests and questionnaires of the PARAKH RS 2024 are described. In PARAKH RS 2024, two types of data will be collected: 1) achievement data via cognitive tests, and 2) background and contextual data through pupil, teacher, and school questionnaires. The achievement data will be used to obtain distributions of student proficiency in the different subjects. The background and contextual data will be used to disaggregate student proficiency results into meaningful insights linked to key characteristics of students (e.g., demographics), teachers (e.g., experience), and schools (e.g., infrastructure). These insights can then help guide educational policy decisions at the national, state, and district levels.

The PARAKH RS 2024 target grades are III, VI, and IX in the following subjects: language and mathematics (for grades III, VI, and IX), the world around us (for grades III and VI), and science and social science (for grade IX). The current analysis framework is developed based on the evaluation and refinement of methodologies from the two previous assessment cycles: NAS 2017 (NCERT, 2019) and NAS 2021 (NCERT, 2022).

The analysis plan is based on the assumption that the data collection for PARAKH RS 2024 will adhere to a robust sampling methodology, ensuring a significant and representative sample suitable for the goals of PARAKH RS 2024 as outlined in the assessment framework.

2 Overview of analysis

Prior to scaling and population modelling, thorough data quality checks will be conducted to ensure alignment with test design requirements and to verify that the data reflects the intended structure and quality. The test design for PARAKH RS 2024 utilizes a balanced incomplete block (BIB) design, implemented across six test forms for grades III and VI, and eight test forms for grade IX. This test design allows for efficient content coverage while reducing student burden. Each student receives a randomly assigned test form, ensuring that variations in average test performance across forms are not attributable to differences in student proficiency levels.

Although test forms were designed to be equally difficult using field test data, differences in test form difficulty can still appear in the main study. This means that simple total-score statistics (e.g., percentage of correct responses) based on Classical Test Theory (CTT) cannot accurately and comprehensively compare student performance across forms. To overcome this limitation, Item Response Theory (IRT) modelling will be applied, enabling reliable comparisons of student performance across varied item sets by capturing the response patterns and the underlying proficiencies. This approach enables the characterization of student proficiency on a common scale, even with varied item assignments across forms. Therefore, calibrating the IRT model represents the first foundational step in analyzing the PARAKH RS 2024 data to establish *the* measurement model.

To further improve measurement accuracy and reduce potential biases when estimating relationships between proficiency and background variables, a population modelling approach, similar to that in PISA will be employed, which combines Item Response Theory (IRT) with latent regression modelling. This method adjusts for background variables from the pupil questionnaire (PQ), enhancing the precision of proficiency estimates. Ignoring such background information can lead to bias in the results, which is the reason why most large-scale assessments perform population modelling. Estimating the population model forms the second foundational step in analyzing the PARAKH RS 2024 data.

Once the population model is established, multiple plausible values will be generated for each student, drawn from a posterior proficiency distribution that accounts for uncertainty in the data. This process allows for an accurate representation of student proficiency distributions across the target populations.

The current analysis framework is organized as follows:

1. **Data Quantity and Quality Assessment:** Procedures to evaluate the collected data for PARAKH RS 2024 to confirm its adequacy and suitability for IRT scaling and population modelling.
2. **Modelling and Methodology:** Explanation and implementation of IRT models, latent regression techniques, and generation of plausible values.
3. **Development of Proficiency Levels:** Defining the discrete proficiency levels with clear skill descriptions, providing structured interpretation of the continuous proficiency scale.
4. **Analysis of Contextual Relationships:** Approach and methods used to analyze the relationships between background variables and student achievement scores.

This analysis plan ensures that the PARAKH RS 2024 data is robust, enabling accurate scaling and reliable insights into student proficiency and contextual factors.

3 Item and test analysis

Under the assumption that the data has been collected according to the design described in the PARAKH RS 2024 Assessment Framework document[1], the proposed analyses allow for:

a) Evaluating the quality of the cognitive test data collected: This includes examining test form distributions, response rates, item completion rates, and proportions of missing data.
b) Evaluating the quality of the questionnaire data collected: This includes examining response rates, item completion rates, and proportions of missing data.
c) Producing classical item analysis (IA) statistics such as item difficulty (proportion correct, *P*+), item response category distributions (distractors), item-total biserial correlations, and missing rates (omit and not-reached).
d) Evaluating the dimensionality of each measure to verify that the expected raw score inter-item correlations patterns are realized and to identify items that may not conform or may violate the assumption local independence.
e) Evaluating the extent to which item position (e.g., in the beginning or end of the test form) has an impact on item difficulty and/or missing rates.
f) Producing scaled item response theory (IRT) item parameters.
g) Producing item DIF statistics for key groups such as gender, school-management type, region, and language of the assessment.

These analyses are designed to offer multiple sources of evidence to identify potential issues in the data collection process that could impact the quality of the test items. The goal is to ensure that the items used for calculating student scores optimize construct coverage and measurement accuracy.

3.1 Data cleaning and management

The first step in the analysis of the data will be to evaluate the quality of the data and to apply appropriate data treatment(s) as needed. The evaluation procedures for the cognitive assessment data and the questionnaire data are described separately. For both, it is expected that the data collected will be provided in the form of excel or csv (comma-separated values) files.

[1] https://ncert.nic.in/parakh/pdf/Assessment_Framework.pdf

For the assessment data, it is assumed that the data follows the booklet design and that each item can be matched with the booklet(s) in which it should appear. Similar to the data received in the pilot, it is expected that the main survey cognitive test data will include the following information:

- Student ID
- Test form ID (booklet)
- School ID
- Grouping variables, including
    - Gender
    - Geographical data (region, state, and district)
    - School-management type (central government, state government, private, and government-aided)
    - Area (rural/urban)
    - Social group (SC, ST, OBC, general)
    - Language of instruction
    - Language of assessment
- Item raw responses such as A, B, C, D for the response options

Once the files are available, the amount and quality of the data collected is to be evaluated overall and by relevant grouping variables. First, it will be evaluated whether the data were collected properly and according to the booklet design. This will be done at various levels (school, state, language of assessment, and national) for all booklets, subjects, and grades to verify the quality of the collected data. For example, duplicate student IDs will be checked as well as the distributions of test forms. Furthermore, data quality checks will include reviewing the distributions of valid item responses, invalid item responses (e.g., two or more marked response options), omitted item responses (i.e., skipping an item), and not-reached items (i.e., students not being able to finish the test).

When the data pass these quality checks for all subjects and grades, scored data will be produced. In scoring the data, invalid responses (coded as 7) and omits (coded as 8) will be treated as incorrect while not-reached responses will be coded as missing (9). Coding invalid responses, omits and not-reached as 7, 8, and 9, respectively, is recommended because it allows for a clear distinction between valid raw responses (1, 2, 3, 4, if used instead of A, B, C, D) and scored responses (0, 1).

For the pupil, teacher, and school questionnaires, the conformity of the data to the codebook will be checked. Unexpected data would be reviewed, and data treatment recommended as appropriate. Furthermore, missing data will be evaluated (e.g., students not being able to finish the questionnaire). Finally, for the pupil questionnaire, it is recommended to check certain response patterns (e.g.,

straight lining) to evaluate potentially problematic response behavior (Khorramdel, von Davier, & Bertling, 2017).

3.2 Classical item analysis

Item analysis (IA) based on classical test theory (CTT), also referred to as classical item analysis, is recommended for each cognitive measure (language, mathematics, the world around us for grades III and VI, and language, mathematics, science, and social science for grade IX). IA is also be conducted for the pupil, teacher, and school questionnaires.

These item analyses aim to provide results to identify: a) the best items, b) items that have appropriate measurement characteristics, c) items that do not function appropriately but which may be revised if subject matter experts are able to identify and correct potential issues with a good chance that the item will perform well in the main study instrument, and d) items that perform poorly and cannot be fixed.

More specifically, item difficulty (proportion correct, P+), item-total biserial correlation (using block number-correct as the criterion score), item-response-category proportions, item-response-category biserial correlation, and proportions of missing responses (invalid, omitted, and not reached) will be produced. Table 1 illustrates the results to be produced.

Table 1: Example of IA results.

Item_ID	Key	Difficulty	Icor_bis	A_pct	B_pct	C_pct	D_pct	A_cor_bis	B_cor_bis	C_cor_bis	D_cor_bis	Miss_pct
item01	A	0.71	0.46	0.71	0.08	0.08	0.13	0.46	-0.36	-0.46	-0.14	0.01
item02	D	0.53	0.38	0.14	0.15	0.17	0.53	-0.04	-0.21	-0.34	0.38	0.01
item03	C	0.81	0.64	0.05	0.06	0.81	0.07	-0.48	-0.47	0.64	-0.43	0.01
item04	B	0.57	0.35	0.14	0.57	0.14	0.14	-0.16	0.35	-0.15	-0.29	0.01
item05	C	0.65	0.43	0.08	0.12	0.65	0.15	-0.40	-0.24	0.43	-0.19	0.01

IA item plots will also be produced to provide more detailed item functioning information. Such plots, as illustrated in Figure 1, along with the IA statistics illustrated in Table 1, will be particularly useful to help item developers identify potential issues with the item (e.g., wrong key specification, poor distractor, poor translation/adaptation from the source version, etc.) that may be responsible for poor measurement characteristics.

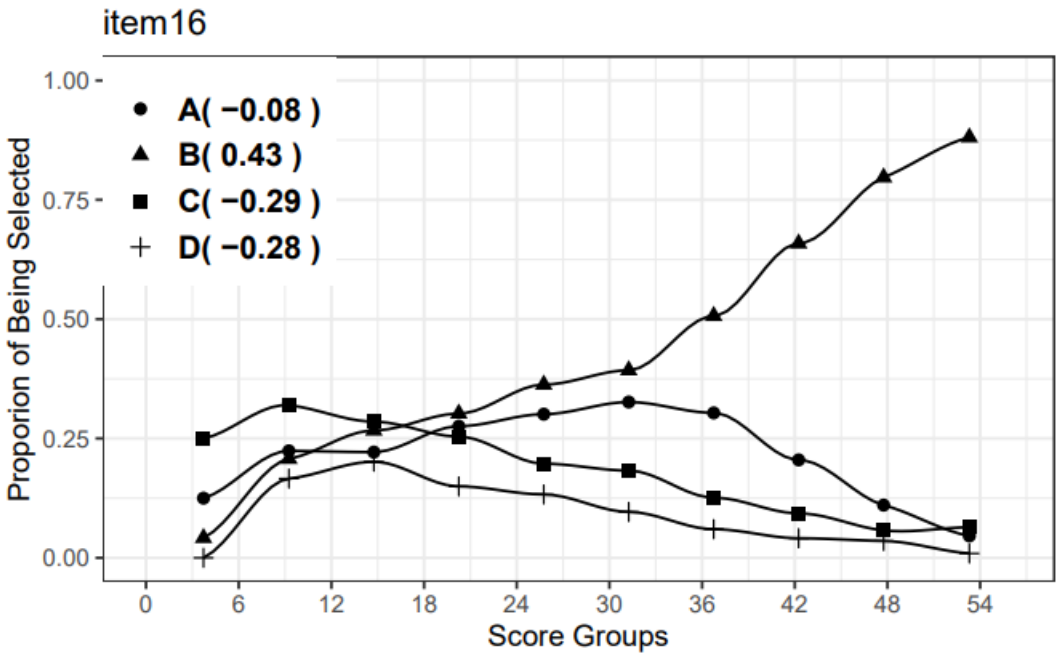

Figure 1: IA plot example

In addition to producing these analysis outcomes, it is recommended that IA criteria such as those used in PISA (OECD, PISA 2022 Technical report, Chapter 11) be applied to flag poorly functioning items (see Table 2). These flags will help the item review process in ensuring that consistent criteria are used to identify the items that pass these criteria so that they can be used in the main study. At the same time, the criteria flag the items with poor psychometric characteristics, which makes the review process more efficient by focusing on the most problematic items.

Table 2: IA flagging criteria.

IA Statistic	Criteria for flagging items
Item-total biserial correlation	0.3
Distractor-total biserial correlation	≥ item-total biserial
Minimum P+	0.2
Maximum P+	0.9
Maximum percentage omitted responses	10%
Maximum percentage not-reached responses	10%

After the analysis the using the IA flagging criteria mentioned in Table 2 the items are further classified as:

pBis range	Color coding	Recommendation on item quality
Less than 0.15	Red	Poor
Between 0.15-0.25	yellow	need review
Between 0.25-0.3	Green	Good
Greater than 0.3	Blue	Good (as PISA cutoff)

Items coded in red will be excluded from the generation of item parameters and subsequent use in student score calculations for PARAKH RS 2024, based on the cut-offs. Items coded in yellow will be included only if they pass a comprehensive review by subject-matter experts (SMEs). Items coded in blue and green will be included without requiring additional SME evaluation.

3.3 Dimensionality

Inter-item raw score correlations will be used to verify that the dimensionality of each assessment and questionnaire instrument conforms to the assessment framework's expectations. This involves examining how items are correlated with each other. Positive correlations are expected among items intended to measure the same overall construct and stronger correlations among items that belong to the same potential sub-scale may be observed. In comparison with the average inter-item correlation, particularly low or negative correlations or stronger correlation with items belonging to a different sub-scale, may indicate that certain items are not effectively measuring the targeted construct. Local item dependence will also be checked. Items that share a common stimulus, or items that may provide clue(s) to one-another may be detected when an item pair shows a noticeably higher correlation than observed amongst most of the item pairs. This information will be carefully reviewed by subject-matter experts and psychometricians to decide whether any item is identified as problematic, can be updated, needs to be removed before IRT analyses are conducted. If significant dimensionality issues are found, it is important to address them to ensure the validity of the assessment and questionnaire instruments.

Figure 1 shows an example of a 12-item correlation matrix for a questionnaire measure displayed. It shows that all the items correlate positively with each other—the expected result for a one-dimensional measure. But it also shows that the correlation between items 3 and 4 (0.74) is very high

compared to the other inter-item correlations, suggesting that these two items may not be independent.

	Item 1	Item2	Item3	Item4	Item5	Item6	Item7	Item8	Item9	Item10	Item11	Item12
Item1	1	0.39	0.55	0.54	0.34	0.27	0.33	0.26	0.4	0.3	0.34	0.3
Item2	0.39	1	0.51	0.48	0.24	0.41	0.31	0.27	0.42	0.32	0.3	0.26
Item3	0.55	0.51	1	0.74	0.35	0.33	0.52	0.39	0.56	0.45	0.41	0.37
Item4	0.54	0.48	0.74	1	0.33	0.29	0.44	0.43	0.54	0.37	0.37	0.31
Item5	0.34	0.24	0.35	0.33	1	0.4	0.53	0.43	0.31	0.28	0.22	0.17
Item6	0.27	0.41	0.33	0.29	0.4	1	0.52	0.44	0.31	0.27	0.21	0.17
Item7	0.33	0.31	0.52	0.44	0.53	0.52	1	0.6	0.37	0.35	0.26	0.24
Item8	0.26	0.27	0.39	0.43	0.43	0.44	0.6	1	0.3	0.27	0.17	0.17
Item9	0.4	0.42	0.56	0.54	0.31	0.31	0.37	0.3	1	0.44	0.43	0.33
Item10	0.3	0.32	0.45	0.37	0.28	0.27	0.35	0.27	0.44	1	0.28	0.23
Item11	0.34	0.3	0.41	0.37	0.22	0.21	0.26	0.17	0.43	0.28	1	0.47
Item12	0.3	0.26	0.37	0.31	0.17	0.17	0.24	0.17	0.33	0.23	0.47	1

Figure 2: Example of raw inter-item correlation pattern of a questionnaire

3.4 Item position effects and speededness

As indicated in the PARAKH RS 2024 Assessment Framework document, domain position effects associated with students' decreased engagement or fatigue towards the end of the test session can be expected to some extent. To mitigate such position effects on performance and on measurement errors, PARAKH RS 2024 is implementing a balanced incomplete block (BIB) design. However, while such item position effects are mitigated by a balanced design, it remains important to ensure that the decrease in performance in each domain when taken last in a test form when compared to taken first is not too large.

Possible block position effects will be examined for each domain by looking at the item difficulties as well as missing responses across the different positions in the test form. For example, in the booklet design for Grade III, language appears in the first section of the test in booklets 1, 2, and 3, in the second section in booklets 5 and 8, and in the third section in booklets 4, 6, and 7. Since each item appears in two booklets, position effects can be evaluated by comparing the item difficulties across sections. In addition, it can be differentiated whether position effects are more apparent for difficult items than for easy items. Furthermore, a similar evaluation can be conducted for missing responses to get an indication of potential fatigue effects (e.g., more omitted items towards to the end of the test) or speededness (e.g., more not-reached items in the third section of the test). These analyses on position effects and speededness for the cognitive items will contribute to the assembly of the final cognitive assessment. For example, forms can be created such that position effects and speededness are minimal.

3.5 IRT scaling

Unidimensional IRT calibrations will be carried out separately for each of the cognitive assessments (language, mathematics, the world around us for Grades III and VI, and language, mathematics, science, and social science for Grade IX). Since all items are dichotomously scored, the 2-parameter logistic model (2PLM)—the same model as used in NAS 2021. For more background on IRT, see, for example, the texts by de Ayala (2013) and Livingston (2020).

IRT calibrations can initially be conducted using the R TAM package—a well-known package that has been developed for large-scale assessment surveys such as PISA or TIMSS (Robitzsch, Kiefer & Wu, 2024). Each calibration run is to be carefully reviewed to ensure appropriate convergence of the calibration procedure and proper overall model-data fit. The item parameters and the item fit statistics produced should be reviewed and poorly functioning items identified. Additionally, a visual check of the item plots showing the empirical and model-based item characteristic curves (ICCs) should be reviewed for another check on item functioning (i.e., in addition to the classical IA). Figure 3 provides an example of the item plots to be reviewed. The blue curve is the model-based curve, and the black curve is the empirical curve

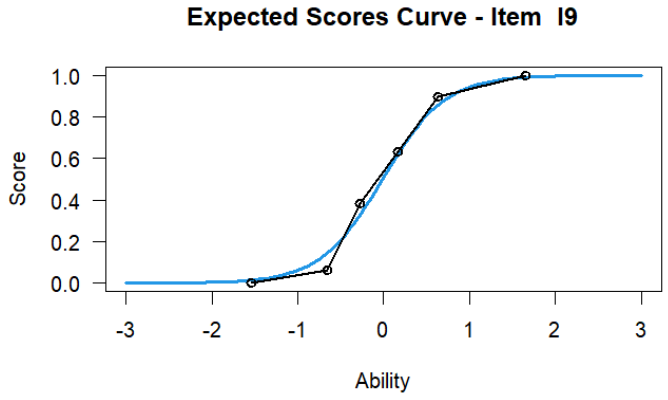

Figure 3: IRT item plot example

IRT scaled scores should also be produced to show the distributions of performance overall, and for the major subgroups of interest. Weighted likelihood estimates (WLEs) (Warm, 1989) of student ability, typically used in international large-scale surveys, can be used for that purpose. IRT reliability should also be computed to get an indication of the measurement precision for each scale (Kim, 2012).

3.6 Differential item functioning

The evaluation of differential item functioning (DIF) is a critical step in ensuring that proficiency scores are comparable across major subgroups such as gender, language, or state. Because of the relatively complex design required for large-scale assessment such as PARAKH RS, PISA, and others, it is most

efficient to estimate DIF using IRT. Thus, we propose to follow the same well-established procedure as used in PISA (Joo et al. 2021; OECD, PISA 2022 Technical report, Chapter 11; Oliveri & von Davier, 2014).

More specifically, the DIF statistics that are produced in this procedure are the mean deviation (MD) and the root mean square deviation (RMSD) statistics. These measures quantify the magnitude and direction of deviations in the observed data from the estimated ICC for each item. While the MD is most sensitive to the deviations of observed item difficulty parameters from the estimated ICC, the RMSD is sensitive to the deviations of both the observed item difficulty and item slope parameters. Both statistics are provided by the R package TAM.

Figure 4 shows a typical plot of a case for the 2PLM to illustrate how the item parameters based on the total data fit data from one particular group. In Figure 4, the solid black response curve represents the fitted (i.e., model-based) 2PLM item response curve; the other curves represent observed proportions of correct responses at various points along the proficiency scale for each group of the groups for which DIF is to be evaluated. This plot indicates that the observed proportions of correct responses, given proficiency, are quite similar for most groups. However, the data for one group indicated by the far-right yellow curve shows a noticeable departure from the model. This item is far more difficult in that particular group than the model expects, which indicates DIF.

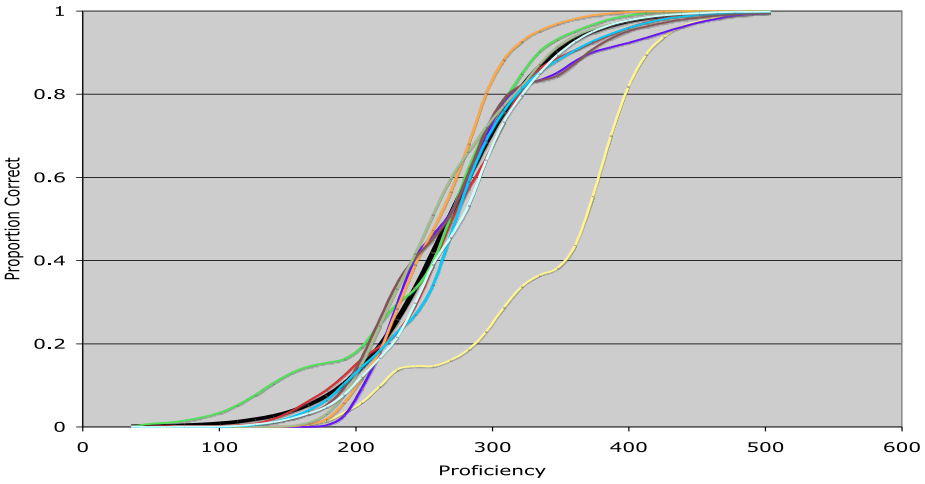

Figure 4: DIF plot

3.6.1 Analysis of linguistic equivalence

While robust linguistic quality assurance is essential, it alone cannot confirm the equivalence of different language versions of a test. Equivalence must be established through statistical analysis. Even with careful translation, there is an assumption that translated tests function independently of the

language, meaning test takers should have a similar experience regardless of the test language. However, this assumption must be tested. If statistical analysis reveals discrepancies between language versions, comparisons between language groups may not be valid.

3.7 Combined results for item review to generate scores

Although the items for PARAKH RS are selected based on pilot testing, any items that fail to meet the required standards in the main survey will be excluded and handled separately from the scoring process.

To facilitate the review of all the items and score calculation for the final item pools from which the final test forms are assembled, the results from classical IA, dimensionality analysis, item position effects, speededness, IRT scaling, and DIF analysis are to be put together in a set of tables and figures. In addition to showing each item's statistics and model parameters, criteria to identify poorly functioning items are to be produced (see Table 2). These will enable the collaboration of subject matter experts and psychometricians to quickly, yet thoroughly, evaluate the number of appropriately functioning items across competencies, subjects and classes, and to identify the items that may need to be treated differently or removed from the score calculation process.

4 Scaling methodology

PARAKH RS 2024 compared to NAS 2021 is to make use of balanced incomplete block (BIB) designs. BIB designs have a long history in combinatorial mathematics and the design of experiments (Fisher, 1935; Bose, 1939). These designs are highly efficient for the estimation of summary statistics (e.g., means). One of the first major applications of BIB designs in educational assessment was in the National Assessment of Educational Progress (NAEP; Messick, Beaton, Lord, 1983; Knapp, 1968). Since then, BIB designs have been commonly used in many large-scale survey assessments, including NAEP, PISA, PIRLS, TIMSS, and many others. Such designs are also referred to as matrix sampling designs or, in case of the assessment of multiple subjects such as in PISA, multiple matrix sampling designs (Frey et al., 2009; Rutkowski et al., 2014).

When these tests are administered, students are administered a randomly selected test form so that differences in the average test performance on forms consisting of different sets of items are not due to differences in student proficiency. However, the test forms can be of different difficulty, which means that the performance of groups measured through different sets of items cannot be directly compared using total-score statistics such as the average number or percent of items that the student responded to correctly. The limitations of using the number or percent of items correct to score assessments that are designed with BIB or administered through Multiple-Stage Adaptive Testing (MSAT) can be overcome by modelling the item responses through item response theory (IRT). When students respond to a set of items in a common subject or domain, their response patterns should show regularities that can be modelled using the underlying commonalities among the items. This regularity can be used to characterize the students and items on a common scale, even when students take different sets of items. However, IRT is only the first step in the scaling of PISA data that makes it possible to describe the distributions of student performance in populations or subpopulations, to estimate the relationships between proficiency and background information.

4.1 Test equating

In the PARAKH RS, test forms will be linked through common linking items, which will consist of a block of shared items across each test form. To generate scale scores that accurately reflect student performance across different levels and test forms, a linking process is necessary.

To ensure comparability across different test forms by accounting for variations in difficulty rather than content, the concurrent calibration method will be used for horizontal linking. This method combines data from multiple test forms into a single dataset, calibrating all items simultaneously,

which enhances the accuracy of cross-form comparisons. Reliable linking depends on having a sufficient number of high-quality linking items across all test forms.

In PARAKH RS analyses, this approach will be used to link six test forms (Forms 31 to 36 and Forms 61 to 66) for Grades III and VI, and eight test forms (Forms 91 to 98) for Grade IX. Following the PARAKH RS assessment framework, the Balanced Incomplete Block (BIB) design will be implemented to strengthen the linkage between forms. A thorough review process will assess the effectiveness and performance of these linking items across different forms to ensure robust and consistent results.

Vertical equating, which links student performance across different grades, will not be within the scope of proposed PARAKH RS 2024 analyses.

4.2 Scaling methods

The two-parameter logistic (2PL) model was used to calibrate the items in the NAS 2017 and 2022 and this model will again be used for the PARAKH RS 2024. In the 2PL model (Birnbaum, 1968), the probability of a correct response for item *j* is given by

$$\Pr(Y_j = 1|\theta) = \frac{\exp\left[a_j(\theta - b_j)\right]}{1 + \exp\left[a_j(\theta - b_j)\right]},$$

where θ is the latent student proficiency, a_j is the item discrimination parameter, and b_j is the item difficulty. A benefit of using this IRT model is that student proficiency and item difficulty are on the same scale. Under the assumption of local independence, the probability of a vector consisting of *m* item responses is given by

$$\Pr(\boldsymbol{y}|\theta) = \prod_{j=1}^{m} \Pr(y_j|\theta).$$

The parameters of the 2PL model can be estimated by maximizing the marginal likelihood given by[2]

$$L(\mathbf{Y}) = \prod_{i=1}^{N} \int \Pr(\boldsymbol{y_i}|\theta) f(\theta) d\theta\,,$$

where $f(\theta)$ is the standard normal density. Several existing software packages can be used to estimate the 2PL model. A multigroup version of the 2PL model is obtained by allowing the mean and variance of θ to differ across predefined groups. For example, in PISA, groups are defined by the

[2] Sampling weights are left out of the equation for convenience.

combination of country/economy and language (e.g., Canada-French). In PARAKH RS, the groups are defined by the combination of state/union territory and language.

Since the 2PL model is based on key assumptions related to the shape of the item response function, local independence, and unidimensionality, model fit will be evaluated. In LSA, this is generally done by evaluating item fit statistics and tests of local independence and unidimensionality. The evaluation of model fit and item-by-language interactions in PARAKH RS 2024 will assess how allowing item parameters to vary for specific language groups enhances overall model performance.

4.3 Conditioning and plausible values

Once item difficulties are placed on the scale, students' scale scores can be computed, with ability estimates determined through plausible values (PVs). PARAKH Rashtriya Sarvekshan 2024 will utilize plausible values (PVs), following the approach outlined by von Davier, Gonzalez, & Mislevy (2009), for reporting results. While weighted maximum likelihood estimates (WLEs) were used to report the results for NAS 2021, PVs are more commonly employed in large-scale survey assessments, such as PISA, PIRLS, TIMSS, and NAEP.

PVs are technically multiple imputations drawn from students' posterior distributions of proficiency, based on their item responses and background characteristics (Wu, 2005; Marsman, Maris, Bechger, & Glas, 2016). These values represent a plausible range of proficiencies that a student might reasonably possess, capturing the inherent uncertainty in the assessment process. An advantage of PVs is their suitability for group-level reporting and for relating proficiency to background variables. PVs, in combination with sampling weights, are also commonly used to estimate variances of reported statistics.

Using item parameters anchored to their estimated values from the calibration sample, plausible values are random draws from the marginal posterior of the latent proficiency distribution. Estimations will be based on the conditional item response model and the population model, which includes a regression on background variables used for conditioning. The variables in the background questionnaires are incorporated as regressors in the conditioning model.

Typically, in LSA, the following procedure is used to create PVs: A regression model is developed to predict students' latent ability from their responses to the test and their background characteristics. The prediction model is imperfect, so a measure of the remaining uncertainty in ability is also created. Using the prediction model, the measurement of uncertainty and a student's responses and background data, a set of plausible values for each student is generated. These plausible values can then be used to estimate quantities of interest such as the mean scores for student groups of interest,

e.g., states and UT or gender groups, etc., the percentiles of the ability distribution, or the proportion of students who are proficient.

PVs provide a more robust representation of population performance compared to point estimates such as WLE or MLE. A set of *K* plausible values will be generated for each student for each scale. Population statistics should be calculated using each PV separately, and the reported statistic will be the average of these *K* estimates.

For example, if the analysis aims to assess the mean difference in performance between boys and girls, the mean score for boys will be computed based on each PV and averaged across the *K* values. The same will be done for girls, and the group means will then be compared to assess the performance differences.

Development of the prediction model starts with the creation of a set of conditioning variables from the background questionnaires. Then, two approaches can be distinguished: In the first approach, followed in NAEP, a set of key reporting variables is identified which are entered directly in the conditioning model. The remaining set of variables can be put through a principal component analysis (PCA) to reduce the number of variables to be added in the conditioning model while explaining most of the variance in these variables (e.g., 80-90%). In the second approach, followed in PISA, the full set of conditioning variables is put through PCA and the number of components that explain 80% of the variance is kept in the model with a maximum of 5% of the sample size (i.e., 315 components can be kept with PISA's target sample size of 6,300 students per country/economy).

A downside of this PCA approach is that it ignores the relation with proficiency. An alternative method that can be used is a partial least squares (PLS) regression approach in which proxies of student proficiency (e.g., a weighted maximum likelihood or expected a posteriori estimate) are entered as well (see, e.g., Robitzsch, Pham, & Yanagida, 2016). This leads to components that explain both a large part of the variance in the conditioning variables and correlate as highly as possible with student proficiency.

After either one of these procedures procedures, the latent regression IRT model can be estimated. The latent regression of the multidimensional proficiency $\boldsymbol{\theta}$ (e.g., language, mathematics, and the world around us for Class 3) on the predictors $\mathbf{x}$ (e.g., direct covariates + selected components) can be stated as

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\Gamma}'\mathbf{x}_i, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Gamma}$ is a matrix of regression parameters and $\boldsymbol{\Sigma}$ is the conditional covariance matrix. This produces latent regression parameter estimates with which the posterior distribution of student proficiency can

be determined to draw plausible values (Wu, 2005; von Davier, Gonzalez, & Mislevy, 2009). This posterior distribution is given by

$$h(\boldsymbol{\theta}_i|\mathbf{y}_i, \mathbf{x}_i) = \frac{Pr(\boldsymbol{y}_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\mathbf{x}_i)}{\int Pr(\boldsymbol{y}_i|\boldsymbol{\theta}_i) f(\boldsymbol{\theta}_i|\mathbf{x}_i) d\boldsymbol{\theta}_i}.$$

Multiple plausible values (PVs) for each student can be drawn from the above posterior (e.g., NAEP use 5 PVs; PISA uses 10 PVs).

Suppose we are interested in some summary statistic, such as the population mean. Then, first, one has to compute the statistic for each set of PVs. The estimate of the population statistic is then simply the average of the statistics computed for each set of PVs. For example, for the mean, it would be: $\hat{\mu} = K^{-1} \sum_{k=1}^{K} \hat{\mu}_k$, where *K* is the number of PVs. The standard error can be found by calculating the sampling error and the imputation error. The imputation error variance is computed as: $\mathrm{Var}_{\mathrm{imp}}(\hat{\mu}) = \frac{1}{K-1} \sum_{k=1}^{K} (\hat{\mu}_k - \hat{\mu})^2$. The sampling error $Var_{sampling}(\hat{\mu})$ can be estimated with, for example, the jackknife or balanced-repeated-replication technique, taking into account within-school and between-school variation. The standard error of the mean is then found to be $SE(\hat{\mu}) = \sqrt{Var_{sampling}(\hat{\mu}) + Var_{imp}(\hat{\mu})}$. The approach to obtain estimates for other statistics such as variances, percentiles, correlations, and regression parameters would be the same.

4.4 Plausible values in educational assessments

As noted, PVs provide a statistical approach for representing a range of potential abilities in LSAs. To understand this concept, we can consider an analogy from measuring tree heights in an urban survey. Suppose a city's environmental department is tasked with estimating the average height of trees in urban parks. Due to time constraints, surveyors use a simplified method: measuring tree heights to the nearest whole meter using a laser device. They report heights as 5 meters, 6 meters, 7 meters, and so on.

However, in reality, tree height is a continuous variable. For instance, a tree reported as 6 meters tall might actually measure 5.85 meters or 6.25 meters. Two types of errors affect these measurements:

- Rounding Error: Heights are recorded to the nearest whole number, leading to rounding inaccuracies.
- Measurement Error: Device inaccuracies or environmental factors might slightly distort the recorded values.

Trees with actual heights close to 6 (e.g., 5.9 or 6.1 meters) are more likely to be reported as 6 meters. Trees near the midpoint (e.g., 6.5 meters) are equally likely to be reported as 6 or 7 meters.

This variability reflects the uncertainty in measurement, a concept central to PVs. The methodology assigns plausible values by:

1. Mathematically modeling posterior distributions around reported values.
2. Randomly drawing values from these distributions to reflect the range of possible true heights.

For example, a tree reported as 6 meters tall might be assigned plausible heights like 5.95, 6.10, or even 6.25 meters.

This analogy applies to student assessments. For instance, in a test comprising six dichotomous items, a student's cognitive ability, a continuous latent variable, can be represented as a discrete total score based on the number of correct responses. The possible total scores on these six items are limited to integers: 0, 1, 2, 3, 4, 5, or 6.

However, a student's true ability exists on a continuous scale and is shaped by various factors, including:

- Testing Conditions: Environmental factors and the overall testing environment.
- Mental and Physical Readiness: The student's state of mind and physical health on the test day.
- Question Difficulty: The alignment of test item difficulty with the student's skill level.

There are significant overlaps in the posterior distributions. In the tree height analogy, measurement error in the posterior distributions is assumed to be independent of the city or the trees being measured. However, in educational assessments, measurement error often depends on students' proficiency levels. It tends to be smaller for average-performing students and larger for low- or high-performing students, relative to the test's difficulty. Additionally, in this context, the posterior distributions for extreme scores, such as 0 and 6, are notably skewed, indicating that they are not normally distributed, as illustrated in Figure 5.

Generating plausible values in educational assessments involves drawing random samples from posterior distributions. This approach highlights that plausible values are not suitable for evaluating individual performance. For example, a student with a raw score of 0 might receive plausible values ranging from -3 to -1, while another with a score of 6 could receive values from 1 to 3.

Simply put, PVs represent the range of abilities a student might reasonably possess. Rather than directly estimating a student's ability with a single point, such as a WLE, a probability distribution of potential abilities is determined. A strength of using this approach is that, apart from the item responses, PVs can be informed by additional information from the questionnaires, which generally

increases their precision. Plausible values are then random draws from the estimated proficiency distribution conditional on item responses and background information (Wu, 2005).

PV methodology transforms discrete test scores into a continuum of possible proficiencies, mitigating bias in reporting group-level results when measuring an underlying ability with a limited set of test items. Additionally, an individual ability estimate can be derived from posterior distributions using the Expected A Posteriori (EAP) estimator. Unlike plausible values, which are random draws, the EAP assigns the mean of the posterior distribution as a single estimate for a student. Consequently, the EAP can be seen as the average of an infinite set of plausible values for a given student (refer Figure 5).

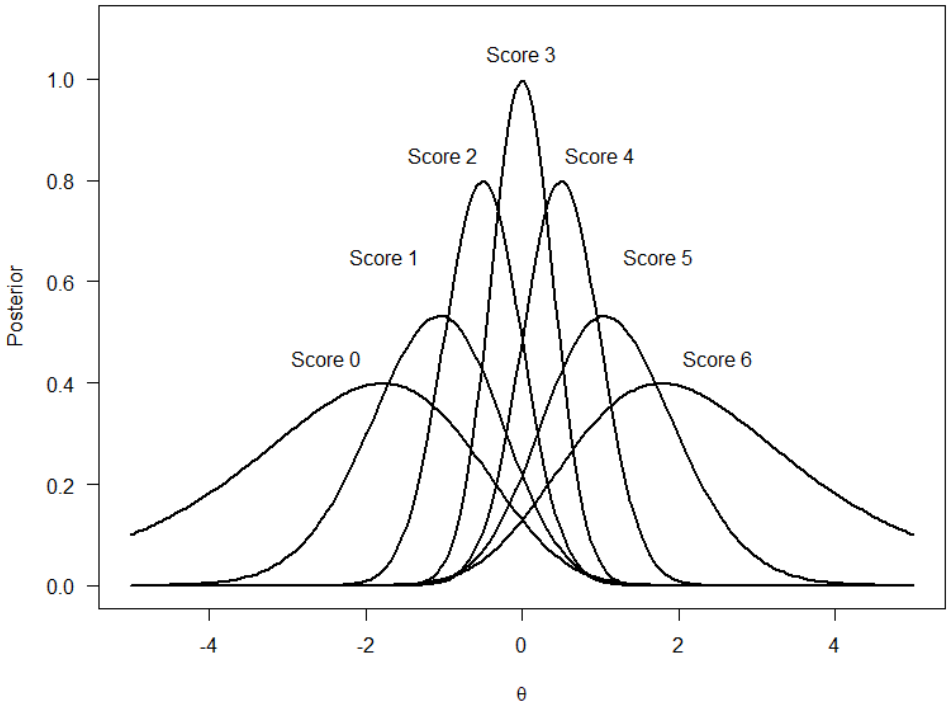

Figure 5. Different posterior distributions of student proficiency for different scores.

Since the EAP estimator assigns a single value per posterior distribution, it too is a discontinuous variable. However, there are key differences between EAP estimates and WLEs. EAP requires an assumption about the population distribution, whereas WLE does not. Additionally, while a specific response pattern on a test will always correspond to a unique WLE, multiple EAP values may be associated with the same response pattern. These variations depend on the regressors used as conditioning variables.

4.5 Estimating student ability

PARAKH RS 2024 will utilize PVs to generate student scores, with the intention to produce 5 PVs for each student on each scale or subscale. PVs provide a more robust representation of population

performance than point estimates like WLE or MLE, making them the preferred method for PARAKH RS 2024.

Educational assessments generally serve two purposes:

1. Measuring Individual Knowledge and Skills:

This focuses on individual student performance, often impacting academic or career outcomes. Minimizing measurement error at the individual level is critical, and methods like Weighted Likelihood Estimation (WLE) are typically used for precise individual scoring.

2. Assessing Population Knowledge and Skills:

This focuses on evaluating the performance of populations rather than individuals, with the aim of minimizing errors in population-level estimates. Large-scale assessments like PARAKH RS 2024 fall under this category, where PVs are essential for accurate group-level analysis.

By capturing the uncertainty in proficiency estimates, PVs ensure robust and reliable insights for population-level reporting, aligning with best practices in national and international assessments.

In PARAKH RS 2024, PVs are used instead of WLE for the following reasons:

1. Population-Level Accuracy: PVs are designed to reduce bias and improve accuracy when estimating population-level statistics, which is the primary goal of such surveys.
2. Capturing Uncertainty: PVs take into account the uncertainty in a student's proficiency estimate, reflecting the range of possible true scores rather than a single point estimate. This approach is critical for drawing reliable inferences about groups.
3. Alignment with International Standards: International assessments like PISA, TIMSS, and others use PVs for reporting student performance. Using PVs in PARAKH RS 2024 ensures consistency with global best practices.

While WLE provides precise estimates for individuals, its use in population-level analyses can introduce biases. PVs, on the other hand, enable robust estimation of group-level statistics, which aligns with the objectives of the PARAKH RS 2024 survey.

Population statistics will be calculated using each PV separately, and the reported statistic will be the average of these 5 estimates. For example, to assess the mean difference in performance between boys and girls, the mean score for boys is computed for each PV and averaged across the 5 values. The same calculation is done for girls, and the group means are compared to assess performance differences.

4.6 Utilizing correlations between subjects

Since the number of cognitive items for each subject is relatively low (15-20 items per subject), this limits the measurement precision (i.e., test reliability) at the student level. However, since students are tested in three subjects for Classes 3 and 6, and four subjects in Class 9, the correlations among the subjects can be utilized to improve the measurement precision at the group level. These correlations are exploited in the conditioning phase leading to multidimensional latent regression models.

5 Linear transformation

The IRT software will place item parameters on a continuous latent scale measured in logits with mean zero and standard deviation one. As a result, approximately half of the students receive negative scores on this scale. Since it is generally undesirable to report negative scores, the latent scale is typically transformed. After the equating and scaling processes, the scores in logits will be transformed to a scale with a chosen mean and standard deviation by applying a linear transformation.

This allows scores to be reported from a test on a readily understandable scale. For PARAKH RS 2024 scale scores will be calculated with mean of 300 and a standard deviation of 50. In this way, student scores fall within the range of 100 and 500 in general. For the PARAKH RS, there will be different scales will be developed for different all three grades and subjects.

Transformation of logit values into scale score with a set mean and a set standard deviation is computed as-

$$Scale\ score = \{\text{Logit value (on scale)} * Standard\ Deviation\ (Set\ value)\} + Mean\ (Set\ value)$$

An example of using the above formula and calculating the scale score of a grade 6 student having logit score of 1.28 would be -

$$Scale\ Score = (1.28 * 50) + 300 = 364$$

6 Development of proficiency levels

In addition to establishing a metric for the proficiency scale, PARAKH RS will define a set of proficiency levels with substantive descriptions, providing a structured interpretation of assessment results. Although the proficiency scale is continuous, it will be divided into discrete levels, each associated with a distinct description of skills. These proficiency levels will synthesize item content at each level and translate it into descriptions of skill and understanding, making student achievement profiles more accessible and interpretable.

6.1 Methods for setting proficiency levels

1. **Bookmark method:** The Bookmark Method is a widely used standard-setting technique where test items are ordered by difficulty, and panels of content experts review and establish "bookmarks" or cut points where proficiency levels begin and end. This involves setting points where the probability of a student answering items correctly aligns with the proficiency level descriptions. Through this technique, performance standards and proficiency levels are informed by a combination of empirical evidence and expert judgment.
2. **Judgmental and Empirical Methods:** Proficiency standards and cut points are also validated using a judgmental-empirical approach, where expert panels review the item difficulty levels (in logits) and set cut scores based on observed student performance patterns. Empirical data from field studies, item difficulty, and performance expectations for various skills are used to refine these cut points. Experts may adjust cut points to ensure the descriptions match the observed skills of students at each level.
3. **Latent Trait Modelling:** Using Item Response Theory (IRT), latent trait modelling helps calibrate items and identify natural breakpoints in student performance distributions. Through latent regression, proficiency bands are fine-tuned to represent meaningful intervals on the scale, which reflect common patterns in the data and are aligned with instructional standards.

6.2 Proposed method to derive proficiency levels and their descriptors

Defining proficiency standards involves a balanced approach that integrates empirical data analysis and expert judgment, ensuring the standards are statistically robust and substantively meaningful. PARAKH RS 2024 will implement a comprehensive methodology that combines Item Response Theory (IRT), expert panel reviews, and validation techniques.

1. Empirical Item Calibration with IRT

- Fit the 2PL IRT model to estimate item parameters (difficulty and discrimination). This process results in item response functions (IRFs), mapping item difficulty along the latent scale, which will become the foundation for setting cut-off points.
- **Interpreting Item Thresholds** Analyze the item difficulty values to determine where each item aligns on the proficiency scale, which will enable the identification of preliminary cut-off points for proficiency levels.

2. Define Initial Proficiency Cut Points
    - Arrange items by difficulty, creating a spectrum of item challenges along the latent scale continuum. This arrangement will provide insights into natural groupings of item difficulties, facilitating the identification of preliminary cut-off points.
    - Locate natural breakpoints in the distribution of item difficulties. These breakpoints will serve as preliminary cut points to segment the proficiency scale into bands that distinguish varying competency levels, ensuring they correspond with coherent and meaningful performance standards.
    - In alignment with **NAS 2021**, PARAKH RS will utilize three proficiency levels:
        I. **Basic**: Demonstrates partial knowledge and skills in subject matter.
        II. **Proficient**: Shows solid understanding and ability to apply core concepts.
        III. **Advanced**: Displays high competence, including application and synthesis of skills.
    - An initial cut-off value of 1.5 logits will be considered as a starting point for defining the proficiency levels.
3. Draft Proficiency Descriptions
    - Based on item content, draft preliminary descriptions of what students in each level are expected to know and be able to do. This step will ensure that each level corresponds to a coherent set of skills.
4. Expert Panel Review - Bookmark Method
    - **Assemble a Panel of Experts**: Convene a group of subject matter experts and educators to review the draft proficiency levels.
    - **Use the Bookmark Method**: Present ordered item difficulties to the panel and will ask experts to place “bookmarks” where they believe students should demonstrate certain skills to reach the next proficiency level. This technique validates that the cut points correspond to meaningful levels of proficiency as judged by experts.
    - **Refine Descriptions**: Based on expert feedback, proficiency level descriptions will be refined to capture the expected competencies at each level clearly and accurately.
5. Validation with Latent Regression Analysis

- Implement latent regression to model relationships between proficiency levels and demographic/background variables (e.g., from the pupil questionnaire). This step will verify that proficiency cut-offs accurately reflect different student performance profiles and demographic backgrounds.
- Based on latent regression findings, cut-off points will be adjusted if biases appear in performance across subgroups, ensuring the final cut points maintain fairness and validity across student populations.

6. Finalize and Document Proficiency Standards

Final checks will be conducted before finalizing the proficiency standards for PARAKH RS 2024.

- **Define Final Proficiency Levels:** Each proficiency level will be established, ensuring alignment with the test's purpose and providing clear, educationally meaningful descriptions.
- **Create Proficiency Descriptors:** Descriptors for each level will be developed to ensure they are straightforward and easily interpretable by educators, policymakers, and stakeholders.

7 Utilization and interpretation of scores for reporting

In PARAKH RS 2024, score reporting aims to provide a clear, fair, and meaningful interpretation of results for diverse stakeholders, including policymakers, educators, researchers, and the general public. The following principles and methods will guide the reporting of scores:

7.1 Scale scores and comparisons

- Scores will be scaled using the 2PL IRT model to ensure comparability across assessments and subgroups. The scale scores will be designed to provide a standardized range, with mean and standard deviation values to contextualize performance.
- Mean scale scores and their standard errors will be reported at the national, state/UT, and district levels for each subject and grade combination.
- Differences between subgroup mean scores (e.g., gender, location, school management type, and social groups) will be analyzed. Statistical tests such as t-tests will determine the significance of differences, and effect sizes will be calculated to contextualize the magnitude of differences.
- For contextual analysis, comparisons will incorporate pupil, school, and teacher questionnaire data. Direct comparisons will include variables like gender, age, social group, teacher qualifications, and class size.

7.2 Reporting by performance levels

- Students' performance will be categorized into predefined levels such as "basic," "proficient," and "advanced." Each level will include a description of learner characteristics and competencies.
- Results will be presented with corresponding proportions of students at each performance level, facilitating meaningful interpretation for stakeholders.

7.3 Subgroup and contextual comparisons

- Indices will be created for more complex constructs like socioeconomic status, student attitudes towards learning, and school infrastructure. These indices will be derived using statistical techniques such as Principal Component Analysis (PCA) or Exploratory Factor Analysis (EFA).

- Regression and correlation analyses will provide deeper insights into relationships between contextual factors and student performance.

7.4 Reporting considerations

- **Subgroup Reporting Rules:** A minimum sample size will be established to ensure reliable estimates for subgroups, with larger error margins highlighted for smaller groups.
- **Competency-Level Reporting:** The number of items contributing to each competency will be reported, providing clarity on the basis of competency-level performance.

7.5 Data visualization and accessibility

- Results will be visualized using maps, graphs, and charts to enhance accessibility and understanding. For example, color-coded maps will indicate significant differences between subgroups at various geographic levels.

PARAKH RS 2024 analysis framework will ensure that results are both statistically robust and accessible, supporting informed decision-making and policy formulation.

8 References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Addison-Wesley.

Bose, R. C. (1939). On the construction of balanced incomplete block designs. *Annals of Eugenics, 9*(4), 353–399.

de Ayala, R. J. (2013). The theory and practice of item response theory. Guilford Press.

Educational Testing Service. (2014). *ETS standards for quality and fairness.* Retrieved from https://www.ets.org/about/fairness/review-publications.html.

Fisher, R. A. (1935). *The design of experiments.* Oliver and Boyd.

Frey, A., Hartig, J., & Rupp, A. A. (2009). Booklet designs in large-scale assessments. *Educational Measurement: Issues and Practice, 28*(3), 39–53. https://doi.org/10.1111/j.1745-3992.2009.00153.x

Joo, S., Rutkowski, D., & von Davier, M. (2021). Advancing methodological research for international large-scale assessments. *Educational Measurement: Issues and Practice, 40*(1), 40–50. https://doi.org/10.1111/emip.12345

Khorramdel, L., von Davier, M., & Bertling, J. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study.

Kim, S. (2012). IRT reliability: An indicator of measurement precision. *Applied Psychological Measurement, 36*(4), 291–307. https://doi.org/10.1177/0146621612441548

Knapp, J. (1968). Balanced incomplete block designs for large-scale assessments. *Journal of* Educational Measurement, 5(4), 317–322.

Livingston, S. A. (2020). *Test score scales.* Routledge.

Livingston, S. A., & Dorans, N. J. (2004). A graphical approach to item analysis. *ETS Research Report* Series, 2004(1), i–17.

Marsman, M., Maris, G., Bechger, T., & Glas, C. (2016). What can we learn from plausible values? *Psychometrika, 81*(2), 274–289. https://doi.org/10.1007/s11336-016-9497-x

Messick, S., Beaton, A. E., & Lord, F. M. (1983). Balanced incomplete block designs in educational measurement. Journal of Educational Statistics, 8(3), 143–160.

NCERT. (2019). NAS 2017 Class III, V and VIII: National Report to Inform Policy, Practices and Teaching Learning. National Council of Educational Research and Training.

NCERT. (2022). National Achievement Survey: National Report NAS 2021 Class III, V, VIII & X. National Council of Educational Research and Training.

NCERT. (2023a). National Achievement Survey 2021: Technical Note on Assessment Framework. National Council of Educational Research and Training.

NCERT. (2023b). Notes on Sampling Design for National Achievement Survey (NAS) 2021. National Council of Educational Research and Training.

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score-based comparisons through the use of performance moderators in international large-scale assessments. *International Journal of Testing, 14*(1), 1–22. https://doi.org/10.1080/15305058.2013.874948

Organisation for Economic Co-operation and Development (OECD). (2022). *PISA 2022 Technical Report* (Chapter 11). OECD Publishing.

Robitzsch, A., Kiefer, T., & Wu, M. (2024). A package for large-scale assessment surveys such as PISA and TIMSS.

Robitzsch, A., Pham, G., & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In S. Breit & C. Schreiner (Eds.), Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung (pp. 259–293). LexisNexis.

Robitzsch, A., Pham, H., & Yanagida, T. (2016). Partial least squares regression for large-scale assessment data. *Journal of Educational Measurement, 53*(2), 203–227. https://doi.org/10.1111/jedm.12104

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2014). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Research and Evaluation, 20*(5), 420–446. https://doi.org/10.1080/13803611.2014.949917

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments* (pp. 9–36). Springer. https://doi.org/10.1007/978-1-4419-1258-8_2

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450. https://doi.org/10.1007/BF02294627

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2–3), 114–128. https://doi.org/10.1016/j.stueduc.2005.05.005