International Early Learning and Child Well-being Study

TECHNICAL REPORT

OECD
BETTER POLICIES FOR BETTER LIVES

INTERNATIONAL EARLY LEARNING AND CHILD WELL-BEING STUDY

TECHNICAL REPORT

Table of contents

FIGURES

TABLES

Acknowledgements

The International Early Learning and Child Well-being Study (IELS) was a collaborative effort between participating countries and the OECD Secretariat. To support the technical implementation of IELS, the OECD contracted an international consortium of institutions, led by Maurice Walker at the Australian Council for Educational Research (ACER). The consortium included the International Association for the Evaluation of Educational Achievement (IEA), led by Anja Waschk, and cApStAn, under the lead of Roberta Lizzi.

To help to ensure that the study was rigorous and valid, IELS was guided and supported in its work by a Technical Advisory Group (TAG) and a Measurement Advisory Group (MAG). Both groups provided input to the assessment framework for the study, instrument development for the assessments and contextual information, and the analysis of the findings.

This volume was prepared in close collaboration with the IELS teams from England, Estonia and the United States, led by Frances Forsyth, Tiina Peterson and Mary Coleman respectively. Their input to this technical report is gratefully acknowledged.

A complete list of contributors can be found in Annex A.

April 2020

List of Abbreviations/Acronyms

The following abbreviations/acronyms are used in this report:

1PL	1 Parameter Logistic
ACER	Australian Council for Educational Research
AU	Audio Upload
AUF	Audio Upload - Female
AV	Audio Verification
AVF	Audio Verification - Female
BRR	Balanced Repeated Replication
cApStAn	cApStAn Linguistic Quality Assurance
CAT	Computer-Assisted Translation
CTT	Classical Test Theory
DA	Direct Assessment
DIF	Differential Item Functioning
DME	IEA Data Management Expert
DOB	Date of Birth
EAP/PV	Expected A-Posteriori/Plausible Value
ECEC	Early Childhood Education and Care
ENG	England (United Kingdom)
ELG	Early Learning Group
ERI	Educational Research Institute
EST	Estonia
FC	Final Check
FCF	Final Check - Female
FSM	Free School Meal
FT	Field test
GA	Gender Adaptation
ICC	Item Characteristic Curves
ICT	Information Communication Technology
IELS	International Early Learning and Child Well-being Study 2018
ID	Unique Identifier
IDB	International Database
IEA	International Association for the Evaluation of Educational Achievement

IQAM	International Quality Assurance Monitor
IR	International Review
IRT	Item Response Theory
ISC	International Study Centre
ISCED	UNESCO International Standard Classification of Education
ISO	International Organization for Standardization
LSA	Large Scale Assessment
MAG	Measurement Advisory Group
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte-Carlo
ML	Maximum Likelihood
MNSQ	Mean Square
MOS	Measure of Size
MRCMLM	Multi-dimensional Random Coefficients Multinomial Logit Model
MS	Main study
NA	National Adaptation
NAR	National Audio Review
NARF	National Audio Review - Female
NAV	National Adaptation Validation
NCES	National Centre for Education Statistics
NDM	National Data Manager
NPM	National Project Manager
NQAM	National Quality Assurance Monitor
NR	National Review
NRBA	Non-Response Bias Analysis
NRF	National Review - Female
NSM	National Sampling Manager
OARS	ACER's Online Assessment and Reporting System
OECD	Organisation for Economic Co-operation and Development
PCA	Principal Component Analysis
PCM	Partial Credit Model
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
PSU	Primary Sampling Unit
PUF	Public Use Files
PV	Plausible Value
QQ	Questionnaire
RCMLM	Random Coefficients Multinomial Logit Model
REC	Reconciler
RIS	Release International Source
RUF	Restricted Use Files

SAQ	Survey Activities Questionnaire
SEN	Special Education Needs
SOP	Survey Operations Procedures
SR	Self-Regulation
SRS	Simple Random Sampling
T1	Translator 1
T2	Translator 2
TAG	Technical Advisory Group
TALIS	OECD Teaching and Learning International Survey
TAVM	Translation, Adaptation and Verification Monitoring Workbook
TMS	Translation Management System
TV:	Translation Verification
TVF	Translation Verification - Female
USA	United States
WinW3S	IEA Within-school Sampling Software
WM	Working Memory
XLIFF	XML Localisation Interchange File Format

Chapter 1. Overview

The International Early Learning and Child Well-being Study (IELS) is an international survey that assessed skills of children at age 5 attending early childhood education centres or schools in England (United Kingdom)[1], Estonia and the United States, in 2018. The aim of the study was to identify key factors that drive or hinder the development of early learning (OECD, 2020[1]).

IELS directly assessed children's emergent literacy, emergent numeracy, some self-regulation skills and some social-emotional skills (see Chapter 2.). IELS indirectly assessed, through teacher and parent reports, some cognitive and social and emotional skills and behaviours. Teachers and parents were also asked to report on other contextual and historical factors (see Chapter 4.).

This report documents the technical details of the study. Other important aspects of the study, such as the findings and the assessment framework are detailed in other publications; see a current list of these publications in the final section of this chapter.

1.1. Innovative features of IELS

As the population of interest was 5 year-olds, special consideration was given to engaging the children in the assessment. The assessment was delivered to each child on a tablet, using colourful and interactive content (see Chapter 2. and Chapter 3.).

Children were presented with a variety of stories, vignettes and game-like activities using cartoon-like illustrations. All instructions and stories were delivered through pre-recorded, standardised audio – no reading at all was required (see Chapter 2. and Chapter 6.).

The assessments were delivered in a one-to-one manner, by trained Study Administrators. To reduce the likelihood of assessment burden, the assessments were conducted over two days (see Chapter 7.).

Another innovative feature of the study was its use of multi-dimensional scaling models. For example, a ten-dimensional scaling model was applied to the data to maximise the information available across all the cognitive and social-emotional domains (see Chapter 10. and Chapter 11.). Such a multi-dimensional model has not been applied previously in international studies.

1.2. Structure of the technical report

Chapter 1. provides an overview of this report, highlighting the innovative features of IELS. The management responsibilities of IELS are outlined and the IELS publications are listed.

Chapter 2. describes the design and development of the IELS assessments. This includes detailing the three phases of item development: item trial, field test and main study. The final assessment design is provided.

Chapter 3. describes the delivery platforms on which the assessments and questionnaires were delivered. The chapter includes descriptions of all the delivery and supporting components: content management; translation management; candidate management; assessment delivery; and, questionnaire delivery. An outline of data security measures is provided.

Chapter 4. describes the development of the contextual questionnaires. It outlines the contextual framework and describes development of the parent and staff questionnaires. An outline of the process for adaptation and quality assurance is provided.

Chapter 5. describes the IELS sample design. The definition of the target population is explained and the characteristics of the target population in each country are described. Exclusions to the target population are detailed and the method for determining sample sizes is described. The process of sample selection is detailed and exemplified.

Chapter 6. describes the linguistic quality assurance procedures in IELS. These include translatability assessment of items, translation and adaptation guidelines, adaptation documentation, linguistic verification of translations and the various review processes. The quality assurance of the audio component of the assessment is outlined.

Chapter 7. describes the field operations of the IELS survey. The operational roles and responsibilities of national staff are described, and the operational manuals are outlined. The operational procedures are detailed including the preparation of the tablets and the one-to-one survey administration guidelines.

Chapter 8. describes the quality assurance measures designed and implemented for IELS. These consisted of three components: the International Quality Assurance Monitoring (IQAM) programme; the National Quality Assurance Monitoring programmes co-ordinated by the National Study Centres; and the Survey Activities Questionnaire (SAQ).

Chapter 9. describes the survey weighting. It describes why sampling weights are needed and explains the various weighting elements that account for differing selection probabilities and non-response. The production of a final student weight is detailed and the rationale for the use of weights in statistical analysis is outlined. The variance estimation method used for this study, Balanced Repeated Replication, is also detailed.

Chapter 10. describes the procedures undertaken to apply Item Response Theory (IRT) scaling and plausible value methodology to the assessment data. The use of practice blocks and stop rules within the direct assessments is described and their use in and effect on the scaling outcomes is detailed. National and international item calibration is outlined and the handling of item by country/language/gender interactions is summarised. The ten-dimensional scaling model is described as is the reporting transformation.

Chapter 11. presents the outcomes of the scaling procedures. The targeting and distribution of item difficulties is illustrated. Reliabilities by dimension are provided, and treatments due to Differential Item Functioning (DIF) are detailed.

Chapter 12. describes the IELS data management processes. The various data sources are described and the quality assurance procedures for ensuring data integrity in the national and international datasets are detailed. The processes around data cleaning are outlined.

Chapter 13. describes the sampling outcomes of the IELS survey and presents indicators of the quality of the achieved samples. The achieved centres/school sample sizes for each participating country, the numbers of participating children, parents and staff members are presented. Estimates are provided for the total target population size and the exclusion rates in each participating country are given. Design effects and effective sample sizes are provided for key variables in each participating country.

Chapter 14. documents the data adjudication process. Adherence to the IELS Technical Standards was used to adjudicate the quality of the each participating country's data, particularly those standards around: sample exclusions and participation rates; adaptation, translation and verification of assessment and questionnaire materials; and, adherence to specified operational procedures. The outcomes of each country's adjudication process are presented.

Chapter 15. describes the scaling and construct validation procedures for the contextual questionnaires administered to parents and teachers. The indices derived from these instruments are described.

Chapter 16. describes the international data products. In particular, the format and contents of the publicly available international database are listed.

1.3. Managing and implementing IELS

Overall governance of IELS was through a secretariat based in the OECD Directorate for Education and Skills. The OECD managed the international contractor and liaised with National Project Managers (NPM) and national government representatives. The

OECD authored the main international report and the reports of the findings within each of the three participating countries.

The international contractor was a consortium, led by the Australian Council for Educational Research (ACER) and including the International Association for the Evaluation of Educational Achievement (IEA) and cApStAn. The main areas of each consortium partner's responsibilities are outlined in Table 1.1.

Table 1.1: Main areas of responsibility for IELS consortium partners

Consortium partner	Main areas of responsibility
ACER	Project management Assessment framework Assessment development Contextual questionnaire development Delivery platform development Constructing psychometric scales and indices for cognitive, social-emotional and contextual data Analysis of results and production of statistical tables
IEA	Field operations Sample design and selection, survey weighting International Quality Assurance Monitoring Data management International database construction
cApStAn	Linguistic quality assurance

IELS was guided and supported in its work by a Technical Advisory Group (TAG) and a Measurement Advisory Group (MAG). Both groups provided input to the assessment framework for the study, instrument development for the assessments and contextual information, and the analysis of the findings.

National Project Managers and national government representatives had a critical role in supporting, promoting and organising the study within each participant country. NPMs were responsible for gathering sample information, providing feedback on the assessment development, obtaining and configuring delivery hardware, training Centre/School Co-ordinators and Study Administrators, recruiting education centres/schools, teachers and parents to participate, managing their own data submission and developing their own national reports or other communications.

1.4. IELS publications

The main international results are published in (OECD, 2020[1]) Early Learning and Child Well-being: A Study of Five-year-Olds in England, Estonia, and the United States, OECD Publishing, Paris, https://doi.org/10.1787/3990407f-en.

Reports on the findings within each of the participating countries have also been published.

OECD (OECD, 2020[2]), *Early Learning and Child Well-being in England.* OECD Publishing, Paris. https://doi.org/10.1787/c235abf9-en

(OECD, 2020[3]), *Early Learning and Child Well-being in Estonia*, OECD Publishing, Paris. https://doi.org/10.1787/15009dbe-en.

(OECD, 2020[4]), *Early Learning and Child Well-being in the United States*, OECD Publishing, Paris, https://doi.org/10.1787/198d8c99-en

(OECD, 2021[5]), Assessment framework for the International Early Learning and Child Well-being Study, OECD Publishing, Paris.

Notes

[1] This report uses "England" as shorthand for England (United Kingdom)

Chapter 2. Assessment Design and Development

The IELS assessment framework was informed by the conceptual framework on early learning outcomes developed by the OECD Secretariat in collaboration with interested countries (OECD, 2015[6]).

The development of the assessment framework was influenced by the agreed principles and parameters within which the IELS early learning assessment was to be developed; that is, that the assessment should be:

- Policy relevant – enabling changes in policies and/or practices to be made.
- Practicable – able to be implemented.
- Reliable, valid and comparable across countries, languages, cultural contexts and over time.
- Ethical – ensuring the well-being of children in the study is paramount in all decisions.
- Efficient – limiting the burden on teachers and parents, as well as on children.
- Cost-effective –affordable for a range of countries.
- Sustainable – establishing a strong foundation for the assessment, for possible expansion to later cycles or linkage to other OECD studies and nationally-based assessments.

The development of the assessment framework respected the broad consensus among early childhood education and care (ECEC) experts that early learning and child well-being are formulated best as holistic constructs that include cognitive aspects, social and emotional aspects, and background contexts such as home life and ECEC experiences.

To achieve the above requirements, it was agreed that the IELS assessment would use a combination of direct child assessment and indirect parent/educator assessment.

The direct assessments would be delivered to the children using tablet devices and would assess emergent literacy, emergent numeracy, self-regulation, and social-emotional competence (Shuey and Kankaraš, 2018[7]).

A parent questionnaire would be used to collect information on children's literacy, numeracy, self-regulation, social-emotional skills, socio-demographic characteristics, parental background, home learning environment, and early childhood education participation.

A staff questionnaire would be used to gather information on staff background, parental involvement in the child's schooling, children's literacy, numeracy, social-emotional skills and self-regulation.

Before the commencement of item development, content specialist teams were established to develop and describe each of the assessment sub-domains based upon a review of the literature and in keeping with the principles and parameters listed above.

The sub-domain descriptions and related assessment blueprints were developed in conjunction with the OECD Technical Advisory Group and the Measurement Advisory Group, indicating the appropriate number and types of questions for children of 5 years of age. The proposals were shared with participating countries and modified accordingly. The resulting descriptions may be found within the IELS Assessment Framework (OECD, 2021[5]).

2.1. Phase 1 of item development

The following teams were responsible for the Phase 1 item development:

- Emergent literacy: ACER early years literacy team
- Emergent numeracy: ACER early years numeracy team
- Self-regulation (working memory, mental flexibility and inhibition): Educational Research Institute (ERI), Poland
- Empathy and trust: ACER early years team.

Each team developed a small set of sample items based upon the assessment framework to be used in cognitive laboratories in Australia and Poland in November 2016 and subsequent review by the OECD, MAG and TAG. The feedback received informed the subsequent item development in preparation for a small item trial in March 2017. The feedback related to the length of the assessments, terminology (e.g. touch rather than tap) and clarity of language, quality of images, the need for cohesive themes, etc. Only the United States was able to participate in the item trial.

2.2. Designing the instrument for the item trial

As part of the item development for the item trial, draft versions of the instruments were sent to the OECD and the countries for their consideration in February (for a list of instruments used in this study, see Annex B). Detailed feedback was received from the OECD and the United States with some feedback from Estonia and New Zealand. Feedback was collated and provided to the lead assessment developers for consideration. Items (text and art) were modified where appropriate and items were selected for item trial assessment forms taking the feedback into consideration.

The item trial was conducted over two days as follows:

- Day 1: Emergent numeracy (21 items) and self-regulation (91 items)
- Day 2: Emergent literacy (20 items) and empathy/trust (12 items)

Each domain was designed to last approximately 25-35 minutes including practice sessions.

Table 2.1, Table 2.2, Table 2.3, Table 2.4 give a breakdown of each domain by sub-domain.

Table 2.1: Emergent literacy item trial items

IELS Aspect	Task	Items	Number of Items
Listening comprehension	Story level aural meaning. Listen to an e-book story, answer audio supported items	Answer questions about: Prominent literal meaning, implied meaning, and information that needs to be linked	8
	Sentence-level aural meaning	Identify if a sentence is real or nonsensical. Complete a sentence with a word to make meaning	3*
Phonological awareness	Initial sounds End sounds Rhyme Middle sounds	Identify first sound in words including: blends Identify last sound Identify rhyme Identify middle sounds	9

*Two of these items were later reclassified as 'reading' items and determined inappropriate for the study.

Table 2.2: Emergent numeracy item trial items

Content Component	Description	Number of Items
Numbers and counting	Identify digits and numbers to 20	2
	Count from 1 to 20	2
	Compare numbers and positions (e.g. smaller, more, first etc.)	2
Working with numbers	Add and subtract in informal number story contexts	3
	Solve problems including doubling, halving, informal multiplication (repeated addition) and sharing	2
Measurement	Use everyday language to compare measures (e.g. longer, heavier, more etc.)	3
Shape and space	Use names of common shapes and objects and related language (e.g. sides, same shape, etc.)	2
	Use language of location (e.g. above, between etc.)	2
Pattern	Recognise and create patterns of shapes, objects and numbers	3

Table 2.3: Self-regulation item trial items

Content Component	Description	Number of Tasks	Number of Items
Working memory	Simultaneous storage and manipulation of information	1	17
Mental flexibility	Switch between different concepts/rules and think about multiple concepts/rules simultaneously	1	36
Inhibition	Override habitual/dominant behavioural response	1	38

Table 2.4: Empathy/trust item trial items

Content Component	Task Description	Item Description	Number of Tasks	Number Items
Empathy	Listen to a short story, look at the pictures, and answer audio and picture supported items	Answer each question about the feelings of the story characters by choosing one of the four emoticon faces (i.e. happy, sad, surprise, and angry) Answer each question about their own feelings in the same situation by choosing one of the same four emotion faces	4	8
Trust	Listen to a short story, look at the pictures, and answer audio and picture supported items	Answer each question about the story characters by choosing one of the three picture and audio supported options Answer each question about the themselves being in the same situation by choosing one of the three picture and audio supported options	2	4

For each domain, feedback on the following aspects was sought from the administrators;

- appropriateness of instruments for children at the age of 5
- clarity of language
- international and cultural applicability of the items
- functionality of the items on the tablets
- feasibility of capturing administrator feedback about child's behaviour
- feasibility of capturing children's feedback about their experience.

2.3. Summary of feedback by domain

A report on the March item trial was provided to the OECD in late April. The report included recommendations on the administration of the study together with the content of the direct and indirect instruments. A summary relating to the direct measures follows.

2.3.1. Emergent literacy

The item types and length of the item trial form (20 items) were appropriate. It was recommended to include some more challenging items in the field test. Additional

recommendations were made in regard to the number of items assessing understanding of the story and the stand-alone items.

It was recommended that

- the phonological awareness items should be unlinked from the story as the translation issues would be too problematic. Countries would need to select the sounds in their language that are the most distinct and commonly used for these items.
- generic 'encouragement' images should be used for the phonological awareness such as different smiley faces.
- the sentence-level tasks of identifying if a sentence is real or nonsensical should be replaced with more challenging sentence-level listening comprehension tasks.
- an additional comprehension story should be developed and that 30 items are selected for the field test with each child seeing 20 items.
- that the scope of the framework should be expanded to include some stand-alone vocabulary items.

Recommended composition of the field test items was:

- Story listening comprehension (35%) – 2 texts to go to trial
- Sentence-level listening comprehension (15%)
- Phonological awareness (25%)
- Vocabulary (25%).

2.3.2. Emergent numeracy

The children found the items engaging. Recommendations were made to fix three technical issues relating to the drag-and-drop items. The item types and length of the item trial form (21 items) was appropriate. It was recommended to include some more challenging items in the field test. Specific item comments were received relating to the content and suitability of 14 of the 21 items and modifications were made where appropriate.

2.3.3. Self-regulation

Eleven children took part in the self-regulation part of the item trial which comprised three tasks - Odd-One-Out (working memory), Shifting (mental flexibility) and Stop-Go (inhibition). It was found that:

- codes should be used to differentiate the reasons for session extension. In future, categorical variables should be used. e.g.: 1 = technical problem, 2 = break during WM, 3 = break during flexibility.
- slow loading of screens was a major problem for obtaining response times. Offline application is vital for any valid measure of automatic processing.
- all graphical elements should be one picture (animals, background and response buttons).
- audio scripts should be abbreviated.

2.3.4. Empathy and trust

Eleven children responded to 5 practice items, 4 empathy tasks (8 items) and 2 trust tasks (4 items) in the March item trial. To ensure children responded appropriately the empathy questions needed to be better anchored (specifically reference the events of the story). For example, the audio asks, "How do you feel about Tom?" Many children picked happy faces. Possibly this was because they liked Tom and so felt positive about him. If the question had directed the child to the story, possibly the answers may have been different. For example, the audio could say, "How do you feel about what just happened to Tom?"

Recommendations were made regarding clarity, artwork, number of tasks, scoring, reporting, and administrator training.

2.4. Phase 2 of item development

In order to act on the findings and recommendations from the March item trial, a second round of item development occurred. This involved retiring some items, modifying some art, text, audio scripts and functionality of the remaining items and developing some additional items to broaden the construct or extend the range of difficulty. Enough items for two field assessment forms were developed for each domain. Feedback on new items was again received from the OECD, MAG, TAG and the countries.

2.5. Selecting the items and designing the instrument for the field test

The final selection of items for the field test was made in conjunction with the OECD, MAG and TAG, taking into account framework coverage and estimated difficulty.

For *emergent literacy*, the construct was broadened to include a separate vocabulary component (Synonyms – select the word that means the same – 25%) and the proportions of other components adjusted accordingly. An additional story was developed. A small set of more challenging listening comprehension story and sentence-level items and harder vocabulary items were written. Altogether 34 emergent literacy items (across two 20-item assessment forms) were developed and selected for the field test. There were 6 items common to each assessment form and the assessment forms also included common practice items.

For *emergent numeracy*, the framework was extended to include repeated addition. To make the instruments more challenging, some items were modified and a small set of new, more challenging items were written. Altogether 32 emergent numeracy items (across two 22-item assessment forms) were developed and selected for the field test. There were 10 items common to each assessment form and the assessment forms also included common practice items.

For *self-regulation*, two identical assessment forms were developed, each with three components: working memory (139 items including training items), mental flexibility (103 items including training items), and inhibition (104 items including training items). After working through a set of practice items, each child was presented with 30 working memory items, 32 mental flexibility items or 60 inhibition items.

All three components included a branching structure that introduced more advanced items if a child performed above a certain level of achievement.

For *empathy and trust*, altogether 26 items (across two 18-item assessment forms) were trialled - 16 empathy items and 10 trust items. There were four empathy items and six

trust items common to each assessment form. The assessment forms also included common practice items.

The field test was conducted over two days in three countries during the months of November and December in 2017 with each child scheduled to take either assessment form 1 or assessment form 2. Each assessment consisted of six instruments as shown in Table 2.5.

Table 2.5: Timetable for field trial

Form	First session	Break	Second session
Assessment form 1	Emergent literacy	5 minutes	1.2 Emergent numeracy
Assessment form 2	1.3 Empathy/trust	5 minutes	1.4 Self-regulation – working memory 1.5 Self-regulation – mental flexibility 1.6 Self-regulation - inhibition
Form	First session	Break	Second session
Assessment form 1	2.1 Empathy/trust	5 minutes	2.2 Self-regulation – working memory 2.3 Self-regulation – mental flexibility 2.4 Self-regulation - inhibition
Assessment form 2	2.5 Emergent numeracy	5 minutes	2.6 Emergent literacy

Local adaptations to the instruments were negotiated, translation (into Estonian) and linguistic quality assurance took place, and the field test was conducted in England, the United States and Estonia with approximately 30 centres and 400 children in each country.

2.6. Phase 3 of item development

The psychometric analysis of the field test data was developed jointly with the OECD and the MAG. While most items had acceptable psychometric properties, some did not. For example, it was clear that the emergent literacy and emergent numeracy instruments were still too easy. It was agreed that some items were far too easy and should not be used for the main study, that some could be modified further to make more challenging, and that a very small subset of more challenging items should be developed. Because it would not be possible to trial the modified or new items it was agreed that another small pilot be conducted in Melbourne. A small number of items were flagged for showing Differential Item Functioning by country.

For *emergent literacy*, it was agreed to retain nine items without modification and to modify 11 items. From these, 18 items were selected for the main study. A further four more challenging items were written and piloted making 22 emergent literacy items in total for the main study. Three of the new emergent literacy items were categorised as 'Listening comprehension – sentence-level aural meaning'. The other new item was categorised as 'Vocabulary – synonyms'.

For *emergent numeracy*, it was agreed to retain 15 items without modification, to modify five items and to develop and pilot two more challenging items making 22 emergent numeracy items in total for the main study. The two partial credit items were modified to be dichotomous given that the analysis did not support the existence of two viable response categories. The two new items were both categorised as 'Working with numbers – Apply'.

For *self-regulation*, after the field test, it was found that some aspects of the administration needed modification for the main study. These included a reduction in the number of practice items, a change in the order of administration so that inhibition (easier) came first and mental flexibility (harder) came last, and a relaxation of the time limits for mental flexibility to four seconds to reduce the amount of missing data.

The following specific changes were made to questions based on data from the field test and feedback from experts.

For *mental flexibility*: animal images set to cow and sheep (re-sizing and flipping of sheep image required, background set to blue circle and square (rather than desert and forest), and new introductory screens introducing the sheep and the cow were created.

For *working memory*: new four window item cluster created and re-ordered to the pattern of Hard-Easy-Hard, changed all audio uses of the term 'Odd-One-Out' to 'Zebra'.

For *inhibition:* a new introductory screen was created, modifications were made to art so that greater cognition to differentiate would be required, and modified choice buttons.

For *empathy and trust*, the psychometric properties of the majority of field test items were not sufficient to permit progression to the main study. For empathy, the number of challenging items was insufficient and the information provided by the instrument on the children with the highest levels of empathy was less than desirable.

The OECD, guided by the TAG recruited Janet Strayer (1993[8]) and William Roberts (2018[9]), authors of the Empathy Continuum Measure for Young Children, prepared a revised empathy research model for young children, based on Strayer's original work outlining six vignettes proposed as suitable for use in IELS. These included practice vignettes, and a description of the scoring system and suggestions for how the material might be adapted for use on tablets. ACER assessment developers collaborated with the TAG, Strayer and Roberts to further revise the vignettes, in particular to greatly reduce the time required for administration by paring down the stimulus, especially in the use of illustrations to support meaning, reducing the wordiness and number of items and develop some new vignettes, applying the same measurement principles that underpin all the IELS items.

The OECD guided by the TAG also identified the Trust Direct Assessment Scale developed by Lucy Betts as a potential source of IELS items. Betts provided a set of 48 trust items from her shortened scale.

Roberts and Strayer also prepared a report for the TAG, '*Empathy and Trust: Scales for teachers, parents and administrators*' outlining an indirect measurement approach to

empathy and trust. An indirect measure of trust was adapted from the model proposed by Roberts and Strayer in collaboration with ACER survey and questionnaire development experts (Roberts, 2018[9]).

It was decided to develop and pilot a new set of empathy items for the main study and to measure trust indirectly via teachers and parents questionnaires.

The new empathy items consisted of eight 'vignettes' or scenarios where children were asked to choose a face indicating how the character felt, how they felt and then choose one of three reasons why they felt that way – 24 scored items in all.

2.7. Main study assessment design

As for the field test, the main study was conducted over two days. The direct assessments were administered October to December 2018 as seen in Table 2.6.

Table 2.6: Timetable for main study

Form	First session	Break	Second session
Assessment form 1	Emergent literacy	5 minutes	Empathy
Assessment form 2	Emergent numeracy	5 minutes	Self-regulation - inhibition Self-regulation – working memory Self-regulation – mental flexibility

As a result of the third phase of item development that ensured from the field test, as described in the above section, 125 items were taken to the main study. Table 2.7, Table 2.8, Table 2.9 and Table 2.10 present the number of items taken to the main study, broken down by the four domains and 13 sub-domain.

Table 2.7: Emergent literacy main study items

IELS aspect	Task	Items	Number of items
Listening comprehension	Story level aural meaning. Listen to an e-book story, answer audio supported items	Answer questions about: Prominent literal meaning, implied meaning, and information that needs to be linked	7
	Sentence-level aural meaning	Answer questions about: Prominent literal meaning, implied meaning, and information that needs to be linked	7
Phonological awareness	Initial sounds End sounds Rhyme Middle sounds	Identify first sound in words including blends Identify last sound Identify rhyme Identify middle sounds	3
Vocabulary	Identify a synonym for a word	Hear a word in a sentence and identify a synonym for a word from a list	5

Table 2.8: Emergent numeracy main study items

Content component	Description	Number of items
Numbers and counting	Identify digits and numbers to 20	1
	Count from 1 to 20	4
	Compare numbers and position (e.g. smaller, more, first etc.)	1
Working with numbers	Add and subtract in informal number story contexts	4
	Solve problems including doubling, halving, informal multiplication (repeated addition) and sharing	3
Measurement	Use everyday language to compare measures (e.g. longer, heavier, more etc.)	2
Shape and space	Use names of common shapes and objects and related language (e.g. sides, same shape, etc.)	2
	Use language of location (e.g. above, between etc.)	2
Pattern	Recognise and create patterns of shapes, objects and numbers	3

Table 2.9: Self-regulation main study items

Content component	Description	Number of tasks	Number of items
Working memory	Simultaneous storage and manipulation of information	1	22
Mental flexibility*	Switch between different concepts/rules and think about multiple concepts/rules simultaneously	1	35
Inhibition*	Override habitual/dominant behavioural response	1	12

* Reaction times in milliseconds are also measured. This was used in the scoring algorithm to differentiate fast response times.

Table 2.10: Empathy main study items

Content component*	Task description	Item description	Number of tasks	Number of items
Emotional identification	Listen to a short story, look at the pictures, and answer audio and picture supported items	Answer each question about the feelings of the story characters by choosing one of the four emoticon faces (i.e. happy, sad, surprise, and angry)	4	8
Emotion attribution	Listen to a short story, look at the pictures, and answer audio and picture supported items	Answer each question about their own feelings in the same situation by choosing one of the same four emotion faces	4	8

*Each one of the empathy tasks includes both components of emotional identification and emotional attribution. Hence, unlike Table 2.9, the 'Number of tasks column' should not be read accumulatively

.

Chapter 3. The delivery platforms

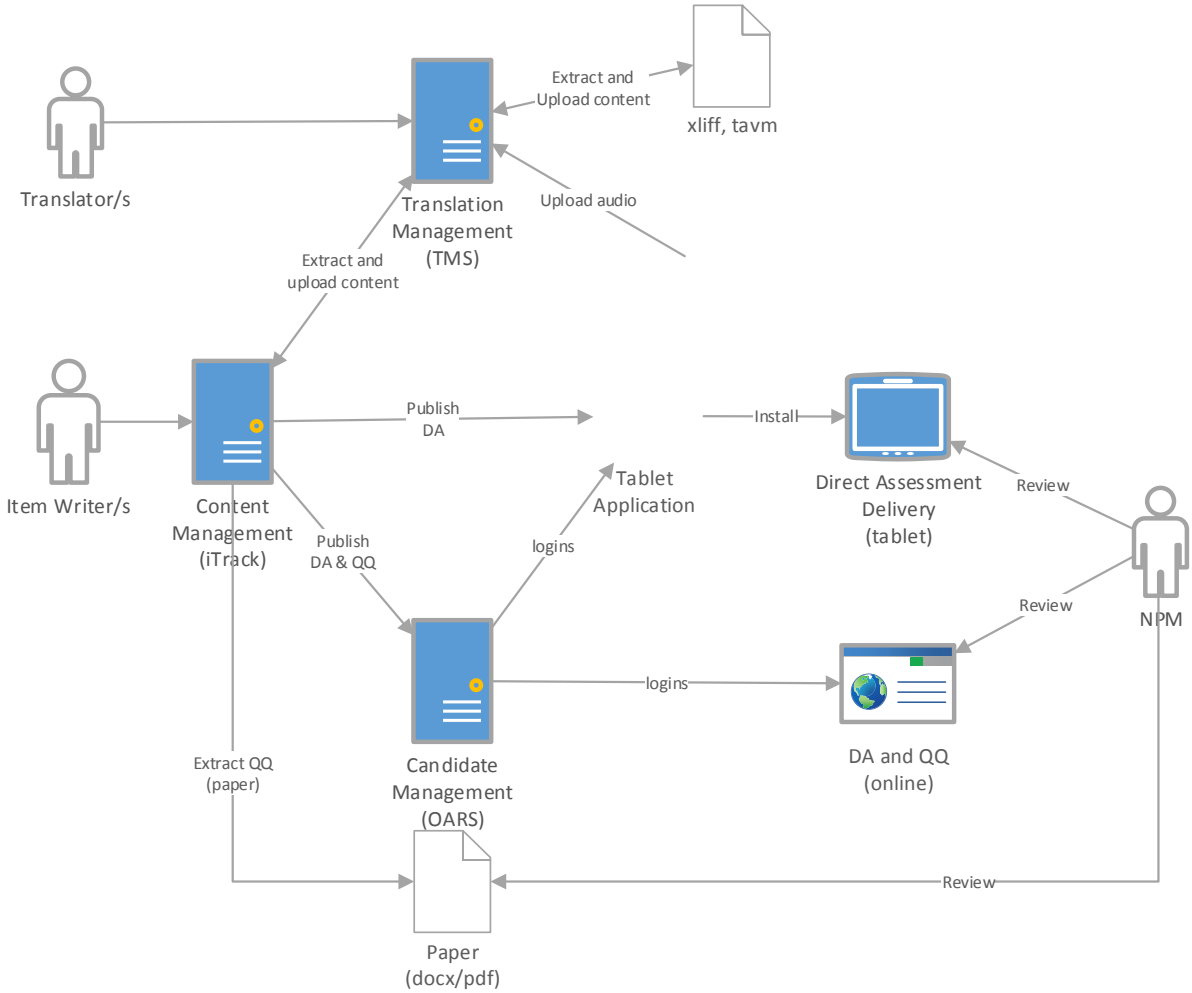
This chapter provides an overview of the delivery platforms used in the main study and associated solutions employed during the study. The platforms were designed as a series of interrelated but distinct solutions allowing for a more integrated means to deliver the various aspects of the IELS assessment while maintaining data security and privacy of participants.

Delivery platforms and supporting solutions:

- Content Management
- Translation Management
- Assessment Delivery
- Questionnaire Delivery
- Candidate Management.

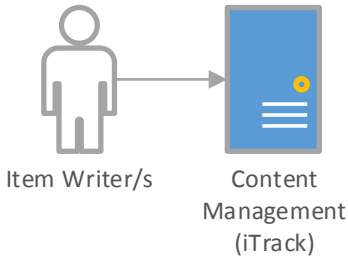
The following diagram shows the schema for the solutions for all the content development, translation and delivery solutions.

Figure 3.1: Solutions supporting content development, translation and verification



3.1. Content management

Figure 3.2: Content management solution



ACER's web-based content management solution 'iTrack', was used by ACER personnel to author assessment and questionnaire item content, and manage its use between country-specific assessments and questionnaires.

iTrack supports collaboration between item writers, assessment developers, psychometricians and project personnel during content development and review. For example, multi-author access to an item, tracking and visibility of changes made during development, and collection of comments made during the panelling process as outlined in Chapter 2. .

iTrack also supports export or publishing of content to various systems, including the Translation Management Solution (with audio text), the assessment and questionnaire delivery solutions, and to MS Word, MS Excel and Adobe PDF file formats.

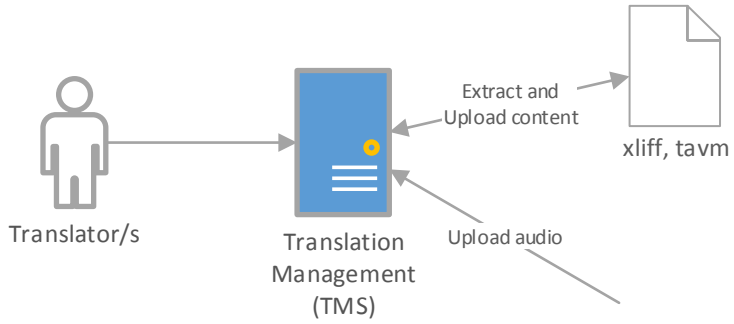
With clearly defined user permissions, auditable version control and support of business processes, iTrack provides consistent management of assessment content and associated metadata.

3.1.1. Key features:

- **Item types:** Stimulus, simple multiple choice, free form, short response items, re-usable templates, and complex dynamic items interactive on-screen elements.
- **Metadata:** Standard and customisable metadata. For example, curriculum classification, item status.
- **Static content:** Text, tables, images, and equations.
- **Interactive content:** Audio (including transcripts), video, drag-and-drop, highlighted and stackable content.
- **Configuration:** Navigation, straightforward linear flow, randomised clusters, trial clusters, opt in/out rules, branching and adaptive algorithms.
- **Scoring definition:** Item choice grouping, reasoning and credits.
- **Version control:** Auditable, with option to restore previous versions. Every change is recorded – by user, item and assessment.
- **Interoperability:** Content can be used in other systems. For example, translation management, marking, online assessment and reporting, and external review systems.

3.2. Translation management

Figure 3.3: Translation management solution



ACER's web-based Translation Management Solution (TMS), was used to manage translation and adaptation of source assessment content, for participating countries.

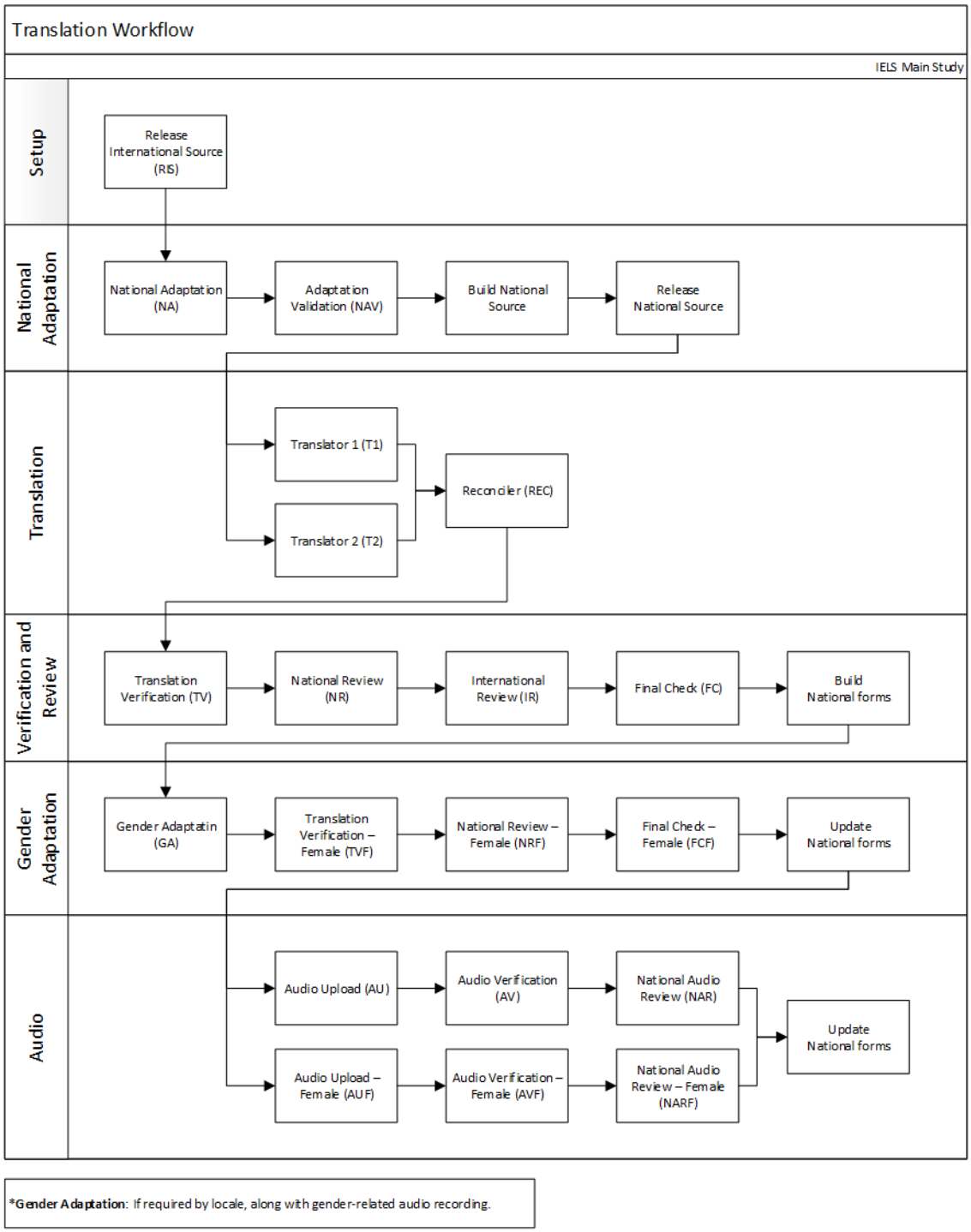
The TMS supports collaboration during translation and review of assessment content, locally and across an international user base. For example, tracking and visibility of changes made at each stage and collection of comments made during the translation process.

With flexible workflows to facilitate translation and review of content, and integration with other systems to meet translation-related quality assurance measures, the TMS provides a central solution for managing translation activities.

Once country-specific content was translated, it was uploaded to the content management solution, iTrack, and country-specific forms were created for delivery in those jurisdictions.

The workflow is outlined in the diagram below:

Figure 3.4: Solutions supporting participant access and data provision

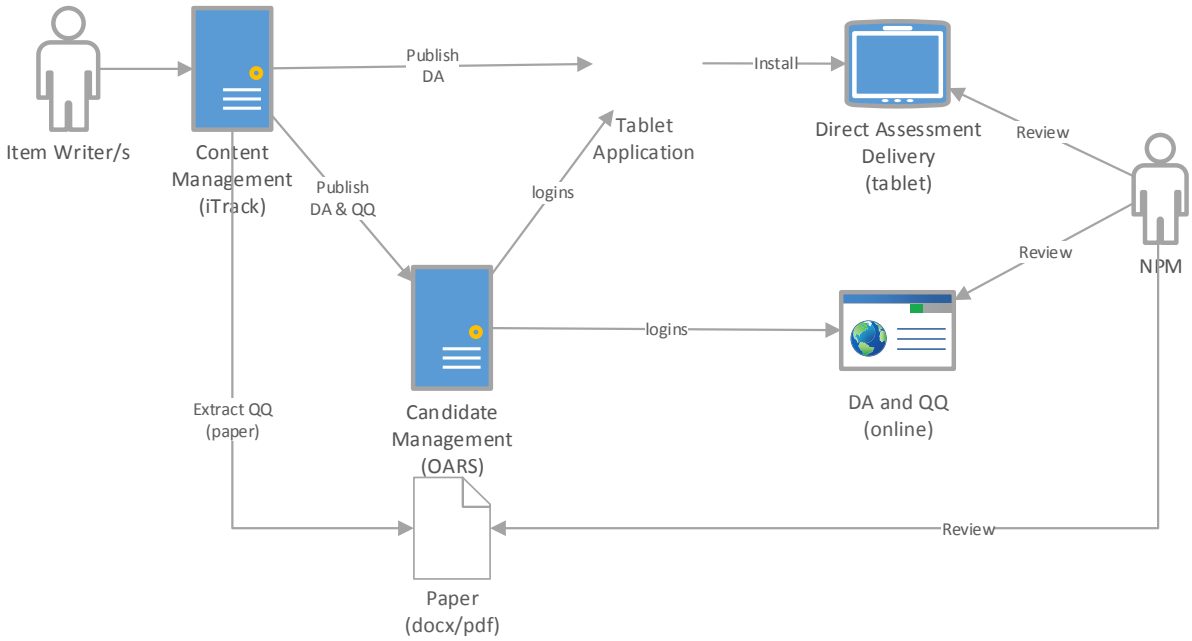


3.2.1. Key features:

- **Editing:** Support for all item types, including interactive items; international source content displayed in plain text, including audio content; support for translation guidelines per group of items; translation to locale content in plain text; mark-up support; collection of comments per item; static preview of source and locale content; on-screen comparison of source and locale content; and ability to upload supporting documentation, for example Translation, Adaptation and Verification Monitoring (TAVM).
- **Languages:** Support translation of participating country languages. Including, for example, character sets, fonts and text directions.
- **Workflow:** Configurable workflows by project, locale, adaptation and domain; configurable user roles and permissions; ability to approve (progress) or reject changes made in previous step; ability to restore from previous step or fast-track workflow; and collection and display of comments made at each step.
- **Interoperability:** Assessment content can be used in other systems. For example, content management, and systems that support localisation interchange file format (XLIFF).

3.3. Assessment and questionnaire delivery

Figure 3.5: Assessment delivery solution



3.4. Direct assessment

ACER's Online Assessment and Reporting System (OARS) Delivery Tablet Application (the application) was used to deliver IELS direct assessments to an internationally distributed user base.

Due to inconsistent Internet connectivity across participating countries and their centres, the application was developed to deliver assessments offline. Internet connectivity was only required to install the application and upload data for completed sittings, which could occur when consistent Internet connectivity was available.

The application was designed to be installed on compatible devices. To ensure consistency of behaviour, tablet specifications and recommended devices were identified.

Several iterations of the applications were developed, in response to feedback from the consortium and participating countries (for example, to improve audio playback). During development, assessments were also made available online, to the consortium and countries for content verification and familiarisation.

Assessment content developed using the content management solution, iTrack, was embedded in the application. Each application was country-specific and contained direct assessments for each domain, in each language used in the jurisdiction for the assessment.

Initially the application contained international source content, and was distributed to participating countries for system testing and familiarisation. When preparation of

national assessments was complete, country-specific versions of the application were released.

Assessment data from the tablets was collected in OARS in preparation for data analysis.

OARS also provided on-demand participation reports for the direct assessments, by country and language.

3.4.1. Key features of OARS, assessment delivery:

- Assign assessments in the appropriate language to participants
- Remove assessments assigned to participants in error
- Offline delivery of assessments
- Consistent assessment configuration regarding visual layout, user interface (UI) elements, navigation and flow
- Display stimuli and items, and capture of participant responses
- Allow participants to resume incomplete assessments
- Manual upload of data for started and completed sittings.

3.5. Questionnaire

Questionnaire content developed using the content management solution, was imported to OARS and made available to participants via the Candidate Management solution.

ACER's Online Assessment and Reporting System (OARS) web application was used to facilitate delivery of England and Estonia online questionnaires to participating countries.

OARS is designed to provide a seamless assessment experience across various devices and network environments. OARS supports completion of questionnaires online within web browsers on desktop and on tablet devices.

Reponses for questionnaires completed online were collected in OARS in preparation for data analysis. OARS also provided on-demand participation reports for the online questionnaires, by country and language.

3.5.1. Key features of OARS questionnaire delivery:

- Online delivery of instruments
- Consistent assessment configuration regarding visual layout, UI elements, navigation and flow
- Display stimuli and items, and capture of participant responses
- Allow participants to restart or resume incomplete instruments
- Instant collection of data for started and completed sittings.

United States questionnaire delivery

In accordance with data protection requirements, United States questionnaires were delivered and response data managed using solutions provided by the United States National Centre.

Paper-based delivery

Following validation of content online using OARS, country-specific questionnaire content was exported from iTrack to Microsoft Word. Documents were provided to countries for final adjustment and distribution to offline questionnaire participants.

3.6. Candidate management

The candidate management feature within OARS was used to manage participant access and generate on-demand participation reports.

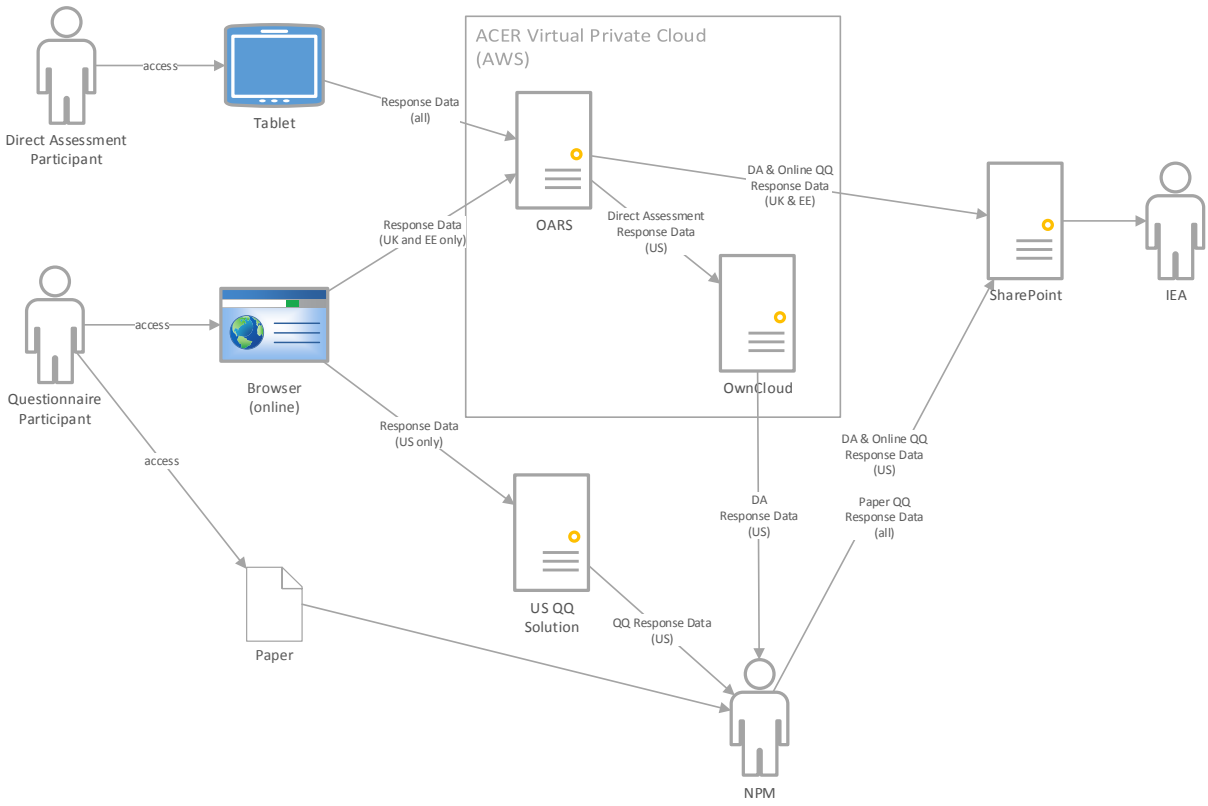
3.6.1. Key features of OARS, candidate management:

- Create participant logins for the direct assessments delivered offline
- Create participant logins for the questionnaires delivered online
- Arrange anonymised participant records by country and centre
- Allocation of online questionnaires to participants
- Issue login credentials to allow access to direct assessments and online questionnaires
- Re-allocation or restart of a participants online questionnaire
- Create administration-level logins for NPMs to oversee country participation
- View completion status of participants direct assessments or online questionnaire
- Generate on-demand participation reports for direct assessments and online questionnaires.

3.7. Data security

Data security solutions were implemented in accordance with response data protection requirements.

Figure 3.6: Solutions supporting participant access and data provision



3.7.1. England and Estonia response data

A separate version of OARS was set up in Dublin for delivery and collection of questionnaire response data, and collection of direct assessment response data from England and Estonia:

- ACER provided England and Estonia assessment and questionnaire response data from OARS directly to IEA Hamburg for processing;
- Single export of data to country-specific csv files containing response data for the entire study;
- Files were encrypted and provided via SharePoint, with appropriate access restrictions in place.

United States response data

In accordance with data protection requirements, ACER and the United States NPM implemented a secure solution for managing direct assessment data. A separate instance of OARS (OARS-US) was set up in the US for collection of assessment response data, with appropriate access restrictions in place:

- ACER provided response data from OARS-US to the NPM for disclosure risk analysis
- Nightly automated export of csv files containing response data for the previous day, with earlier exports retained
- Encrypted and signed files provided via OwnCloud, with appropriate access restrictions in place
- The National Project Manager provided the United States questionnaire and direct assessment response data direct to IEA Hamburg for processing.

Chapter 4. Contextual questionnaire development

This chapter describes the development of the three contextual questionnaires designed for both parents and staff of the target population for IELS.

The parent questionnaire was designed to capture information about the child's personal and home background as well as parent reports of their child's skills and behaviours. It included questions on:

- Child's background characteristics (including gender, age)
- Parent estimates of child's capacity in the assessment language for literacy and numeracy skills
- Parent estimates of how their child is developing in social skills, emotional skill`s, trust, empathy, self-regulation, gross and fine motor skills, expressive and receptive language skills, and numeracy skills
- Parent reports on issues or difficulties that the child has experienced
- Indirect assessment of prosocial behaviour, trust and non-disruptive behaviour
- The child's previous early childhood education and care attendance and arrangements
- Activities undertaken with the child
- Home background characteristics (including the number of people in the household, language and country background, parental education, parental occupation, household income)

Each staff member from an early childhood education setting or school who was nominated as knowing each of the target children best from the setting was asked to complete two sections of a questionnaire. The first section contained questions designed to capture characteristics of the staff members themselves including:

- Staff background characteristics (including age and gender)
- Staff education and specializations
- Staff years of experience and current employment status.

The second section contained questions about the assessed child. Staff were asked to complete this for each assessed child assigned to them (to a maximum of 15 children). The questions in this section included:

- The length of time they have known the target child
- Perceptions of parent involvement in activities undertaken at school
- Knowledge of additional support provided to the child
- Staff estimates of child's capacity in the assessment language for literacy and numeracy skills (equivalent to question in the parent questionnaire)
- Staff estimates of how the child is developing in social skills, emotional skills, trust, empathy, self-regulation, gross and fine motor skills, expressive and receptive language skills, and numeracy skills (equivalent to question in the parent questionnaire)

- Indirect assessment of prosocial behaviour, trust and non-disruptive behaviour (equivalent to question in the parent questionnaire)

This chapter will provide a brief overview of the two questionnaires and detail their development. Information on the delivery of the computer-based versions of these questionnaires can be found in Chapter 3. .Chapter 15. describes the procedures for index construction and scaling of context questionnaire data.

4.1. Contextual framework for IELS

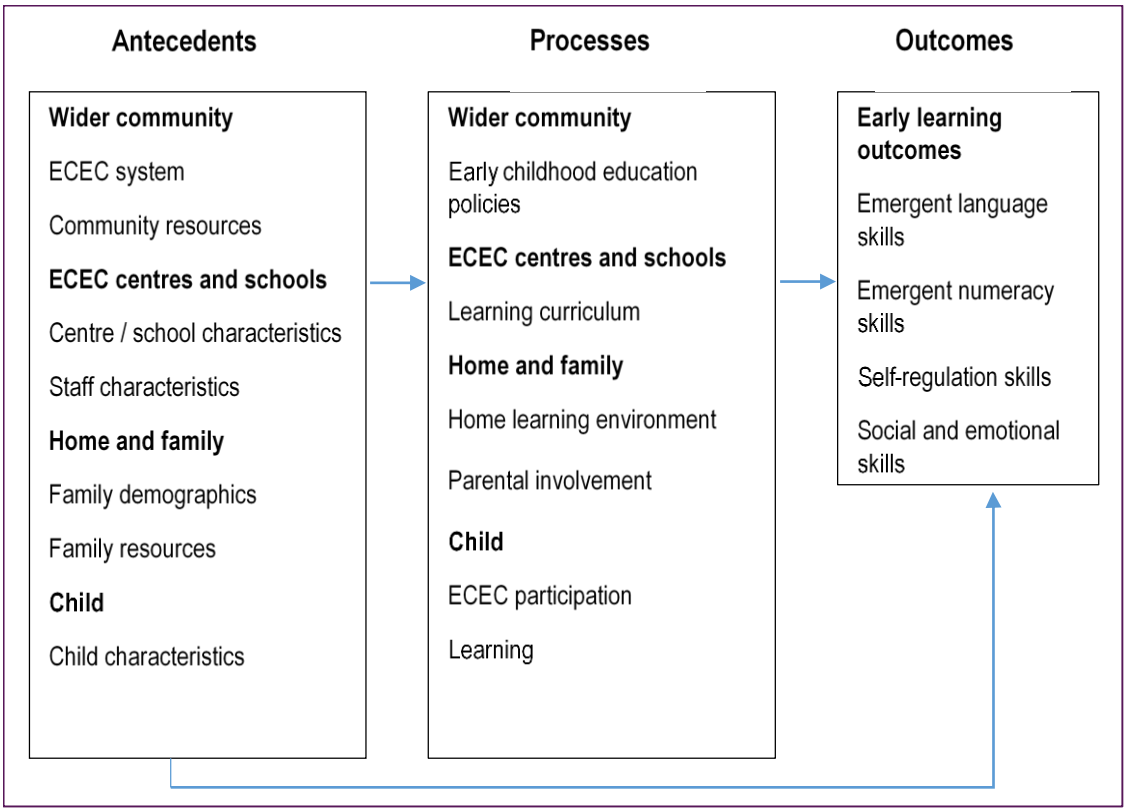
The contextual questionnaires used in IELS were developed using the assessment framework to provide a conceptual underpinning the contextual factors that are believed to influence study outcomes (OECD, 2021[5]). The assessment framework is organised according to the domains assessed (both directly and indirectly) in IELS. The questionnaires contain some content related to the direct assessment chapters (for instance items related to the child's development and skills in areas related to emergent literacy, emergent numeracy, self-regulation, social and emotional competence (including empathy, trust and prosocial behaviours)).

The contextual framework identified the context variables that reflect the environment in which early learning takes place. It assumes that the home environment for young children has a deep influence on the development of their cognitive, social and emotional skills. The framework suggests that contextual influences on early learning outcomes are conceived as either antecedents or processes. Antecedents refer to general background variables that affect learning outcomes (e.g. through context factors such as socio-economic status of the household). Process-related variables are those factors shaping early learning outcomes more directly (e.g. the home learning environment). The model used in IELS suggests that:

- both antecedents and processes can influence outcomes directly
- antecedents can influence processes and thus have a transmitted influence on outcomes
- processes can mediate (explain the relationship between an antecedent and an outcome) and moderate (affect the strength of the relationship between an antecedent and an outcome) the influence of antecedents on outcomes.

Figure 4.1 displays the conceptualisation for IELS of the influence of antecedents on both process variables and early learning outcomes in addition to the influence that process-related variables have on those outcomes.

Figure 4.1: Contexts for developing children's early learning



4.2. Contextual item development (parent and staff questionnaires)

The development of the contextual questionnaires used in IELS was co-ordinated by ACER in liaison with the OECD, consortium members, participating countries, the Technical Advisory Group (TAG), Early Learning Group (ELG) (which included the OECD, participating countries, interested countries, technical representation from the consortium) and external consultants. The development work included extensive reviews and discussions at different stages of the process with different project stakeholders including the ELG and TAG.

4.2.1. The development process involved three phases:

Phase 1: This phase included the initial development of field test material guided by the assessment framework and involved the item trial. The questionnaire item trial comprised of the trialling of the entire questionnaire for both parents and staff as well as qualitative feedback obtained through focus groups.

Phase 2: This phase involved the international field test, conducted in all three participating IELS countries, and subsequent analyses of field test data to inform judgements about the suitability of questionnaire material for the main study.

Phase 3: During this final phase, ACER discussed the field test results with the OECD, ELG and TAG. Phase three concluded with a final selection of main study items.

4.3. Initial development

The parent questionnaire as used in the item trial included items about the child (Section A), the child's activities outside of the home (Section B), the child's activities in the home (Section C), and their family composition and characteristics (Section D). Questions were either newly developed for the survey (e.g. those related to early childhood education attendance and care), adapted from existing surveys of social-emotional development (e.g. the Adaptive Social Behaviour Inventory (ASBI)) (Hogan, 1992[10]), or adapted from other large-scale educational surveys such as OECD Programme for International Student Assessment (PISA) (including questions related to home background).

In total, the version of the questionnaire for the item trial consisted of 27 questions with 133 items. Based on the number of items, it was anticipated that this version of the questionnaire would take approximately 30 minutes to complete.

The staff questionnaire comprised two sections. The first section contained questions about the staff's own background and experience working in ECEC (Section A), whereas the second section contained questions about the children in their classroom who were participating in the study (Section B). The questionnaire included 12 questions and 69 items.

Staff participating in the item trial were asked to complete both sections (for the purpose of the item trial, they only had to complete section B once based on their experiences with one child). Based on the number of items, it was anticipated that this version of the questionnaire would take approximately three minutes for section A, and approximately 10 minutes for section B.

The purpose of the item trial was to trial the entire questionnaire for both parents and staff, with a focus in particular on qualitative feedback. The limited quantitative data assisted with the revision process, but, more importantly, the qualitative feedback collected from respondents provided valuable insights into whether the questions functioned as intended.

Countries were given some item trial guidelines outlining a suggested approach for conducting the focus groups, including what materials to use, suggestions on how to recruit participants, plans and preparations, administration and how to capture and submit feedback. Countries were also asked to adapt the questionnaires in the questionnaire adaptation spreadsheet. Details on how to complete this spreadsheet were also provided in the guidelines.

The item trial yielded valuable qualitative information on how the two questionnaires could be improved. Many of the suggestions relating to question instructions and wording were incorporated into the international field test source versions. In addition, the exercise helped countries and the consortium identify issues that would impact the national versions of each questionnaire (such as those relating to adaptations, translations and national questions).

A translatability report on the item trial versions of the questionnaires was also supplied to the questionnaire development team at ACER by cApStAn. Although this was not technically part of the item trial, the feedback on question wording was incorporated into revisions of the questionnaires.

4.4. Field test

The final parent questionnaire used in the field test consisted of 33 questions containing 164 items, to be completed within the targeted time of 40 minutes (based on completion times of the questionnaire in the item trial). The field test parent questionnaire consisted of the following four sections:

***About your child*:** Including child characteristics, child's capacity in literacy and numeracy skills, digital device experience, social, emotional, cognitive and motor skills, child issues or difficulties and an indirect assessment of prosocial behaviour and disruptive behaviour

Activities outside the home: Including neighbourhood resources, activities outside of school, extra language tuition, ECEC attendance, informal care providers

***Activities in the home*:** Including home learning environment, home resources (such as electronic devices, Internet access, number of children's books in the home)

***About you and your family*:** Including characteristics of the child's parents, people living in the home, and other home background details.

The final staff questionnaire consisted of two sections. Section A consisted of six questions containing 24 items and Section B consisted of six questions containing 62 items to be completed within the targeted time of six minutes for section A and 15 minutes for section B for each child.

4.5. Main study

Consortium staff undertook a full psychometric analysis of all questionnaire data. This included the following steps:

- Analysis of response frequencies and missing data
- Exploratory and confirmatory factor analysis
- Correlations and triangulations between indirect and direct outcomes measures.

Following the analyses of field test data and in consultation with the OECD and Technical Advisory Group, refinement of the questionnaires resulted in the removal of 5 questions from the parent questionnaire and no removal of questions from the staff questionnaire. These omissions were made in order to reduce the overall length of the questionnaires and allow for new content. Some changes were made to reflect modifications to the assessment framework, while others were affected based on advice from project stakeholders over the relative importance of the content. The psychometric performance of questions (including lack of response variance, poor association with indirect and direct outcomes variables and poor fitting confirmatory models) had a large role in deciding which items to remove from the field test.

The questions removed from the parent questionnaire included content related to neighbourhood resources available to the child, activities undertaken by the child outside of school (although the main study parent questionnaire did ask about the frequency with which children were brought to special or extra-cost activities outside of the home, such as sport, dance, etc.), whether they received extra language tuition, resources available in the home (including ICT devices and software, Internet access, reading material for older children or adults).

Major refinement of the following questions in the parent questionnaire following consultation with the OECD and TAG were:

- Question on the child's capacity for literacy and numeracy skills (modifications made to include more complex activities to address high proportion of respondents indicating that the child was capable).
- Indirect assessment of child's prosocial behaviour and non-disruptive behaviour (confirmatory models supported a satisfactory factor structure similar to the ASBI (Hogan, 1992[10]). It was deemed necessary to add additional items related to trust at expense of several existing items that were excluded based on conceptual and psychometric grounds.
- Question on activities at home designed to capture information about the child's home learning environment (originally adapted from Niklas et al. (2016[11]) were modified in consultation from the TAG to align the items more closely with other studies of early years leaning and development including the Effective Provision of Pre-School Education Project (Melhuish, 2008[12]).
- Early childhood education and care setting attendance (in order to reduce the burden on respondents, this series of questions was largely redesigned for the main study while still collecting information on type of attendance, age of attendance and intensity of attendance).

Corresponding changes made to questions on the child's capacity in literacy and numeracy and indirect assessment of prosocial behaviour, non-disruptive behaviour and trust were also made in the staff questionnaire. The final parent questionnaire consisted of 24 questions containing 177 items, to be completed within the targeted time of 40 minutes. The final staff questionnaire consisted of two sections. Section A consisted of six questions containing 24 items and Section B consisted of six questions containing 62 items to be completed within the targeted time of five minutes for section A and 15 minutes for section B for each child.

4.6. Adaptation and final checks

After the release of the source version of the questionnaires, participating countries were asked to make a series of adaptations (both structural and non-structural) in the Translation Adaptation Verification Monitoring (TAVM) workbooks. Non-structural adaptations in the questionnaires included:

- Unit of measurement for the child's birth weight
- Appropriate term(s) for the centre/school where the assessment took place
- Country of the assessment
- An appropriate term for the parent/guardian
- Nationally appropriate terms for supervision and care types and early childhood education and care settings
- Nationally appropriate terms for International Standard Classification of Education (ISCED) levels.

4.6.1. Structural adaptations in the questionnaires included:

- National extension questions (e.g. ethnicity)

- Necessary addition or subtraction of appropriate terms for examples of ECEC settings and ISCED levels to suit national context.

Countries were provided with detailed notes on required adaptations, and all proposed national adaptations (both non-structural and structural) were reviewed by consortium members and negotiated with countries (if necessary). Some common issues identified during the adaptation review were:

- Difficulties in establishing appropriate adaptations for questions related to ECEC settings and ISCED levels
- Structural changes that would make comparison with other countries impossible
- Inconsistent adaptation of terms within an instrument.

Once the adaptation and translation process had been finalised (and countries had implemented any changes following translation verification), the final national questionnaires were built. The questionnaires were first built in an online format. They were then converted to a paper-based format by consortium staff members, and checks were undertaken to ensure that the content across both modes was comparable (where applicable). Layout or formatting issues that were identified in both the online delivery system and in the paper-based format were fixed. In some instances, errors were identified in translated terms, or it was apparent that previous feedback from adaptation or translation verification had not been implemented. In these instances, the consortium informed the country of necessary changes to be made. Countries had an opportunity to review the final instruments, and notify the consortium of any changes that needed to be made, before administering the instruments. All these measures were implemented to assure comparability, reliability and high quality of the collected questionnaire data. The steps described have been implemented following the quality control framework of IELS.

Chapter 5. Sample design

This chapter details the sampling design applied in IELS. First, the definition of the target population of children is explained. Next, the terms "centres/schools", "parents" and "staff members" are clarified. Information about the country-specific characteristics of the target population is given, including information on parts of the target population that could not be covered. The rationale behind the determination of sample sizes is provided, and an overview of the country-specific sample sizes is given. Lastly, the process of selecting centres/schools and children, using a stratified two-stage probability design, is detailed.

5.1. Overview of the sampling design

IELS aimed to study children at the age of five. Accordingly, children who had experienced their fifth birthday, but not yet their sixth birthday, at the time of the assessment were eligible for the study. The target population was limited to children registered in early education and care centres or schools.

IELS employed a stratified two-stage probability sampling design. At the first stage, centres/schools were randomly sampled from a list of centres/schools that were expected to provide education and care for children of the target age. The probability of a centre/school being sampled was proportional to the expected number of eligible children in the centre/school. In the second stage, children were randomly selected from lists of all eligible children within the sampled centres/schools.

5.2. Target population

5.2.1. Definitions of centres/schools, children, parents and staff

In order to allow National Study Centres to identify the correct units for sampling and study administration, it was crucial to develop unambiguous definitions for all sampling-related terms. Centres and schools were defined as institutional (officially registered) settings providing education and care for children at the age of five. In order to be classified as a "centre" or "school" for IELS, settings needed to provide education or care for at least two hours per day and 100 days a year. Both early childhood education and care centres and schools needed to be included in IELS, if they accommodated children of the target age group. In practical terms, however, in Estonia all children at the age of five were found in centres, while in England and the United States they were in schools.

Informal arrangements to look after children were not in scope of IELS, nor were those children under the care of helpers employed in the household where the child lives.

The aim of IELS was to cover all children in centres/schools at the age of 5 at the time of the assessment. The definition had to be operationalised considering that:

- the assessment was administered on two different days for each child

- for each sampled centre/school, multiple days were needed to assess all sampled children
- the assessment period for a specific centre/school was usually unknown at the time of listing and sampling the children.

National Study Centres were instructed to determine the day in the middle of their administration period and to define all children as eligible who were at least five years, but not six years old on that day, resulting in a birth date range for each participating country. To facilitate this definition, the participating countries were allowed to round the birth date range to the nearest whole month, and all countries participating in IELS chose to do so.

Although participating countries could choose their administration period within the international administration period of October to December 2018, the administration periods of all three participating countries were close to each other, using almost the whole period between October and December 2018, and the resulting birth date ranges were November 2012 to October 2013 for all three participating countries.

For each sampled child, the parents or primary caregivers were asked to complete a parent questionnaire. Additionally, the staff member at the centre/school who knew the child best or who was the first contact person was asked to complete a questionnaire about each sampled child. If a staff member was unable or did not wish to participate, they could be replaced by another person who could provide information about the child.

5.3. Characteristics of the target population by country

In England, schooling starts with reception classes before continuing with year 1. Usually, children who are five years old before 1 September enter year 1 on that day. Therefore, children who were born between November 2012 and August 2013 were found in year 1 of schools; children who were born in September or October 2013 were found in reception classes of schools.

In Estonia, children enter school if they have reached the age of seven years before 1 October of the respective year. Attending early education and care centres is not compulsory, but very common, with enrolment rates above 90%. Therefore, the target population consists of five-year-old children in pre-school institutions.

In the United States, there are differences in enrolment options between the states. Often, children enter kindergarten class within a school if they are five years old by the start of the school year. In this case, children born between November 2012 and August 2013 were attending kindergarten classes, while children born in September or October 2013 were most likely to be in pre-kindergarten classes.

5.4. Exclusions and the national study population

Due to specific circumstances in the participating countries, it was not feasible to access all eligible children. For example, it might be difficult to access specific centres/schools. Similarly, in some circumstances the instruments would not have been appropriate to the specific needs of some children without major customisation (e.g. translation). Therefore, national centres had the opportunity to reduce the coverage to a specified part of the target population, thereby creating a national study population. The technical standards required that no more than 5% of children in the target population be excluded.

Within sampled centres/schools, all 5 year-old children were listed and eligible for sampling. However, it was recognised that some children might not be able to fulfil the basic requirements of the study due to severe special education needs or very limited experience in the language of administration. In these cases, sampled children were given the participation status "Did not participate due to special education needs". These children were counted as within-sample exclusions (refer to Table 13.7).

Table 5.1, Table 5.2 and Table 5.3 provide an overview of the target populations, study populations and exclusions at centre/school level prior to the centre/school sample selection.

Table 5.1: Target and study population in England

	Number of schools	Number of children	Percentage of children
Target population	17 730	695 173	100.0%
Exclusions	1 114	6 722	1.0%
Very small schools	*482*	*1 605*	*0.2%*
Special needs schools	*632*	*5 117*	*0.7%*
Study population	16 616	688 451	99.0%

Note: In England, 'very small schools' were defined as those with fewer than six enrolled children eligible to participate in IELS.

Table 5.2: Target and study population in Estonia

	Number of centres	Number of children	Percentage of children
Target population	625	13 070	100.0%
Exclusions	6	55	0.4%
Special needs centres	*3*	*32*	*0.2%*
English speaking centres	*3*	*23*	*0.2%*
Study population	619	13 015	99.6%

Table 5.3: Target and study population in the United States

	Number of schools	Number of children	Percentage of children
Target population	78 315	3 400 000	100.0%
Exclusions	13 346	200 000	5.9%
Alaska, Hawaii, Puerto Rico	*1 590*	*53 700*	*1.6%*
Overseas schools organised by the Department of Defense	*58*	*4 000*	*0.1%*
Centres/schools without kindergarten class[2]	*1 882*	*130 000*	*3.8%*
Very small schools	*9 816*	*12 300*	*0.4%*
Study population	64 969	3 200 000	94.1%

Note: In the United States, 'very small schools' were defined as those with fewer than five enrolled children eligible to participate in IELS.

5.5. Required sample sizes

The IELS sample design is one where sampled children are clustered within centres or schools. It can be expected that children in the same centre/school are likely to share characteristics as they share the same environment and staff, but also due to effects of self-selection and interaction. This is called the "clustering effect". Together with other characteristics of the sampling design, such as stratification or unequal selection probabilities, a "design effect" can be estimated. The design effect estimates the ratio of the sampling variance within the given sample design to the sampling variance expected from a simple random sample (Kish, 1965[13]).

Those involved in the planning of IELS initially assumed a design effect of 4, although it was expected that the design effect would vary for different participating countries and different variables. This assumption was based on information from previous studies with similar designs and target populations.

It was desired that the estimates based on the IELS data should be as precise as a simple random sample of 400 assessed children per participating country, which is called an "effective sample size". Therefore, with the assumed design effect of 4 and assumed response rates of 75% at centre/school and child level, the required sample size was derived as shown in Table 5.4.

Table 5.4: Deriving sample sizes for IELS

Effective sample size	a	400
Design effect	b	4
Actual sample size	c = a x b	1 600
Children per centre/school	d	15
Minimum child response	e	75%
Minimum response per centre/school	f = d x e	11
Number of centres/schools required	g = c / f	145
Minimum centre/school response	h	75%
Sample size for centres/schools	i = g / h	200

The sample sizes were set to a minimum of 200 centres/schools. In each sampled centre/school, at least 15 children needed to be sampled. In small centres/schools with fewer than 15 children, all children were selected.

The number of sampled children was set to 3 000 (derived by calculating 200 centres/schools x 15 children), without considering non-response. Therefore, if there were centres/schools with fewer than 15 children, a higher number of centres/schools needed to be selected to maintain the overall number of 3 000 children.

Participating countries were free to increase their sample sizes by selecting more centres/schools, or more children within the sampled centres/schools.

5.6. Selecting centres and schools

5.6.1. Centre/school sampling frames

It was each National Study Centre's responsibility to provide a list of all eligible centres/schools according to the definitions of IELS. This list had to be current and

complete. Variables that needed to be included in the centre/school sampling frame were:

- a national centre/school identifier
- variables that could be used for stratification
- a measure of size (MOS).

Ideally, the measure of size was the number of 5 year-old children in the centre/school. If this information was unavailable, a different indicator such as the size of an equivalent grade was used.

Some key information about the provided centre/school sampling frames is given in Table 5.5.

Table 5.5: Information about sampling frames

Country	Year	Data source	Measure of size
England	2016/17	Department of Education	5 year-old-children
Estonia	2016/17	Estonian Educational Information System	5 year-old-children
United States	2015/16	US National Centre for Education Statistics (NCES)'s Common Core of Data NCES's Private School Universe Survey File	Children in kindergarten class

5.7. Stratification

Stratification entails the grouping of units on the sampling frame by certain characteristics. This is generally done to improve the efficiency of the sampling design.

Stratification can be done explicitly, or implicitly. For explicit stratification, the list of centres/schools is divided by the categories of the stratification variable. Then for each explicit stratum, the number of centres/schools to be sampled is assigned, and samples are selected independently for each explicit stratum. For implicit stratification, the units on the sampling frame are sorted by the stratification variable, thereby ensuring that each implicit stratum will be represented proportionally. Explicit and implicit stratification can be combined within a participating country by using some variables for explicit stratification, and other variables for implicit stratification within explicit strata.

IELS did not require the use of stratification; however, it was strongly recommended. The strategy for stratification was determined by participating countries with the help of the sampling team.

The variables used for stratification, together with their categories, are presented in Table 5.6.

Table 5.6: Stratification variables by country

Country	Explicit stratification	Implicit stratification
England	School type (maintained, academy, independent) Free school meal (FSM) eligibility (lowest 20% FSM, 2nd lowest 20% FSM, middle 20% FSM, 2nd highest 20% FSM, highest 20% FSM, unknown FSM)	Region (North, Midlands, Greater London, South)
Estonia	Language (Estonian, Russian) *For Estonian speaking centres:* Location (urban, rural)	*For Russian speaking centres:* Location (urban, rural)
United States	School type (public, private) *For public schools:* Free lunch (less than 60% free lunch, at least 60% free lunch, unknown free lunch) Region (Northeast, Midwest, South, West)	Location (city, suburb, town, rural) State (Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, Washington, West Virginia, Wisconsin, Wyoming, District of Columbia)

The number of centres/schools to sample was distributed between the explicit strata following a suitable allocation agreed by the participating country and the sampling team. Usually, the sample was allocated proportionally to the share of the target population in each explicit stratum. Alternatively, it was also possible that the sample could be allocated disproportionally to select more centres/schools in certain explicit strata and fewer centres/schools in other explicit strata without biasing the results when applying weights. In IELS, disproportional allocation was only used to ensure a sufficient number of centres/schools were sampled from each explicit stratum. Table 5.7 shows the sample allocation for each country.

Table 5.7: Sample allocation by country

Country	Sample allocation across explicit strata
England	Proportional to the number of children at the age of five
Estonia	Proportional to the number of children at the age of five
United States	Proportional to the number of children in kindergarten class, then adjusted to have a minimum of eight schools in the smaller explicit strata

5.8. Sample sizes by country

By using the agreed-upon sample allocation, the sample sizes for each participating country and explicit stratum were determined.

Table 5.8: Sample sizes by country and explicit strata

Country	Explicit stratum	Sample size
England	Maintained – lowest 20% FSM	26
	Maintained – 2nd lowest 20% FSM	30
	Maintained – middle 20% FSM	32
	Maintained – 2nd highest 20% FSM	32
	Maintained – highest 20% FSM	24
	Academy – lowest 20% FSM	8
	Academy – 2nd lowest 20% FSM	8
	Academy – middle 20% FSM	10
	Academy – 2nd highest 20% FSM	12
	Academy – highest 20% FSM	12
	Independent – unknown FSM	8
	Total	202
Estonia	Estonian – urban	110
	Estonian – rural	78
	Russian	32
	Total	220
United States	Public – less than 60% free lunch – Northeast	12
	Public – less than 60% free lunch – Midwest	20
	Public – less than 60% free lunch – South	22
	Public – less than 60% free lunch – West	18
	Public – at least 60% free lunch – Northeast	8
	Public – at least 60% free lunch – Midwest	10
	Public – at least 60% free lunch – South	28
	Public – at least 60% free lunch – West	20
	Public – unknown free lunch – Northeast	8
	Public – unknown free lunch – Midwest	8
	Public – unknown free lunch – South	8
	Public – unknown free lunch – West	8
	Private – Northeast	8
	Private – Midwest	8
	Private – South	8
	Private – West	8
	Total	202

5.9. Sample selection

All centre/school samples were selected by the sampling team at IEA Hamburg. The method used was systematic random sampling with probability proportional to size (PPS) within explicit strata.

The process of the sample selection is detailed below; a numerical example is included later in this section.

In a first step, the centres/schools were sorted by implicit stratification and measure of size within each explicit stratum. The sorting by measure of size was done in a serpentine manner, with the order alternating between increasing and decreasing magnitudes of the measure of size. This had the effect that two adjacent centres/schools

within the same explicit stratum were of similar size even if they were in two different implicit strata. The goal of this sorting routine is to make sure adjacent centres/schools in the frame are as similar to each other as possible. This procedure can potentially decrease sampling variance and allows effective replacement of centres/schools.

The next step of the centre/school sample selection process is to determine the preliminary sampling interval, I_g, for each explicit stratum g. It is calculated as the total measure of size, S_g, divided by the number of centres/schools to sample, D_g:

$$I_g = \frac{S_g}{D_g} \tag{5.1}$$

In case the share of centres/schools to sample, based on the number of existing centres/schools, is high, or some centres/schools are very large, it might be possible that the measure of size of some centres/schools exceeds the sampling interval I_g. When this happens, it is possible for a centre/school to be selected multiple times, which should be avoided. Therefore, these centres/schools were treated as certainty centres/schools, meaning they were set aside, i.e. added to the sample and removed from the sampling frame. The values for S_g, D_g and I_g were then recalculated. In case that, again, some centres/schools were larger than the resulting sampling interval, this process had to be reiterated until all remaining centres/schools had a measure of size smaller than the sampling interval.

IELS employed a so-called "self-weighting design", i.e. aiming for approximately similar selection probabilities and weights for all sampled children. This was achieved by selecting primary sampling units (centres/schools) with probabilities proportional to size, and secondary sampling units (children within centres/schools) with systematic random sampling with equal probabilities. For small centres/schools, i.e. those with fewer than 15 eligible children, a different approach led to this goal. All small centres/schools within an explicit stratum were selected with the same probabilities. All children within these centres/schools had a secondary selection probability of 1, leading (in an ideal scenario) to identical selection probabilities. To synchronise the selection probabilities with the large centres/schools in the explicit stratum, the measure of size was averaged for all small centres/schools by adding the individual measures of size for all centres/schools with fewer than 15 children and then dividing this number by the number of small centres/schools.

Next, one random number, RN_g, per explicit stratum was generated that was larger than 0 and less than or equal to 1. Selection numbers were calculated using the formula: $Z_{gj} = RN_g \times I_g + (j_g - 1) \times I_g$, with j_g denoting the centre/school to sample within the explicit stratum. This translates to $Z_{g1} = RN_g \times I_g$ for the first centre/school to sample, $Z_{g2} = RN_g \times I_g + I_g$ for the second centre/school to sample, $Z_{g3} = RN_g \times I_g + 2 \times I_g$ for the third centre/school to sample, etc.

For each centre/school in the sampling frame, the cumulative measure of size within each explicit stratum, C_g, is calculated. For each selection number Z_{gj}, a centre/school S_g is selected if the cumulative measure of size C_{gS} is larger or equal to the selection number ($C_{gS} \geq Z_{gj}$) and the cumulative measure of size of the immediately preceding centre/school is smaller than the selection number ($C_{gS-1} < Z_{gj}$).

5.10. Replacement centres/schools

In order to maintain the sample size in case of refusals and to reduce the risk of non-response bias, two replacement centres/schools were assigned to each originally selected centre/school. The replacements were the centres/schools immediately following and preceding the original centre/school on the sorted sampling frame, as long as these centres/schools were within the same explicit stratum. In case the first or last centre/school of an explicit stratum was selected, the replacements were the two centres/schools following or preceding the selected centre/school.

Centres/schools that already belonged to the original sample could not be selected as replacements. In the same way, an assigned first replacement could not be selected as a first or second replacement for another sampled centre/school. Therefore, it was not always possible to assign two replacements for large centres/schools in Estonia. In some cases, no replacement could be assigned because adjacent centres/schools were sampled as well.

5.11. Example of a sample selection

In what follows, the sample selection process is illustrated by means of an example. In a hypothetical explicit stratum g (frame in Table 5.9 below), there are 15 centres/schools, and 4 centres/schools are to be sampled ($D_g = 4$). The total measure of size S_g is 383, which gives a sampling interval I_g of $\frac{383}{4} = 95.75$. The random number RN_g used is 0.9000.

The first selection number Z_{g1} is $0.9 \times 95.75 = 86.175$. The first centre/school to be selected is the third one in the list because its cumulative measure of size is the first to reach or exceed the selection number.

The second selection number Z_{g2} is $0.9 \times 95.75 + 95.75 = 181.925$. This leads to the sixth centre/school being selected.

The third selection number Z_{g3} is $0.9 \times 95.75 + 2 \times 95.75 = 277.675$, and the final selection number Z_{g4} is $0.9 \times 95.75 + 3 \times 95.75 = 373.425$.

The centres/schools following and preceding each selected centre/school are assigned as replacements. Since the last centre/school in the frame has been selected, its second replacement is the centre/school preceding its first replacement.

Table 5.9: Example of a sample selection

Centre/school	MOS	Cumulative MOS	Selection number Z_{gj}	Selection
1	40	40		
2	35	75		R2
3	32	107	86.175	Selected
4	30	137		R1
5	28	165		R2
6	27	192	181.925	Selected
7	26	218		R1
8	25	243		
9	24	267		R2
10	24	291	277.675	Selected
11	22	313		R1
12	20	333		
13	18	351		R2
14	17	368		R1
15	15	383	373.425	Selected

Note: MOS = measure of size.

5.12. Centre/school IDs

Each selected centre/school and each replacement received a unique four-digit identifier (ID). For the originally selected centres/schools, the identifier started with a "1", followed by a three-digit sequential number. For instance, a sample of 200 centres/schools had the IDs 1001 to 1200. The replacement centres/schools had the same ID as their original, increased by 1000 for the first replacements (e.g. 2001 to 2200), and by 2000 for the second replacements (e.g. 3001 to 3200).

5.13. Sampling for the field test

From October to December 2017, the instruments and procedures were tested in each participating country. Per country, at least 30 centres/schools were sampled for this purpose. Both the field test sample and the main study sample were selected simultaneously by selecting a sample of at least 230 centres/schools to begin with. In a second step, those centres/schools were randomly distributed between the field test and the main study. This was achieved by selecting 200 out of 230 centres/schools for the main study, using generally the same routines as described above. For the field test, only one replacement was assigned to each selected centre/school.

By selecting both samples in a combined routine, the overlap between the two study phases could be controlled to avoid the same centres/schools being selected twice, which might have had a negative impact on the willingness of centres/schools and staff to participate.

5.14. Selecting children within centres/schools

The within-centre/school sampling procedures were carried out by the National Study Centres, using the Within-school Sampling Software for Windows (WinW3S) provided by the IEA.

After the sample of centres/schools had been selected, the National Study Centres contacted the selected centres/schools and asked them to complete the Child-Staff Linkage Form. Within this form, all eligible children had to be listed, together with their date of birth, gender and special education needs. Children unable to participate due to limited experience in the language or special education needs were not to be left off the listing.

Figure 5.1 shows an example of a fictitious completed Child-Staff Linkage Form.

Figure 5.1: Child-staff linkage form

Child Name or Number	Sequence Number	Date of Birth			Gender	Special Education Needs	Staff Information									
		DD	MM	YYYY			Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:	Staff's Name:
		01	04	2013	1		x									
		30	09	2013	2				x							
		15	06	2013	1						x					
		18	05	2013	2		x									
		21	07	2013	1					x						
		13	04	2013	1			x								
		04	08	2013	2		x									
…		…	…	…	…	…	…	…	…	…	…					

Note: Gender (Column 4): 1 = Female; 2 = Male\Special Education Needs (Column 5):
1 = Child with functional disabilities – child has a moderate to severe permanent physical disability;
2 = Cognitive, behavioural, or emotional disability – in the opinion of qualified staff, child has a cognitive, behavioural, or emotional disability;
3 = Limited assessment language experience – child is not a native speaker of any of the languages of the assessment in the country and has limited proficiency in these languages

After importing the information from the Child-Staff Linkage Form to the sampling software, children were sorted by gender and age, and then a systematic random sample of 15 children was selected with equal probabilities from each centre/school.

The information about special education needs did not have an influence on the sample selection; children with special education needs could be sampled. However, it was recognised that some children were not able to fulfil the basic requirements of the assessment situation. For these children, it was possible to assign the participation status "Did not participate due to special education needs".

For each sampled child, additional information was collected from parents and staff. In the Child-Staff Linkage Form, each child was connected to a staff member who knew the child best. If the child was selected, the staff member was given a questionnaire about the child.

Notes

[2] Although there are some differences across states, children in the United States often enter kindergarten class within a school if they are five years old by the start of the school year. As not all elementary schools in the United States have a kindergarten class, those without kindergarten were excluded from sampling.

Chapter 6. Linguistic quality assurance

The direct assessments were administered in English (England and USA), and Estonian and Russian (Estonia). The questionnaires were administered in English (England and USA) and Estonian. In the main study, both questionnaires were also administered in Russian (Estonia) and the parent questionnaire was also administered in Spanish (USA). The purpose of verification is to ensure that the instruments are functionally equivalent across different cultural contexts.

To that end, the translation process was conceived as a collaborative effort between national research partners and the consortium. The following actions were undertaken in the translation process:

- A translatability assessment was performed before the direct assessments and questionnaires were finalised. Linguists from different language groups identified and reported potential translation/adaptation issues. Possible solutions were discussed with the authors. Item-per-item translation and adaptation notes were drafted.
- The national research partners in the participating countries appointed their own translation teams.
- National adaptations were discussed and agreed before the translation process.
- The countries' translation teams were trained to use an open-source Computer-Assisted Translation (CAT) tool called OmegaT. Additional support was provided in the form of a step-by-step user manual, video tutorials and an interactive online training. The teams received technical support throughout the translation process.
- The translated/adapted versions of the direct assessments and the questionnaires were submitted to cApStAn for translation verification. The verification feedback was sent to national teams for discussion. All decisions taken were documented in a centralised monitoring tool designed by cApStAn.

6.1. Translation preparation

6.1.1. Translatability assessment

Before the master versions of the direct assessments and the questionnaires were finalised, the source texts were submitted to three linguists from different language groups (German, Polish, Estonian) for translatability assessment.

The purpose was to identify the challenges translators would face if they had to translate the direct assessments and questions in the proposed form. Their feedback was then collated in a report by a senior linguist at cApStAn who focused on issues that could be generalised across several languages.

As a result of this process, the senior linguist proposed;

- a translation/adaptation note to clarify a given term or expression, or to indicate the type of adaptation that may be necessary; and/or
- an alternative wording i.e. a new formulation that circumvented the problem without loss of meaning.

The translatability report was sent to the direct assessment/questionnaire authors, who had the opportunity to eliminate ambiguities, to address cultural issues or avoid unnecessary complexity.

6.2. Translation and adaptation guidelines

One of the outcomes of the translatability assessment was that it produced a subset of translation and adaptation notes. These item-per-item notes were reviewed and validated by the direct assessment developer and were entered by cApStAn into the centralised monitoring tool.

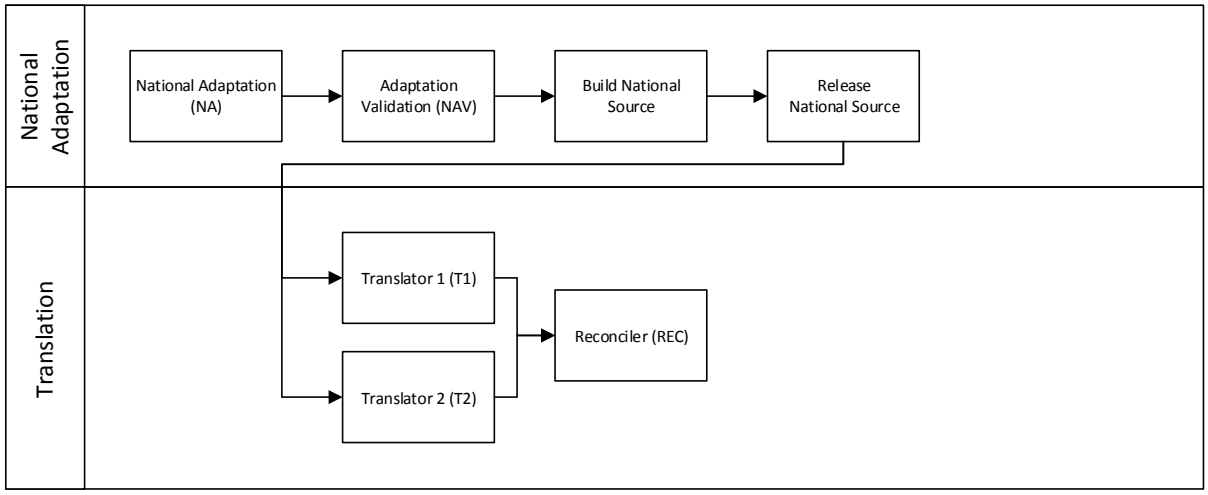
They were provided in addition to the general guidelines integrated in the Survey Operations Procedures (SOP) document, unit 3, "Instrument preparation". The translation teams were encouraged to document how they addressed the item-per-item guidelines. The verifiers systematically checked whether they had been addressed in a satisfactory fashion and reported possible deviations.

cApStAn designed a follow-up Excel workbook called TAVM (Translation, Adaptation and Verification Monitoring) for the documentation of the entire production of each target version. In the TAVM, the translation, adaptation and verification history of each language version of each item was documented; each decision or corrective action was noted. The TAVM contained the source text, item-per-item guidelines and one column for each step performed in subsequent order in the workflow (from adaptation to sign-off, including the audio check) so that the person responsible for that step could add comments and remarks. This form was released to the national centres together with the files for translation and accompanied the files throughout the entire workflow.

For the main study, the TAVM also included a column indicating the source updates from the field trial to the main study to facilitate the task for the translators.

6.3. National adaptation and translation

Figure 6.1: National adaptation and translation workflow



A presentation about the adaptation, translation and verification procedures was given at the National Project Managers meeting. Additional webinars with the translation teams were organised to explain the procedure in greater detail and translators were trained to use the open-source Computer-Assisted Translation tool OmegaT. Additional support was provided in the form of step-by-step online user manuals. The teams also received technical support throughout the translation process.

6.3.1. Translation procedure for direct assessments

- The National Study Centre proposed any national adaptations to the direct assessments in the TAVM.
- After the validation of the national adaptations by the International Study Centre, the National Study Centre double-translated the source script (this step did not apply to English versions).
- The two translations were then reconciled by a third linguist, called the Reconciler (this step did not apply to English versions).
- Once the files were verified at cApStAn following the verification procedure described below, they were signed off by the National Study Centre.
- After the assessment developers' sign-off and the verifier's final check, the script was recorded at the National Study Centre.
- A parallel check of the script and the audio files was then performed by cApStAn verifiers. This was followed by a possible re-recording and then sign-off by the National Study Centre.

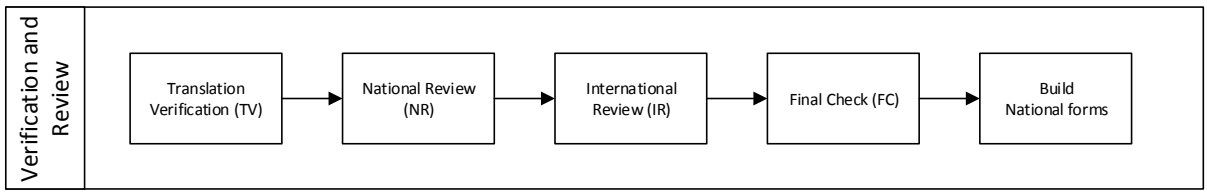
6.3.2. Translation procedure for questionnaires

- The National Study Centre proposed any national adaptations to the questions in the TAVM.

- After the validation of the national adaptations by the International Questionnaire Team, the National Study Centre double-translated the source questionnaire (this step did not apply to English versions).
- The two translations were then reconciled by a third linguist, called the Reconciler (this step did not apply to English versions).
- Once the files were verified at cApStAn following the verification procedure described below, the National Study Centre performed a verification review.
- The International Questionnaire Team performed a final content check which was followed by a final check by cApStAn.
- The National Study Centre signed off the questionnaires.

6.4. Verification and review

Figure 6.2: Verification and review workflow



At the verification stage, the main role of the verifier was to double-check that:

- the target version was free from language mistakes (spelling, grammar, and syntax);
- the target version was equivalent to the international English source;
- the agreed adaptations had been implemented correctly.

To that end, the verifiers were instructed to compare the source and target materials sentence by sentence, by looking at the preview on the TMS portal. In doing this, they referred to the item-per-item guidelines in the TAVM at all times.

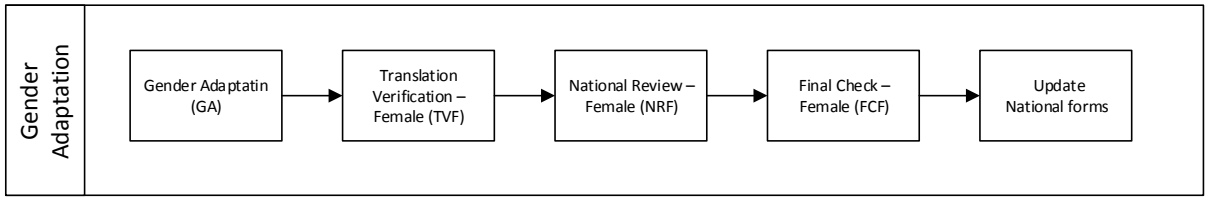
They were also asked to react to any item-per-item guideline and the agreed adaptations.

Before uploading the xliff file on the TMS portal, they were required to preview their changes.

At the final check, the verifier's role was to check that any pending issues had been addressed in a satisfactory way.

6.5. Gender adaptation

Figure 6.3: Gender adaptation workflow



The source script for direct assessments included sentences directly addressed to the respondents or expressed from the point of view of the respondent (e.g. "Are you ready?", "I feel surprised"). In languages with grammatical gender, such as Russian, these sentence constructs may need to be changed depending on whether the respondent is a male or a female. In order to guarantee gender equality, the Russian script needed to be recorded in two versions: a version for male respondents and a version for female respondents.

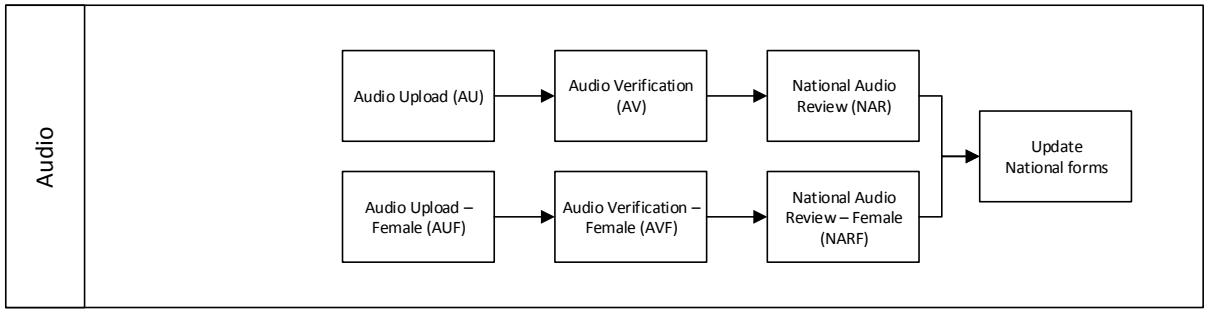
For Russian, once the script for male respondents was translated, reconciled and verified, the text was adapted for female respondents by the reconciler. At this step, the main role of the verifier was;

- to focus only on the gender adaptations by cross-checking them via the diff reports (files which tracked the differences between the two versions);
- to make sure no changes were made to the script except for gender equivalence reasons;
- to correct any language issues in the gender adaptations and document them in the TAVM.

After the final check on residual issues, the script for female respondents was recorded.

6.6. Audio

Figure 6.4: Audio workflow



For the audio check of the direct assessments, the verifier's main role was to ensure that;

- the voice was fluent, understandable for all children regardless of their geographical location, appropriate for small children in terms of pace and intonation;
- the audio files corresponded to the script on the TMS portal;
- the audio files corresponded to the agreed script in the TAVM.

To that end, while listening to the audio, they kept the Excel script file open and checked that the recording corresponded to what was agreed. At the same time, they checked any guidelines in the TAVM as well as any comments from the voice-over artist.

If they encountered any issue, they described their findings in the relevant column in the TAVM, by starting their comments with one of the following labels: Noise, not suitable for target age, accent, pronunciation/word stress, voice gender/voice age, not matching the script, missing preview, or other issue.

At final check, the verifier's role was to check that any pending issues had been addressed in a satisfactory way.

Chapter 7. Field operations

Successful administration of the IELS assessment depended heavily on the contributions of the study's National Project Managers and their National Study Centre staff.

The IELS International Consortium developed internationally standardised field operations procedures to assist the NPMs and to aid uniformity of their assessment administration activities. The international team designed these procedures to be flexible enough to simultaneously meet the needs of individual participants and the high-quality expectations of survey standards.

7.1. Overview of responsibilities

7.1.1. National Project Managers

One of the first steps that all countries participating in IELS had to take when establishing the study in their country was to appoint an NPM. Each NPM had primary responsibility for implementing the study at the national level.

The International Consortium recommended that each NPM should appoint a National Data Manager (NDM) to oversee and implement all data and technical tasks, and a National Sampling Manager (NSM) to help the NPM manage the complex sample design and potential national additions and extensions to the survey samples.

The number of staff members in the National Study Centres varied from one country to the next depending on the country size and how it chose to organise the national data collection work. Some NPMs worked with external survey organisations to conduct scientific and/or operational tasks.

NPMs, data managers and/or sampling managers;

- established overall preparation and administration schedules in co-operation with the International Consortium
- attended NPM meetings in order to become familiar with all instruments, materials and survey procedures
- provided the International Consortium with up-to-date national sampling frames
- discussed nationally-specific aspects, such as international and national options, with the International Consortium
- established procedures for maintaining confidentiality of respondents and security of materials at all times
- performed within-centre/school listing and child sampling
- prepared the national versions of the assessment items, questionnaires, Centre/School Co-ordinator and Study Administrator Manuals, forms, script, and coding guides

- documented required national adaptations of the instruments on Translation, Adaptation and Verification monitoring Workbooks (TAVMs)
- appointed experienced translators to produce the national versions of the international instruments
- organised and prepared tablets for online data collection
- prepared parent and staff questionnaires for paper-based data collection
- identified and trained Centre/School Co-ordinator from each of the sampled centres/schools
- recruited and trained Study Administrators
- appointed and trained National Quality Assurance Monitors (NQAMs)
- nominated possible International Quality Assurance Monitors (IQAMs) working on behalf of the International Consortium and supported their work (main study only)
- monitored the return of parent and staff questionnaires
- arranged the data entry of responses from parent and staff questionnaires administered on paper
- organised coding of the occupational data on questionnaires
- completed Survey Activities Questionnaires after data collection
- submitted data and documentation to the International Consortium and responded to data queries during data processing and analysis.

7.2. Centre/school co-ordinators

In co-operation with centre/school leaders, National Study Centres identified and trained Centre/School Co-ordinators for all participating centres/schools. The Centre/School Co-ordinator could be a staff member of the centre/school or a staff member of the National Study Centre. Centre/School Co-ordinators acted as the main contact person within centres/schools and were responsible for organising IELS in the centres/schools.

In this role Centre/School Co-ordinators;

- liaised with the National Study Centre
- identified eligible children belonging to the target population
- co-ordinated the accurate listing of children, their characteristics and their main responsible staff on the Child-Staff Linkage Form
- planned the assessment days (e.g. arranged rooms, scheduled the assessment sessions and breaks)
- informed centre/school staff, children and their parents about the study and the assessment schedule
- received and completed the lists of sampled children (Child Tracking Forms) and the lists of selected staff (Staff Tracking Forms)
- liaised with all staff members who needed to respond to staff questionnaires and the parents of the sampled children

- ensured that the materials received from the National Study Centre were complete and kept in secure places and confidential at all times
- distributed access credentials for the online staff and parent questionnaires
- distributed the corresponding paper staff and parent questionnaires
- ensured that the paper staff and parent questionnaires were returned in time
- co-operated with the Study Administrators.

7.3. Study administrators

7.3.1. Recruitment

To minimise burden on the participating centres/schools and to encourage standardisation in the administration of the assessment sessions, Study Administrators were recruited to administer the assessment within the sampled centres/schools[3].

The Study Administrators could not be recruited from staff of the centres/schools in which they administered the assessment. Administrators were expected to have experience in working with children in the target age cohort, good communication skills, clear pronunciation, and pedagogical experience. Additionally, it was required that the Study Administrators had high sensitivity with regard to the direct assessment situation and were computer literate.

7.3.2. Training

Prior to the assessment days, Study Administrators were trained by the NPM or National Study Centre. It was required that all Study Administrators were trained in person, so NPMs could ensure that Study Administrators met the required conditions. The aim of the training was to familiarise Study Administrators with the purpose of the study, as well as the procedures for administering the direct assessment to the sampled children.

The Study Administrator training comprised, but was not limited to, the following topics;

- introduction to IELS
- characteristics of the age group
- preparation of the assessment
- general guidelines and the script for administering the assessment
- handling of the tablets
- parent consent forms (if required)
- formats of responding
- rapport building with the children and practice sessions
- schedule of the assessment sessions
- Child Tracking Form and procedures for its completion
- trouble shooting options on how to handle potential challenges
- general guidelines to act as a representative of the National Study Centre

- nationally-specific issues, such as protocols for entering centres/schools or communicating with centre/school staff.

7.3.3. Responsibilities

Study Administrators were responsible for administering the assessment within the centres/schools, complying with all guidelines as outlined in the Study Administrator Manual and the Study Administrator training. Additional responsibilities contained;

- confirming assessment plans and schedule with the Centre/School Co-ordinators
- verifying that the tablets and equipment received from the National Study Centre were complete
- bringing the tablets and accompanying equipment to the centres/schools
- arranging tablets and accompanying materials on the assessment days
- reviewing the accurate listing of children and their characteristics on the Child-Staff Linkage Forms
- administering the sessions to the sampled children as indicated by the Child Tracking Forms
- timing the assessment sessions
- completing the Administration Report Forms (feedback on the children's behaviour, technical issues, course of the administration, etc.)
- reviewing and updating the Child Tracking Forms
- uploading the response data collected on the tablets to the cloud-hosted central database
- returning the tablets and accompanying equipment to the National Study Centre.

7.4. Survey Operations Procedures units and manuals

During all phases of IELS, the National Study Centres adhered to the standardised procedures prepared by the International Consortium. The International Consortium outlined these procedures in the following documents, which they released to National Study Centres prior to the field test. Before the main study updated versions of these documents were released.

- The NPM Manual provided an overview of IELS, details of the tasks that NPMs and National Study Centres would carry out, and information about key milestones and deliverables.
- Survey Operations Procedures Unit 1 (sampling of centres) specified the actions and procedures required to develop a national sampling plan in compliance with the international IELS sample design.
- Survey Operations Procedures Unit 2 (working with centres) contained information about how to work with centres/schools in order to plan for a successful administration of the IELS assessment.
- Survey Operations Procedures Unit 3 (instrument preparation) described the process of national adaptations, translation, translation and layout verification of and audio recordings for the IELS instruments within the country.

- Survey Operations Procedures Unit 4 (data collection) dealt with the processes involved in preparing for, supporting and monitoring the IELS data collection in centres/schools and provided guidelines for selecting and training Study Administrators.
- Survey Operations Procedures Unit 5 (data capture procedures) comprised the post-data collection processes and tasks. These included data capture from the paper questionnaires, data quality control and coding parental occupational data.
- The Centre/School Co-ordinator Manual detailed the role and responsibilities of the Centre/School Co-ordinator. NPMs were responsible for translating the manual into the language(s) in which they would administer the assessment and for adding national information where necessary.
- The Study Administrator Manual described the role and responsibilities of the Study Administrators. Additionally, the manual included the script with detailed instructions for the Study Administrators on how to conduct the practice and the assessment sessions to ensure they were conducted the same way in all centres/schools. NPMs were responsible for translating the manual into the language(s) in which they would administer the assessment and for adding national information where necessary.
- The National Quality Assurance Monitoring Manual specified the role and responsibilities of the National Quality Assurance Monitors (NQAMs), timelines, actions and procedures to be followed in order to carry out the national quality control programme. This programme related closely to the programme carried out by the international monitors. However, NPMs were free to adapt the manual and procedures according to their country-specific needs.
- The International Quality Assurance Monitoring Manual (only main study) was delivered directly to the International Quality Assurance Monitors (IQAMs), all of whom were contracted by the IEA. The manual outlined the tasks the monitors needed to complete in order to check the quality of the survey operation procedures within the participating countries.
- The Tablet Application User Guide specified how to use the tablet application delivering the child assessment.

During the meetings with the NPMs and data management training sessions, the International Consortium described and explained the field operation procedures outlined in the survey operation procedures units, manuals and guidelines and provided guidance on how to use the software packages.

7.5. Field operations procedures

7.5.1. Working with centres/schools

In IELS, the assessment administration required close co-operation between the National Study Centres and Centre/School Co-ordinators. Once NPMs had obtained a list of the centres/schools sampled for IELS (for centre/school sampling, please refer to Chapter 5. of this report), NPMs encouraged the centres/schools to take part in the assessment, a process that often involved obtaining support from national or regional educational authorities or other stakeholders, depending on the national context.

High response rates are one of the prerequisites for obtaining reliable data. Therefore, it was very important to establish and maintain good co-operation with centres/schools, staff and parents from the outset of the study.

National Study Centres requested a list of all eligible children from each participating centre/school. The National Study Centres used this information to sample children within the centres/schools. For the child sampling, the NPMs were asked to use the IEA WinW3S software prepared and provided by the International Consortium (for child sampling, please refer to Chapter 5. of this report). The software generated the list of sampled children for each centre/school, the so-called Child Tracking Form, being the central documentation instrument and linking children, direct assessment and parent questionnaires.

7.6. Shipping materials to centres/schools

Prior to the assessment days, National Study Centres provided each Centre/School Co-ordinator with the following instruments and key documents;

- a sealed envelope for each staff member with the cover letter containing login credentials for the online staff questionnaire and the corresponding paper staff questionnaire, or instructions on how to request a paper questionnaire (as applicable)
- a Child Tracking Form listing all sampled children
- a Staff Tracking Form listing all staff members designated for IELS
- a number of Staff-Child Assignment Forms according to the number of staff assigned to the sampled children to be used as reference
- a sealed envelope for each sampled child's parents with the cover letter containing the URL and login credentials for the online parent questionnaire and the corresponding paper parent questionnaire, or instructions on how to request a paper questionnaire (as applicable)
- a set of labels (and preferably pre-paid envelopes) addressed to the National Study Centre to facilitate the return of all study materials
- a clear indication of the date by which the online questionnaires had to be filled and the completed paper questionnaires had to be returned to the National Study Centre
- a list of materials to check against.

7.7. Assigning questionnaires to parents and staff

Administering the questionnaires is one of the main tasks of the Centre/School Co-ordinators. However, NPMs needed to monitor as closely as possible the on-going participation rates of parents and staff. The participation rate estimates produced by IEA WinW3S helped with this task.

Upon receipt, the Centre/School Co-ordinators assigned the questionnaires to parents and staff. In detail, they distributed the staff questionnaires as indicated on the Staff Tracking Form and the parent questionnaires as indicated on the Child Tracking Form. The cover page of each questionnaire indicated the name and/or identification code of the individual to whom the questionnaire should be administered. It was crucial that

each designated participant received the questionnaire assigned to him or her. Parents were not to be replaced under any circumstances. It was the responsibility of the Centre/School Co-ordinator to make sure that all questionnaires were handed out to the designated individuals.

Centre/School Co-ordinators collected the completed questionnaires as soon as possible, but no later than the return date given, recorded the participation of staff on the Staff Tracking Form and the participation of parents on the Child Tracking Form.

7.8. Shipping materials to Study Administrators

NPMs were requested to provide two tablets for each Study Administrator: one tablet used for assessing the child and one back-up tablet. NPMs downloaded the app on the tablets and disabled as much software on the tablet screen as possible, to ensure that the screen was free of distractions. In addition, they were asked to make replacement materials (adaptors, power cords, portable batteries, etc.) available in case of damage or technical defects.

Apart from the tablets needed for the administration of the direct assessments, each Study Administrator needed to receive the following key documents in time for the assessments:

- Study Administrator Manual including all appendices (one copy for reference)
- Tablet Application User Guide (one copy for reference)
- Study Administrator Checklist (one copy per assessment day)
- Administration Report Form (one copy per child)
- a set of labels (and preferably pre-paid envelopes) addressed to the National Study Centre to facilitate the return of all study materials
- if applicable: study information material (e.g. leaflets, brochures)
- materials receipt form.

7.9. Study administration

Administering the assessment sessions with the sampled children was the main task of the Study Administrators. After arriving at the centre/school on the assessment day, the Study Administrator was required to review and update the Child Tracking Form and parts of the Administration Report Form. Afterwards, the Study Administrator set up the room and materials for the assessment sessions and made him- or herself familiar with the on-site conditions.

7.10. Preparation of tablets

Before the start of the assessment session, the Study Administrator was requested to;

- ensure that the tablets were fully charged and the screen was clean
- check the tablets for functionality and that the sound was working
- change the sleep mode to 30 minutes
- lock the screen orientation to landscape mode

- ensure that the app was ready.

7.11. Assigning the assessment to children

The Study Administrator assigned the assessment forms to the children, following the guidelines in the Tablet Application User Guide. Each child was logged in for the scheduled domain by using the login details provided on the Child Tracking Form.

7.12. Administering the assessment

7.12.1. Rapport building and practice session

The Study Administrator was asked to start the session by building rapport with the child and preparing a comfortable and engaging environment. This could be done by asking about the child's favourite animal, hobbies, siblings, holidays, etc., engaging in a brief conversation, and/or explaining the purpose of the session.

When the practice programme was launched, the Study Administrator had to make sure that the child understood the audio and reminded the child to handle the tablet gently. During the practice session, the Study Administrator ensured that sufficient help and encouragement were provided to make the child feel comfortable when responding on the tablet or reporting any difficulties. Additionally, the Study Administrator ensured that the child understood the words used in the audio as well as the instructions and rules, and showed capability on how to handle the tablets; i.e. touch the screen, use drag-and-drop, operate the audio button, and move between screens.

7.12.2. Assessment session

During the assessment sessions, the Study Administrator had to;

- ensure that the administration room was quiet, well-lit, comfortable and free of distractions
- create a similar administration environment for all sessions
- support the child when needed with regard to tablet functions, operations of buttons, etc.
- provide scaffolding for the assessment questions, if required, but without providing the response to the child
- observe the child as he/she responds
- capture the child's feedback in the Administration Report Form
- record the child's participation on the Child Tracking Form.

Study Administrators were required to follow strictly the instructions described in the Study Administrator Manual to achieve comparable data and standardised assessment procedures across participating countries; adhering to the timing for the assessment sessions, shown in Table 7.1, was especially important.

Table 7.1: Timing of the assessment sessions

Administration day 1	Time
Rapport building with the child	5 minutes
Conducting the practice session for the first domain (Emergent literacy)	2-5 minutes
Administering the assessment (Emergent literacy)	15 minutes
Break between domains	0-5 minutes
Conducting the practice session for the second domain (Self-regulation4)	2-5 minutes
Administering the assessment (Self-regulation)	15 minutes
Total for two domains	39-50 minutes

Administration day 2	**Time**
Rapport building with the child	5 minutes
Conducting the practice session for the third domain (Emergent numeracy)	2-5 minutes
Administering the assessment (Emergent numeracy)	15 minutes
Break between domains	0-5 minutes
Conducting the practice session for the forth domain (Empathy)	2-5 minutes
Administering the assessment (Empathy)	15 minutes
Total for two domains	39-50 minutes

After the assessment sessions, Study Administrators and the Centre/School Co-ordinator collected all relevant materials and returned them to the National Study Centre.

7.13. Materials receipt at the National Study Centres

The major tasks for NPMs immediately after data collection included retrieving and collating the materials from centres/schools and Study Administrators and verifying their integrity. On receiving study materials from the centres/schools and Study Administrators, NPMs;

- verified that all identification numbers on all questionnaires were accurate and legible
- checked that the participation status recorded on the tracking forms matched the availability of the online and paper questionnaires
- followed up with those centres/schools and Study Administrators that did not return all study materials
- followed up with centres/schools in case forms or paper questionnaires were missing, incomplete or otherwise inconsistent.

National Study Centres recorded all necessary information about centres/schools, children and staff, including the return status of the parent and staff questionnaires, in WinW3S. NPMs then organised the paper questionnaires and corresponding forms for data entry (see Chapter 12.).

Notes

3 The number of required Study Administrators in a centre/school could vary according to how often a Study Administrator was appointed for the assessment sessions and the centre's/school's study administration plan.

4 There were three sub-domains for Self-regulation. They had to be completed in the following order: Inhibition, Working memory, Mental flexibility

Chapter 8. Quality Assurance Monitoring

One of the aims of IELS was to collect robust empirical data about children's social-emotional and cognitive development on the basis of which the participating countries will be able to conduct research and inform policy making in the field of early childhood education.

To facilitate the collection of high-quality data, the IELS Consortium developed detailed procedures and standards to monitor the quality of various processes and activities, including the sampling of ECEC centres, the preparation of national study instruments, and study administration.

The quality assurance measures designed for IELS consisted of the following three components;

- the International Quality Assurance Monitoring programme designed and overseen by the IEA in co-operation with independent International Quality Assurance Monitors (IQAMs) in each of the countries that participated in IELS (conducted during the main study)
- National Quality Assurance Monitoring programmes co-ordinated by the National Study Centres in each participating country (conducted during the field test and main study)
- the Survey Activities Questionnaire designed by the IELS Consortium to elicit feedback from the National Project Managers on the different processes and procedures of the study (administered during the field test and main study).

This chapter provides an overview of the quality assurance measures implemented for IELS, outlining the specifically designed procedures and reporting on the outcomes of each of the three previously mentioned components.

8.1. The International Quality Assurance Monitoring programme

In order to monitor the administration of IELS at the international level, the IEA designed the International Quality Assurance Monitoring programme. At the core of the programme, independent IQAMs contracted by IEA visited a sub-sample of the centres that participated in IELS to observe the administration of the study in the field. The IQAMs interviewed the Centre/School Co-ordinator at each of the centres visited about the implementation of the study.

8.2. Selecting and training IQAMs

The IQAMs were resident in the participating countries and had experience in the field of early childhood education, but no direct affiliation with the National Study Centres tasked with the implementation of IELS at the national level. In order to facilitate the recruitment of IQAMs, each National Project Manager nominated two candidates to serve as IQAMs. Potential candidates would fulfil the following criteria;

- have experience with social science studies
- be familiar with ECEC/early schooling environments and/or the day-to-day operations of these settings
- have ICT literacy
- be fluent in English and the common (state) language(s) of the respective country
- have no professional affiliation with the National Study Centre
- have no personal relation to the NPM or any other staff member at the National Study Centre.

The IEA invited the selected IQAMs to attend an in-person training seminar.[5] The purpose of the seminar was to equip IQAMs with all the information and materials necessary to complete their quality assurance tasks. To that end, representatives of the IELS Consortium provided an overview of the study, outlined the purpose of the quality assurance activities, and explained the IQAM responsibilities and deliverables. In addition, the IQAMs received the following materials that they needed to fulfil their tasks;

- the IELS IQAM Manual
- the IELS Technical Standards
- the IELS Survey Operations Procedures units 1, 2, 3, and 4
- the international versions of the Centre/School Co-ordinator Manual and the Study Administrator Manual (in English)
- the NPM Interview Outline and Question Template structuring IQAMs' visits to the National Study Centres
- a checklist for collecting materials from the NPMs
- a template of the Centre/School Visit Record guiding the centre visits and the interviews with the centre co-ordinators
- a template of the Manual Review Report

8.3. IQAM responsibilities

The International Quality Assurance Monitoring programme entailed the following responsibilities. IQAMs were expected to;

- familiarise themselves with the programme content
- visit the NPM in their country and select 20 centres/schools to be visited
- arrange visits with the 20 selected centres/schools and observe one direct child assessment session and interview the Centre/School Co-ordinator in each centre
- prepare the Manual Review Report
- document their findings and submit the documentation to the IEA.

The following sections will detail the three main IQAM tasks – visiting the NPMs, observing and interviewing in the 20 centres/schools, and completing the Manual Review Report – and provide summaries of the respective IQAM findings.

8.4. IQAM responsibility 1: visiting the NPM

The first major IQAM responsibility was to visit the National Study Centre to select centres for the observation of direct child assessments and to collect materials and documents. In order to structure the meeting with the NPM, the IEA provided the IQAMs with the NPM Interview Outline and Question Template that guided them through the tasks to be completed during the visit.

8.5. Selecting centres/schools

As part of the international quality assurance activities, IQAMs had to visit 20 centres/schools during the study administration window. The selection of the centres/schools was carried out jointly with the NPMs. Taking into account that each child's direct assessment contained two assessment days, IQAMs were expected to select centres/schools in a way that allowed them to make ten observations of Assessment Day 1 and ten observations of Assessment Day 2. IQAMs also had to consider that centres/schools had to be within reachable distance and that centres/schools should not be taking part in the National Quality Assurance Programme. After removing the centres/schools that did not fit these selection criteria, each IQAM was required to randomly select, together with the NPM, 20 centres/schools for the visits. In addition, they selected 20 replacement centres/schools in case any of the initially selected settings could not be visited.

8.6. Collecting materials from the NPM

During their visits at the National Study Centres, IQAMs obtained the following materials from the NPMs;

- a copy of the national version of the Centre/School Co-ordinator Manual
- a copy of the national version of the Study Administrator Manual
- copies of the Child-Staff Linkage Forms (one per centre/school)
- copies of the Staff-Child Assignment Forms (one per sampled staff member at every selected centre/school)
- copies of the Child Tracking Forms (one per centre/school)
- copies of the Staff Tracking Forms (one per centre/school).

IQAMs required copies of the translated and/or adapted manuals to complete the Manual Review Report (see IQAM Responsibility 3). They obtained the various forms in preparation for the observations of the direct child assessments and the Centre/School Co-ordinator interviews (see IQAM Responsibility 2).

8.7. IQAM responsibility 2: observing direct child assessments and interviewing Centre/School Co-ordinators

The second major task of the IQAM programme entailed a total of 20 observations of direct child assessments and interviews with Centre/School Co-ordinators. At each visited centre/school, IQAMs observed one direct child assessment session (either the session on Assessment Day 1 or the one on Assessment Day 2) and interviewed the Centre/School Co-ordinator to obtain information about the implementation of IELS at centre/school level. Table 8.1 provides an overview of the number of planned and achieved centre/school visits (including observations of assessment sessions and Centre/School Co-ordinator interviews) for each of the participating countries. As can be seen in the table, 56 of the 60 planned centre/school visits (93%) took place.

Table 8.1: Number of planned and achieved centre/school visits

Country	Assessment Day 1 sessions	Assessment Day 2 sessions	Total
England	10 (10; 100%)	10 (10; 100%)	20 (20; 100%)
Estonia	10 (10; 100%)	10 (10; 100%)	20 (20; 100%)
United States	12 (10; 120%)	4 (10; 40%)	16 (20; 80%)
Total	32 (30; 107%)	24 (30; 80%)	56 (60; 93%)

Notes: The numbers provided in the parentheses represent the planned number of visits. The percentages represent the ratio of achieved visits over planned visits. In the United States, the IQAM observed 12 instead of ten Assessment Day 1 sessions and four instead of ten Assessment Day 2 sessions, resulting in 16 instead of 20 visited centres/schools. This discrepancy between the planned and the achieved number of visits was due to logistical reasons related to the administration schedule and the travel distance between centres/schools.

IQAMs used the Centre/School Visit Record to document their observations of the assessment sessions and to record the Centre/School Co-ordinators' answers to the interview questions concerning their initial preparations and general impressions. The Centre/School Visit Records consisted of five sections with a total of 54 questions. After completion of the centre/school visits, IQAMs uploaded their findings to the IEA Online Survey System. The following subsections summarise the IQAMs' findings of their centre/school visits as documented in the Centre/School Visit Records.

8.8. The administration of the direct child assessments

The IELS Consortium prepared standardised scripts that Study Administrators were expected to follow when delivering the direct assessments to the children. As indicated by the IQAMs in the Centre/School Visit Records, 53 of 56 observed Study Administrators followed the script for the practice session, and 54 Study Administrators followed the script for the assessment domains.[6] 53 Study Administrators followed the instructions related to providing support or feedback to children during the assessment sessions. Regarding the overall administration of the assessment sessions, IQAMs reported that 42 Study Administrators followed the standardised procedures outlined in the Study Administrator Manual. In the remaining 14 cases, the standardised procedures were followed to some extent and deviations were mostly related to the number and duration of the breaks taken during the assessment sessions.

In respect to the suitability of the assessment environment, the IQAMs considered the conditions in the assessment room as suitable for the children to work without being

distracted in 47 of the visited centres/schools. For the remaining nine centres/schools, the IQAMs reported that distraction was caused by outside noise, by several child assessments taking place simultaneously, or by the rooms used for the assessment being too small.

According to the definition of the child sample, children with special education needs could participate in the assessment. In the interviews with the IQAMs, 14 Centre/School Co-ordinators reported that, at their centres/schools, children requiring special accommodation (for example, children with visual or hearing impairments or dyslexia) participated in the assessment.

8.9. Information about Centre/School Co-ordinators

One of the sections of the Centre/School Visit Record contained questions about the Centre/School Co-ordinators' professional background. All Centre/School Co-ordinators held positions internal to the centres/schools visited by the IQAMs. Most of the co-ordinators were the centre leaders or school principals (37 out of 54) at the visited centres/schools, followed by staff members or teachers (14 out of 54) and auxiliary staff members (3 out of 54). Three Centre/School Co-ordinators indicated that they were co-ordinating the administration of IELS in more than one institution, and 50 Centre/School Co-ordinators had not previously served as co-ordinators for another national or international study.

8.10. Centre/School Co-ordinators' initial preparations for IELS

In order to ensure the smooth implementation of IELS at centre/school level, it was essential that Centre/School Co-ordinators received instructions from the National Study Centres and made some preparations prior to the study administration. As recorded by the IQAMs, six Centre/School Co-ordinators attended training sessions designed for them. Of the remaining 50 Centre/School Co-ordinators, 47 co-ordinators had spoken to the NPM, read the Centre/School Co-ordinator Manual and familiarised themselves with the study and return procedures in preparation for their role in co-ordinating the administration of IELS. Five Centre/School Co-ordinators received a leaflet or brochure from the NPMs explaining the purpose of IELS. Regarding communication with the NPMs, only two of the interviewed co-ordinators experienced difficulties that resulted in delays or unexpected changes.

Regarding their understanding of different aspects of the study procedures, IQAMs reported that two Centre/School Co-ordinators had difficulty understanding the purpose of IELS, while three co-ordinators had difficulty understanding the different components of the study administration. Eight co-ordinators had some difficulty understanding the return procedures of online and offline questionnaires.

8.11. Using listing and tracking forms

In order to facilitate a standardised administration of IELS, the IELS Consortium provided National Study Centres with listing and tracking forms designed to assist with the distribution of study instruments and the monitoring of participation (see Chapter 7.).[7]

In order to check the usage of these forms, IQAMs compared the forms they received from the NPM against the ones used in the visited centre/school. The IQAMs reported that they did not find any inconsistencies between the Child Tracking Forms they had

collected from the NPMs and the Child Tracking Forms used in the visited centres/schools. Four Centre/School Co-ordinators reported to the IQAMs that they did not use the Child-Staff Linkage Form, the Child Tracking Form and the Staff Tracking Form. Five out of 53 co-ordinators experienced difficulties completing them.

8.12. Centre/School Co-ordinators' general impressions of IELS

This part of the Centre/School Visit Record was designed to elicit Centre/School Co-ordinators' general impressions of the administration of IELS in the centre/school(s) they were responsible for. Regarding staff members' attitudes toward the study, 47 out of 54 Centre/School Co-ordinators reported that staff members who took the IELS questionnaire were extremely co-operative, while seven co-ordinators indicated that staff members were moderately co-operative. Regarding parents' willingness to participate in the study, 26 of the interviewed Centre/School Co-ordinators reported that parents who took the IELS questionnaire were extremely co-operative. 21 co-ordinators stated that parents were moderately co-operative. The remaining co-ordinators indicated that parents were somewhat co-operative (8) or hardly co-operative at all (1).

As shown in Table 8.2, Centre/School Co-ordinators reported that they were approached by staff members and parents with questions regarding different aspects of the study.

Table 8.2: Number of Centre/School Co-ordinators who reported that questionnaire respondents (staff members and/or parents) approached them to discuss specific aspects of the study.

Study aspects	Staff members (n = 56)	Parents (n = 55)
Purpose of the study	15	5
Questionnaire return procedures	9	4
Clarification of any items	4	2
An error they spotted	2	0
Any questions they could not answer	4	1
Other questions about the study	3	6

As documented in the Centre/School Visit Records, the co-ordinators moreover reported that none of the staff members at their centres/schools refused to participate in IELS. However, 22 out of 55 Centre/School Co-ordinators indicated that some of the children's parents refused to participate in the study. 49 out of 53 Centre/School Co-ordinators described the staff questionnaire distribution process as very efficient, and 47 co-ordinators felt that the parent questionnaires were distributed without any problems.

8.13. IQAM review

The final section of the Centre/School Visit Record covered questions that asked the IQAMs to reflect upon their impressions of the IELS administration at each centre/school that they visited. With regards to the Centre/School Co-ordinators' preparedness, the IQAMs reported that 46 co-ordinators appeared very well prepared and six co-ordinators were somewhat prepared (with 'very well', 'somewhat' and 'not at all' prepared being the response options). The IQAMs were furthermore certain that 46 co-ordinators applied the study procedures seriously and professionally.

8.14. IQAM responsibility 3: completing the Manual Review Report

The IQAMs' third main responsibility was to review the national versions of the Centre/School Co-ordinator Manual and the Study Administrator Manual and to document their findings. The IELS Consortium provided the participating countries with the international version of the Centre/School Co-ordinator Manual and the Study Administrator Manual describing the respective roles, tasks and responsibilities. It was then the countries' responsibility to translate and/or adapt the manuals and to distribute them to the Centre/School Co-ordinators and Study Administrators. While the translated manuals did not undergo translation verification, the IELS Technical Standards stipulated that the national versions of the manuals had to be consistent with the international English source version; any deviations from the international manuals required the IELS Consortium's approval and needed to be documented adequately.

It was the IQAMs' task to compare the translated and/or adapted Centre/School Co-ordinator Manuals and Study Administrator Manuals against the respective international versions in order to detect and document any deviations with regards to content and instructions for the Centre/School Co-ordinators and Study Administrators. The Manual Review Report asked IQAMs to indicate for every section of the manuals whether the provided information was consistent with the international template.

The countries participating in IELS used a total of three national language versions of the Centre/School Co-ordinator Manual: England and the United States adapted the English source version, and Estonia translated the manual into Estonian. England and Estonia used both the paper and the electronic template of the manual, while the United States used only the latter.

The Study Administrator Manual was translated and/or adapted into four national versions: England and the United States used English versions of the manual, while Estonia translated the international template into Estonian and Russian. As with the Centre/School Co-ordinator Manual, England and Estonia prepared an electronic and a paper version of the manual, and the United States used only the electronic template of the Study Administrator Manual.

8.15. National Quality Assurance Monitoring programmes

In addition to the international quality assurance activities outlined above, the participating countries co-ordinated quality assurance monitoring at the national level. The purpose of these national activities was to enable National Study Centres to receive feedback directly from the study administration in the centres/schools. While the IELS Consortium provided the countries with guidelines on the implementation of national quality assurance activities as well as the National Quality Assurance Monitor Manual, including a template of a Centre/School Visit Record (adapted from the international quality assurance programme), it was the NPMs' responsibility to arrange and co-ordinate the National Quality Assurance Monitoring programmes for both the field test and the main study. Findings from these national quality assurance measures were to be reported to the IELS Consortium in the relevant section of the Survey Activities Questionnaire.

As stipulated in the IELS Technical Standards, National Quality Assurance Monitors (NQAMs) were required to visit at least ten percent of the participating centres/schools (i.e. 20 centres/schools) in order to monitor the administration of the study. NQAM responsibilities included;

- verifying that the lists of respondents (children, parents, staff) had been prepared correctly by the Centre/School Co-ordinator.
- checking that the Centre/School Co-ordinator had prepared the name sheets if confidentiality regulations did not allow respondents' names to be sent to the National Study Centre.
- checking that the Study Administrator had behaved according to the guidelines, administered the assessment session as described in the Study Administrator Manual and followed the wording in the scripts.
- verifying the completeness and security of the materials.
- checking that the Centre/School Co-ordinator took care of the confidentiality regulations.
- checking that the instructions for the documentation of the children's special education needs (SEN) were followed correctly.
- verifying the assignment of the instruments.
- checking that the distribution of materials had been undertaken correctly and that participation had been recorded in the tracking forms.

As documented in the Survey Activities Questionnaire, all three participating countries implemented national quality assurance activities during the field test and the main study. On average, NQAMs visited 20 centres/schools per country to monitor the study administration. In all countries, the NQAMs used the templates provided by the IELS Consortium.

8.16. The Survey Activities Questionnaire

The national and international quality assurance activities described above were complemented by the NPMs' comprehensive feedback on the steps and processes of the study cycle recorded in the Survey Activities Questionnaire. The main purpose of the SAQ was to gain information about all study-related activities and the extent to which the Technical Standards were followed. It also gave the NPMs an opportunity to share valuable feedback on IELS including on instrument preparation, study administration, manuals, guidelines, and software. The following subsections provide a summary of the SAQ results.

8.17. Contacting centres/schools and recruiting Centre/School Co-ordinators

According to the NPMs, National Study Centres usually recruited Centre/School Co-ordinators from the people working at the sampled centres/schools. National Study Centres called or emailed co-ordinators and used the Centre/School Co-ordinator Manual to inform them about their responsibilities. Two of the three NPMs indicated that they made changes to the manual including minor additions, deletions, and revisions to adapt the instructions to the national settings of the IELS administration. In one country, the National Study Centre arranged online trainings for Centre/School Co-ordinators in addition to written instructions. NPMs reported that the most challenging part of training Centre/School Co-ordinators was explaining the schedule of the assessment sessions.

Although final participation rates were considerably high, all three NPMs reported that they experienced difficulties in recruiting centres/schools to participate in IELS due to

the anticipated amount of additional work. While there were some issues in relation to the information provided by Centre/School Co-ordinators, the NPMs' overall feedback regarding centre/school contact was positive.

8.18. Within-centre/school sampling

As outlined in Chapter 5. of this report, the IELS Consortium provided countries with the WinW3S software to facilitate within-centre/school sampling. As reported in the SAQ, none of the participating countries experienced any conditions or constraints that necessitated deviations from the standard within-centre/school sampling design. In one country, national data protection laws restricted the information that the National Study Centre was allowed to put onto the linkage and tracking forms.

Two of the three National Study Centres used alternatives to the Child-Staff Linkage Forms and the Child and Staff Tracking Forms provided by the WinW3S software to list and track children and staff: one National Study Centre used Microsoft Excel workbooks in addition to the provided forms; in the other country, the National Study Centre created so-called Child Assessment Grids combining the information of the linkage and tracking forms. One NPM reported that the linkage and tracking forms were not easy to work with as the forms were difficult to understand and the total number of them was rather high.

Overall, the NPMs indicated that the within-centre/school sampling process using the WinW3S software was not very complex. One country expressed difficulty in using the software to produce forms and labels.

8.19. Documenting and implementing national adaptations

In order to document national adaptations to the study instruments, countries used the Translation Adaptation and Verification Monitoring Workbooks (TAVMs) provided by the IELS Consortium (see Chapter 6.). In the SAQ, NPMs reported that the procedure of documenting adaptations in the TAVMs was mostly clear, with exceptions pertaining to the difficulty of keeping track of the latest verified translation. Overall, the National Study Centres did not find it difficult to adapt the international source version of the study instruments (assessment items and questionnaires). The few difficulties reported in the SAQ related to adapting or mapping internationally classified ECEC levels to national programmes as well as ensuring consistency across multiple national languages of administration.

Most NPMs indicated that the submission process of the TAVMs for verification was clear; however, timelines and return dates could have been communicated more clearly. The adaptation negotiation feedback provided by ACER was considered somewhat useful: according to the NPMs, the utility of the process was limited by time constraints and the method used to provide the feedback (i.e. TAVMs). However, while the process for documenting and implementing national adaptations was challenging, all three NPMs were satisfied with the support they received from ACER and cApStAn.

8.20. Translating instruments and audio recordings

As outlined in Chapter 6. , National Study Centres were responsible for translating the international source version of the IELS instruments to the language(s) of administration if the study was not administered in English. Overall, NPMs' feedback on using the Translation Management System (TMS), the Translation and Adaptation Verification

Monitoring Workbooks (TAVMs) and OmegaT was positive. However, two NPMs reported some issues with respect to the use of XLIFF files. NPMs considered the translation verification feedback provided by cApStAn generally as useful, but some issues occurred with the method of providing the feedback (TAVMs) and the appropriateness of the feedback for the targeted age group. Based on the feedback, the National Study Centres corrected errors identified by the translation verifiers and adopted suggestions for improvement.

In general, countries did not have any difficulty employing translators, reconcilers and adaptors who fulfilled the selection criteria set by the IELS Consortium (see Chapter 6. for details). Most countries experienced difficulties in recruiting voice-over artists and in recording national audio instruments. The difficulties noted by the NPMs related to finding a sufficient number of voice-over artists with appropriate voices, managing the number of required recordings, and dealing with unexpected issues in the recording studio. All NPMs reported that the provided instructions were clear and sufficient.

8.21. Checking layout and preparing instruments for delivery

As summarised in Chapter 6. , the IELS Consortium checked the layout of all IELS instruments after the completion of the translation verification process. Feedback on layout verification was positive. In fact, all three NPMs indicated in the SAQ that there were no difficulties related to preparing the IELS questionnaires (paper and online) and assessment instruments for delivery.

8.22. Administration of direct assessments

Two National Study Centres recruited and employed external staff specifically to serve as Study Administrators for IELS. In one country, Study Administrators included staff members from the National Study Centre, staff members from sampled centres who did not work with the sampled children, and staff members from other centres. In general, it was not difficult to recruit Study Administrators. Across the three countries, between 85 and 206 Study Administrators were recruited, the vast majority of whom received in-person training. In two of the three countries, these training sessions included additional elements that were not required by the IELS Consortium such as providing information on the ethics of working with children and additional scheduling guidance. All three NPMs considered the Study Administrator Manual provided by the IELS Consortium as clear and useful.

The average duration of each session of two direct child assessments varied from 40 minutes to 55 minutes across the countries. While all assessments were administered one-to-one, up to four children were assessed simultaneously in the same room. The average number of children assessed per day ranged from four to 16. It was reported that, in two countries, the children did not require much support to complete the assessment, whereas they needed some support in the third country.

The NPMs indicated that the two-day schedule of the direct child assessments mostly worked well. However, in some instances, only the first part of the assessment could be completed as children were unable to attend the second part due to illness or other reasons. The NPMs furthermore reported that a few children found the assessment repetitive and did not want to return for the second day. Overall, Study Administrators informed the NPMs that they would have liked more flexibility with regards to the timing and duration of assessment breaks.

Regarding the functioning of the different components of the assessment, NPMs shared the Study Administrators' feedback that a few assessment items seemed too difficult or not clear enough for some children to complete. In these cases, the Study Administrators reminded the children of the instructions or repeated their explanations. While children enjoyed the assessment consistently across all three countries and across the different assessment domains, the Study Administrators' feedback reported by the NPMs also included that the assessment was a little too long as some children got tired toward the end of the assessment sessions.

As reported by the NPMs, the instructions on recording data on the tablet and uploading them to the OARS were clear. However, the NPMs reported that the participation reports exported from the OARS did not always show all the data they were expecting. In most countries, NPMs also reported issues with the completion of the Child Tracking Forms and Administration Report Form including incorrectly entered information and missing information. The NPMs reported that in some instances the assessment app froze or closed unexpectedly. Overall, the NPMs indicated that they were satisfied with the administration of the direct child assessments in their countries.

8.23. Administration of questionnaires (online and paper)

Most of the NPMs reported that they experienced difficulties in ensuring that centres/schools returned the completed Child and Staff Tracking Forms as centre/school staff forgot to return the forms, misplaced them, or did not have sufficient time to complete them before the holiday season. Regarding the parent questionnaire, two of the three countries experienced difficulties in achieving high participation. These difficulties were generally caused by the National Study Centres' reliance on Centre/School Co-ordinators to contact the parents. Two NPMs reported that it was difficult to achieve high participation from the staff members, too. The difficulties existed due to some staff members' feeling that completing the questionnaire was too time-consuming and that some questions in the questionnaires were unclear. However, evidenced by the relatively high final participation rates, the National Study Centres were able to overcome these difficulties, as can be seen in Table 13.1.

With regards to the administration of the online questionnaires, two of the three NPMs stated that there were issues with using the OARS for monitoring questionnaire participation, including missing or inaccurate information in the exported files and difficulty exporting the files. Paper questionnaires were either returned by respondents sending them individually to the National Study Centres or by Centre/School Co-ordinators collecting them. Overall, the feedback from all three NPMs about the administration of the online and paper questionnaires was positive.

8.24. Manual data entry and submission

As outlined in Chapter 12. the countries that participated in IELS were responsible for manually entering the data from the paper questionnaires into the IEA Data Management Expert (DME). Overall, NPMs reported that there were no major issues associated with this process.

8.25. Security arrangements

The NPMs confirmed that national data protection regulations affected the implementation of IELS in all three countries, for example requiring National Study Centres to anonymise the data. None of the NPMs were aware of any confidentiality

breaches affecting the assessment materials. All project staff who had access to confidential study materials signed confidentiality forms.

8.26. Conclusion

The IELS Consortium, in co-operation with the participating countries, developed and co-ordinated a range of quality assurance measures, including the International Quality Assurance Monitoring programme, National Quality Assurance Monitoring programmes, and the Survey Activities Questionnaire in order to monitor the quality of the study administration. Taken together, these activities form an important source of information about the implementation of the different steps and processes of IELS. Across all activities, the observations show that all three participating countries generally followed the standardised procedures as outlined in the IELS Technical Standards.

Notes

[5] One of the selected candidates withdrew from the IQAM programme after the training seminar. IEA Amsterdam trained the replacement IQAM remotely using video conferencing and email communication.

6 The numbers reported in this section refer to the Centre/School Co-ordinators that were interviewed during the IQAMs' 56 centre/school visits. Unless stated otherwise, reported numbers are out of a total of 56 responses recorded in the Centre/School Visit Records. Discrepancies in the total number of responses are caused by missing responses.

7 The United States combined the forms provided by the IELS Consortium into so-called "Child Assessment Grids".

Chapter 9. Survey weighting

This chapter describes why sampling weights are needed and explains the various weighting elements that account for differing selection probabilities and non-response. The weighting elements are then combined to produce a final child weight, which needs to be used for statistical analysis to ensure that any obtained estimates are unbiased and reliable. Additionally, information is provided for how to correctly interpret the results from the parent and staff questionnaires. Finally, the variance estimation method used for this study, Balanced Repeated Replication, is detailed. The replicate weights included in the data files always need to be used for calculating unbiased estimates of standard errors.

To correctly analyse the data gathered by IELS, it is necessary to use the sampling weights calculated by the sampling team at the IEA. The weights reflect the sampling design; base weights are computed as the inverse of sampling probabilities, and adjustments account for non-response. The final weights can be viewed as the estimated number of children represented by each participating sampled child.

The sampling design used for IELS aimed to be a self-weighting design; however, different factors made it impossible to achieve identical final weights for all children:

- Participating countries were free to choose oversampling of centres/schools in certain strata, causing unequal sampling probabilities for centres/schools across strata. In small strata, a specified minimum number of centres/schools were selected to ensure sufficient sample sizes even after non-response, also leading to disproportional sample allocation. Minor variations to sampling probabilities across strata were also caused by rounding.
- Sampling probabilities at centre/school level were based on the measure of size given in the centre/school sampling frame. This measure of size needed to be the latest available number of eligible children at the time of the frame preparation or, if this number was not available, an estimate of this number. However, as time elapsed between the production of the centre/school sampling frame (see Table 5.5 for details about the sampling frames) and the actual study administration, there were sometimes changes in the number of eligible children within selected centres, leading to deviation from a self-weighting design.
- Non-response leads to underrepresentation of specific centres/schools or children in the sample. This is corrected by non-response adjustments of the weights. If the non-response patterns differ across explicit strata, this will again cause a deviation from a self-weighting design.

The procedure implemented in IELS for calculating weights to achieve unbiased, reliable results has been used in other international studies in the area of education, among them PISA (OECD, 2017[14]), TALIS (OECD, 2019[15]) and TALIS Starting Strong (OECD, 2019[16]).

Table 9.1 lists the notation used in the formulae in this chapter for calculating the sampling weights.

Table 9.1: Conventional notation used in this chapter

Notation	Denotation
D_g	Number of centres/schools to sample per explicit stratum
enr_i	Number of enrolled children per centre/school
f_{1i}	Non-response adjustment at centre/school level
f_{2ij}	Non-response adjustment at child level
g	Index for explicit strata
i	Index for centres/schools
j	Index for children
I_g	Sampling interval
M_g	Number of centres/schools to sample for the main study per explicit stratum
MOS_i	Measure of size per centre/school
sam_i	Number of sampled children per centre/school
W_{ij}	Final weight for child j in centre/school i
w_{1i}	Base weight at centre/school level
w_{2ij}	Base weight at child level
$\lvert\Gamma_g\rvert$	Number of participating centres/schools per explicit stratum
$\lvert\Delta_i\rvert$	Number of participating children per centre/school
$\lvert X_i\rvert$	Number of sampled eligible children per centre/school
$\lvert\Omega_g\rvert$	Number of sampled eligible centres/schools per explicit stratum

9.1. Calculation of the weights

The final weights W_{ij} consist of four different factors: a centre/school base weight w_{1i}, a centre/school non-response adjustment f_{1i}, a child base weight w_{2ij} and a child non-response adjustment f_{2ij}.

9.2. Centre/school base weight

During the first stage of sampling, a sample of centres/schools was selected with probabilities proportional to size. The centre/school base weight is defined as the inverse of the sampling probability of the centre/school. The sampling probability is calculated as the measure of size MOS_i divided by the sampling interval I_g, as defined during the centre/school sample selection process (see Chapter 5. for details). However, if a centre/school had a measure of size larger than the sampling interval, it was selected with certainty. For small centres/schools with less than 15 expected children, the calculation of the centre/school base weight is based on the mean measure of size in the group of small centres/schools instead of the original one.

For each centre/school i in explicit stratum g, the centre/school base weight is calculated by:

$$w_{1i} = \begin{cases} \frac{I_g}{MOS_i} & \text{if } \mathrm{MOS_i} < \mathrm{I_g} \\ 1 & \text{otherwise} \end{cases} \tag{9.1}$$

In the case that the field test and the main study samples were selected at the same time, the sampling probability of a centre/school being sampled for the main study needs to be considered:

$$w_{1i} = \begin{cases} \frac{I_g}{MOS_i} \times \frac{D_g}{M_g} \text{ if } \mathrm{MOS_i} < \mathrm{I_g} \\ \frac{D_g}{M_g} \quad \text{otherwise} \end{cases} \tag{9.2}$$

D_g denotes the number of centres/schools to sample, while M_g denotes the number of centres/schools to sample for the main study. Since $\frac{D_g}{M_g}$ is equal to 1 if the field test and the main study samples were not selected at the same time, this formula can be used in general.

9.3. Centre/school non-response adjustment

Although replacement centres/schools were assigned for originally sampled centres/schools that refused to participate, there was still some non-response at centre/school level. For example, the originally sampled centre/school and both replacement centres/schools might have refused to participate, or there was no replacement assigned to a large centre/school.

The explicit strata were used as non-response adjustment groups, and the weights of non-participating centres/schools were distributed over participating centres/schools within the explicit stratum. For this purpose, the number of sampled eligible centres/schools $|\Omega_g|$ was divided by the number of participating centres/schools (including those centres/schools that replaced originally sampled centres/schools) $|\Gamma_g|$, for the explicit stratum g containing centre/school i:

$$f_{1i} = \frac{|\Omega_g|}{|\Gamma_g|} \tag{9.3}$$

Some centres/schools were found to be ineligible at the time of the study administration, for example because they were closed or because they did not have eligible children attending. These centres/schools were not considered in the centre/school non-response adjustment, i.e. the numerator in equation (9.3) above included only eligible centres/schools.

Only refusing eligible centres/school could be replaced; ineligible centres/schools were not replaced because they represented other ineligible centres/schools in the centre/school sampling frame.

9.4. Child base weight

Within each centre/school i, a sample of children was selected. The child base weight is calculated as the number of eligible children enr_i within each centre/school divided by the number of sampled children sam_i within the same centre/school:

$$w_{2ij} = \frac{enr_i}{sam_i} \tag{9.4}$$

The child base weight is the same for all sampled children within any one centre/school. The number of eligible children was often not the same number as the measure of size. This may have been because time had passed between the centre/school frame preparation and the child sampling, or because the measure of size was averaged for

small centres/schools, or the measure of size might only have been an estimate of the number of eligible children.

9.5. Child non-response adjustment

Non-response could also occur at child level since non-participating sampled children were not allowed to be replaced. For child j within-centre/school i, the child non-response adjustment is calculated by dividing the number of sampled eligible children per centre/school $|X_i|$ by the number of participating children per centre/school $|\Delta_i|$:

$$f_{2ij} = \frac{|X_i|}{|\Delta_i|} \tag{9.5}$$

Children who were ineligible to take part in the study or who were excluded were not considered in the adjustment. These include children who had since left the centre/school, who were not in the target age range, or who could not participate due to special education needs or limited experience in the language of the assessment.

Centres/schools with a participation rate of less than 50% of sampled children were treated as non-participating because the remaining participating children had a high risk of not being representative of the children in the whole centre/school.

9.6. Final weight

The final weight for each child j within-centre/school i is composed of the four elements:

$$W_{ij} = w_{1i} \times f_{1i} \times w_{2ij} \times f_{2ij} \tag{9.6}$$

For centres/schools with a measure of size smaller than the sampling interval, this translates into:

$$W_{ij} = \frac{I_g}{MOS_i} \times \frac{D_g}{M_g} \times \frac{|\Omega_g|}{|\Gamma_g|} \times \frac{enr_i}{sam_i} \times \frac{|X_i|}{|\Delta_i|} \tag{9.7}$$

For other centres/schools, the final weight is given by:

$$W_{ij} = \frac{D_g}{M_g} \times \frac{|\Omega_g|}{|\Gamma_g|} \times \frac{enr_i}{sam_i} \times \frac{|X_i|}{|\Delta_i|} \tag{9.8}$$

9.7. Using weights for data from staff and parents

Children were the only target population of IELS. Additionally, parents and staff were asked to complete questionnaires. All data arising from those questionnaires must be seen as features of the participating children, need to be weighted accordingly, and need to be interpreted with this perspective.

In other words, the final child weight also has to be used for statistical analysis of data from the parents or staff questionnaires. Data is not representative of a parent population, nor for a staff population, which is why special care has to be applied when interpreting the results.

For example, referring to data from the staff questionnaire, one could state: "The main contact person of X% of 5 year-old children is female." One cannot state: "X% of main contact persons of 5 year-old children are female." This is because there is no one-to-one match between children and staff.

Similarly, when referring to data on parental education, one could state: "X% of 5 year-old children have at least one parent with a university degree." One cannot say "X% of parents of 5 year-old children have a university degree." This is because the parental questionnaire could be completed by both or just one parent and parents can have more than one 5 year-old child (e.g. twins).

9.8. Replicate weights for estimating sampling variance

For studies with complex survey designs, special attention needs to be paid to the correct estimation of the sampling error. For IELS, as well as for PISA, TALIS and TALIS Starting Strong, Fay's variant of the Balanced Repeated Replication has been chosen for estimating sampling variance (Fay, 1989[17]).

Balanced Repeated Replication is a technique that uses subsamples of the realised sample, created in a special way by manipulating the weights. It compares estimates generated from the subsamples with the estimate generated from the whole sample to calculate the sampling error.

To create the replicates, the participating centres/schools from the sampling frame were sorted by explicit stratification, implicit stratification and measure of size, and then paired within each explicit stratum. If the number of participating centres/schools within an explicit stratum was uneven, the last three centres/schools of the explicit stratum constituted a triplet. Centres/schools that were selected with certainty were treated as explicit strata, and the children within these centres/schools were paired. Each pair is called a "pseudo stratum" or "zone", and is numbered sequentially. Within each zone, the first centre/school was randomly assigned a pseudo Primary Sampling Unit (PSU) number 1 or 2, and the second centre/school was assigned the other number. Also, for the third centre/school in a triplet, the pseudo PSU number was assigned randomly.

Table 9.2 shows the first seven centres/schools of an example country with their assigned zones and pseudo PSUs.

Table 9.2: Example of a random assignment of zones and pseudo PSUs

Explicit stratum	Centre/school ID	Zone = pseudo stratum	Pseudo PSU
1	1001	1	1
1	1002	1	2
1	1003	2	2
1	1004	2	1
2	1005	3	1
2	1006	3	2
2	1007	3	1
3	…	…	…

In the Fay's variant of the Balanced Repeated Replication, one of the two pseudo PSUs per zone will have its weight decreased, and the other will have its weight increased. To

know which pseudo PSU will be increased and which will be decreased, a Hadamard matrix is used. As an example, a Hadamard matrix of order 8 is written as:

$$Hadamard_8 = \begin{pmatrix} +1 & +1 & +1 & -1 & +1 & -1 & -1 & -1 \\ -1 & +1 & +1 & +1 & -1 & +1 & -1 & -1 \\ -1 & -1 & +1 & +1 & +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & +1 & -1 & -1 \\ -1 & +1 & -1 & -1 & +1 & +1 & +1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

This matrix gives the information for which pseudo PSU's weight to increase and which pseudo PSU's weight to decrease from each zone, by associating +1 with pseudo PSU 1 and -1 with pseudo PSU 2 when treating each row of the Hadamard matrix as one zone and each column as one replicate. The example Hadamard matrix of order 8 therefore translates into the following table that shows which pseudo PSU will have its weight increased:

Table 9.3: Example of which pseudo PSU's weight to increase per zone

Zone	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Replicate 6	Replicate 7	Replicate 8
1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2
2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2
3	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2
4	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2
5	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2
6	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2
7	Pseudo PSU 1	Pseudo PSU 1	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 1	Pseudo PSU 2
8	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2	Pseudo PSU 2

According to the Fay's variant, each child weight that is to be increased is multiplied by 1.5, and each child weight to be decreased is multiplied by 0.5. For children in triplets, this strategy was adapted by Judkins to ensure that the sum of the factors is equal to 3 (OECD, 2002[18]). In this case, the weights are multiplied by 1.7071 or 0.6464 if the weight of the single unit is to be increased, or they are multiplied by 1.3536 or 0.2929 if the weight of the single unit is to be decreased.

Table 9.4 shows the factors for the first seven centres/schools from the earlier example.

Table 9.4: Example of Balanced Repeated Replication factors

Centre /school ID	Zone	Pseudo PSU	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
1001	**1**	1	1.5	1.5	1.5	0.5	1.5	0.5	0.5	0.5
1002	**1**	2	0.5	0.5	0.5	1.5	0.5	1.5	1.5	1.5
1003	**2**	2	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5
1004	**2**	1	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5
1005	**3**	1	0.6464	0.6464	1.3536	1.3536	1.3536	0.6464	1.3536	0.6464
1006	**3**	2	1.7071	1.7071	0.2929	0.2929	0.2929	1.7071	0.2929	1.7071
1007	**3**	1	0.6464	0.6464	1.3536	1.3536	1.3536	0.6464	1.3536	0.6464

The replicate weights for the three participating countries are included in the data files; by using the IEA IDB Analyser (IEA, 2019[19]), these will be used automatically to correctly estimate the standard errors.

Chapter 10. Scaling IELS data

This chapter outlines the procedures undertaken to apply Item Response Theory (IRT) scaling and plausible value methodology to the data.

10.1. Study design and data yield

IELS used a combination of direct and indirect assessment of ten outcome domains. The ten outcome domains can be grouped under four main developmental areas: emergent literacy, emergent numeracy, self-regulation, and social and emotional skills. The direct assessments consisted of children responding to a series of items on a tablet-based app (see The 'Delivery Platforms' section in Chapter 3. . The direct assessments measured seven domains: emergent literacy, emergent numeracy, inhibition, mental flexibility, and working memory, and emotion identification and emotion attribution. All children in the study undertook direct assessment consisting of the same set of items, irrespective of the centre or country to which they belong (i.e. no rotation). Although assessment was conducted over two separate days, the order of assessments was the same each day. Assessments were conducted in the following order:

Day 1:

Emergent Literacy

Self-Regulation - Inhibition

Self-Regulation - Working memory

Self-Regulation - Mental flexibility

Day 2:

Emergent Numeracy

Empathy – Emotion Identification

Empathy - Emotion Attribution

The indirect assessment consisted of questionnaires for parents and educators designed to gather relevant information about each child in the study in three domains –prosocial behaviour, disruptive behaviour, and trust. Parents and educators were given the option of either completing the surveys via pen and paper or online.

The assessment design for IELS required participating countries to sample 200 centres/schools and 3 000 children, typically at least 15 children within each school/centre. If a school/centre had less than 15 children, then all children were selected. A stratified two-stage sampling design was utilised where centres/schools from participating countries were selected followed by children within each school/centre. Centres/schools were selected using systematic random sampling with probability proportional to size (PPS). More detail on the sampling framework is provided in Chapter 5. . Details of sampling outcomes including unweighted and weighted participation rates for centres/schools, Children, Parents and Educators can be seen in Chapter 13. . A child was considered to have been a respondent in the study if they had provided at least one valid answer in at least two assessment domains. Table 10.1 details the number and proportion of children and educators who responded to at least one

scored item from each domain and those who responded to all scored items within each domain.

For the domains of prosocial behaviour, disruptive behaviour and trust response data from educators was used to yield children's abilities in these domains. There is complimentary parent-level data, that is, parents responded to the same items as the educators. There are factor scores available, generated using parent-responses, in the international data set. These parent-responses were included in the latent regression (population) model. See the section below titled '*Population model'* for a discussion. The implication is that educator-response data (and indeed all 10 IELS outcome domains) is conditioned on the parent-ratings of children's social and emotional skills. Analysis that includes these parent-rated factor scores as a covariate (e.g. one of the 10 IELS outcome domains regressed on parent-ratings of social-emotional skills) will yield correct population estimates of the relationship between Y and X. The parent-ratings of children's social and emotional skills, however, if regressed on another covariate will yield potentially biased estimates of the relationship between Y and X as these parent-ratings of social and emotional skills were not conditioned on the other information in the IELS data set. The use of conditioning variables and implications for secondary analysts is described in more detail in 'Constructing conditioning variables' within the 'Multi-dimensional scaling of 10 latent dimensions (domains) in IELS and generating plausible values' section.

Table 10.1: Data yield per domain by country

Domain	Country	Children who completed at least one item		Children who completed all items[8]	
		N	%	N	%
Literacy	**England**	2 556	99.2	2281	88.5
	Estonia	2 095	99.3	1695	80.3
	United States	2213	99.1	1625	72.7
	Total	6864	99.2	5601	80.9
Numeracy	**England**	2493	96.7	2137	82.9
	Estonia	1984	94.0	1428	67.7
	United States	2082	93.2	1586	71.0
	Total	6559	94.8	5151	74.4
Inhibition	**England**	2189	84.9	1254	48.7
	Estonia	1770	83.9	1284	60.9
	United States	1595	71.4	1013	45.3
	Total	5554	80.2	3551	51.3
Mental flexibility	**England**	2494	96.8	530	20.6
	Estonia	2052	97.3	511	24.2
	United States	2081	93.2	265	11.9
	Total	6627	95.8	1306	18.9
Working memory	**England**	2531	98.2	651	25.3
	Estonia	2065	97.9	568	26.9
	United States	2147	96.1	360	16.1
	Total	6743	97.4	1579	22.8
Emotion identification	**England**	2474	96.0	2324	90.2
	Estonia	1977	93.7	1891	89.6
	United States	2046	91.6	1913	85.6
	Total	6497	93.9	6128	88.5
Emotion attribution	**England**	2470	95.8	1881	73.0
	Estonia	1964	93.1	1383	65.5
	United States	2035	91.1	1135	50.8
	Total	6469	93.5	4399	63.6
Prosocial behaviour	**England**	2310	89.6	2211	85.8
	Estonia	1994	94.5	1888	89.5
	United States	2128	95.3	1869	83.7
	Total	6432	92.9	5968	86.2
Disruptive behaviour	**England**	2310	89.6	2271	88.1
	Estonia	1994	94.5	1951	92.5
	United States	2128	95.3	1979	88.6
	Total	6432	92.9	6201	89.6
Trust	**England**	2309	89.6	2220	86.1
	Estonia	1994	94.5	1904	90.2
	United States	2128	95.3	1823	81.6
	Total	6431	92.9	5947	85.9

Note. Values for prosocial behaviour, disruptive behaviour and trust are based on responses from educators.

More detail on participation rates can be found in Chapter 13. .

10.2. The use of practice blocks and stop rules within assessments

The three domains within self-regulation (i.e. inhibition, working memory, mental flexibility) were conducted using practice and assessment item blocks along with stop rules. Children who did not meet the minimum threshold defined by the stop rules were stopped from taking the remainder of the assessment. Separate rules were applied for each domain within self-regulation. Table 10.2 provides an overview of the number of children who undertook questions from each of the assessment phases for the three domains within self-regulation.

Table 10.2: Number and percentage of children undertaking each phase of assessment within self-regulation domains

Assessment phase	Inhibition		Working memory		Mental flexibility	
	N	%	N	%	N	%
Practice 1	6921	100	6921	100	6921	100
Practice 2	2869	41.5	3877	56.0	1165	16.8
Testlet 1	5568	80.5%	4071	58.8	6158	89.0
Testlet 2	5374	77.6%	2722	39.3	3978	57.5
Testlet 3	-	-	1597	23.1	3320	48.0

Note. Children who did not meet the minimum threshold in Practice 1 were required to do Practice 2, while the remainder progressed straight to Testlet 1.

The inclusion of a missing-by-design structure is similar to other assessments that use rotated booklets, or adaptive designs. Although IELS does not use a rotated or complex booklet design and therefore has an almost complete response matrix (i.e. the m x n matrix of m children's scored responses to n items in IELS) due to all children having the opportunity to complete every item, a missing-by-design structure is introduced in the way that stop rules are employed in the self-regulation domains (see Table 10.2 for response rates for each block of items within self-regulation domain). There is a significant literature exploring the effect of such designs on the estimation of item parameters. This analytic work asks the question of whether it is possible to make unbiased estimates of item difficulty if there are potential missing data structures that lead to higher or lower ability children (or other relevant subpopulations) being excluded from completing items. In the IELS cases, the inclusion of stop rules introduced missing data that varied as a function of ability. That is, lower ability children tended to stop earlier and those children had higher rates of non-response by design. Like in other missing data problems, the analysis of this in the IRT case is a matter of assessing whether the missing data structure is ignorable: that is, that under a single international calibration, the item parameter estimates are unbiased (Rutkowski, 2011[20]).

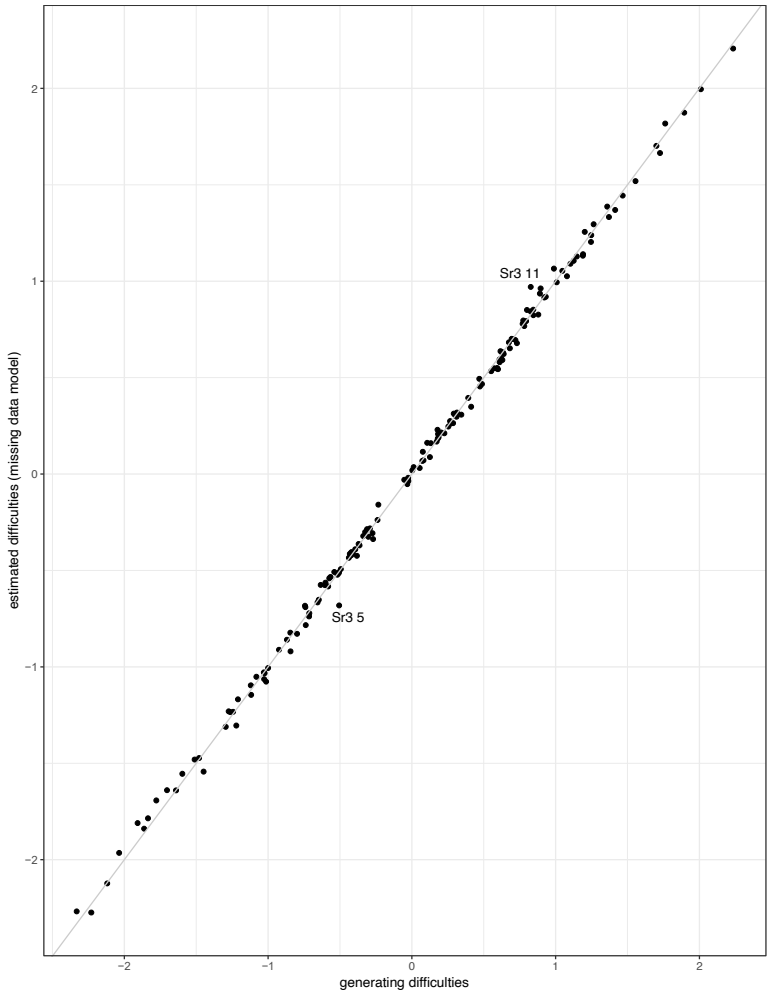
Traditional approaches where subscales are calibrated separately using a marginal maximum likelihood (MML) estimator yields biased estimates of item parameters when there is a missing-by-design response structure. When, instead, calibrations are conducted simultaneously across subscales (i.e. a multi-dimensional model) the ignorability assumption (see e.g. Rosenbaum & Rubin (1983[21])) is met and in turn a

single international calibration is sufficient ((Rutkowski, 2011[20]) (Wang, Chen and Jiang, 2019[22])).

The effect of this missing-by-design structure was also assessed empirically by simulation. The purpose of this simulation was to test the quality of the recovery of known parameters. That is, data was simulated using the conquestr (Cloney, 2019[23]) library in R (R Core Team, 2016[24]) and ACER ConQuest (Adams and Wilson, 2020[25]) from a known multivariate normal distribution and recovery of the underlying "true" values was assessed. A single data set was generated with complete responses. Item responses were generated for the same number of items used in the IELS study (however, generated such that the item responses are dichotomous, fit the 1PL model, and are generated from known distributions of both children's abilities and item difficulties). Preliminary runs from analysis of IELS (field trial) data were used to generate credible parameters for the simulation (variances and covariance's). All means were set to zero. Missing data was added in two ways: (1) under the missing completely at random (MCAR) assumption to the whole data set so that approximately 2% of responses are missing, and (2) the missing data structure of the Self-Regulation (SR) domains in IELS was recreated. That is, stop rules were simulated such that children with higher raw scores completed more items (as they met the thresholds for continuing the assessment).

The results of the simulation show excellent recovery of the underlying item parameters. Only 2 items show more than a 0.1 logit absolute difference from their generating values – both simulated working memory items. Note that the absolute difference between these items is less than 0.175 logits – well below a level, for example, where concerns would be raised in a DIF analysis. These items are both not from any particular block (e.g. the block with the most missing data), and were in both directions (over- and under-estimates of the generating difficulty) and are likely simply perturbations due to unreliability (see Figure 10.1).

Figure 10.1: X-Y scatter of generating and estimated item parameter estimates from simulation of missing-by-design response structure.



10.3. Response timing

Using the tablet-based assessment application, all items were timed in order to collect data about the time taken to respond to each item, and to collect data about the length of the assessment of each domain. The target length for assessment of each major domain in the main study was 15 minutes (Chapter 4.). The set of items that were included in the main study did not present any issues regarding response time and the assessments met the criteria for the time taken to complete the assessment.

In the main study, response times in the self-regulation domains of inhibition and mental flexibility were also used in the scoring of the items. The response time cut-off criteria were selected following an assessment of the distribution of response times and selecting times that produced approximately equal groups, while also attending to the distribution of child proficiencies in the created groups (under the expectation that children's cognitive skill or response time is negatively correlated with their executive function skills: children who take longer to respond to stimulus are expected to have lower inhibitory skills) (Floyd et al., 2007[26]). This was done by secondary checking,

ensuring that proficiencies within those selected groups were ordered in increasing magnitude.

For inhibition, the following scoring rules were applied:

1. Incorrect response or correct response that took longer than 1.3 seconds
2. Correct response in 1.05 to 1.3 (inclusive) seconds
3. Correct response in 0.9 to 1.05 (inclusive) seconds
4. Correct response that took less than 0.9 (inclusive) seconds

For mental flexibility, the following scoring rules were applied:

5. Incorrect response
6. Correct response that took longer than 2.5 seconds
7. Correct response that took less than 2.5 (inclusive) seconds

Average reaction times to non-scored (i.e. non-switching) inhibition items were used to develop a proxy measure of processing speed. This was included as a regressor in the population model (see 'Constructing conditioning variables' within the 'Multi-dimensional scaling of 10 latent dimensions (domains) in IELS and generating plausible values' section). Summary statistics of this measure for each country is provided in Table 10.3.

Table 10.3: Summary statistics or response times by country

Country/economy	Mean (secs)	SE	Median (secs)
England	1.140	0.004	1.092
Estonia	0.995	0.004	1.017
United States	0.990	0.004	1.042

Results of a regression analysis of the three self-regulation domains regressed on the measure of processing speed are provided in Table 10.4.

Table 10.4: Regression of self-regulation domains on processing speed

Country	Domains	Intercept			Slope			R^2
		B	SE	t	B	SE	t	
England	**Inhibition**	888.34	9.61	92.49	-375.35	8.49	-44.19	0.71
	Mental flexibility	636.01	12.78	49.77	-107.85	10.51	-10.26	0.05
	Working memory	585.47	10.88	53.79	-60.85	9.12	-6.67	0.02
Estonia	**Inhibition**	880.16	12.37	71.16	-362.92	11.90	-30.50	0.52
	Mental flexibility	609.11	13.43	45.34	-99.20	12.92	-7.68	0.04
	Working memory	609.03	16.79	36.26	-89.14	16.59	-5.37	0.04
United States	**Inhibition**	801.83	12.08	66.37	-283.25	11.39	-24.87	0.40
	Mental flexibility	564.13	10.24	55.11	-88.19	9.89	-8.92	0.04
	Working memory	513.53	12.75	40.27	-50.29	12.24	-4.11	0.01

Note. Each row represents a separate regression.

Processing speed, measured by average response time, is negatively related to the self-regulation domains. The magnitude of the relationship is near to one standard deviation in mental flexibility, stronger in inhibition, and weaker in working memory. The relationship is particularly strong for inhibition on processing speed. This is likely due to the measure of processing speed being embedded in the inhibition task in addition to the theorised relationship between the cognitive skills of inhibition and processing speed.

10.4. The IRT models for scaling

In large-scale assessment (LSA) design the number of items that are used to capture information relating to each domain of interest is done so with population-level statistics in mind. Because the focus is on population parameter estimates (and not the ability of individual students) the point is often made that the study design may produce unreliable estimates for the ability of individual students while still producing highly reliable and accurate estimates of population parameters ((Adams, 2005[27]); (Wu, 2005[28])). In addition to this, in studies like IELS, background and contextual information can be added to the information provided by the item responses to improve the recovery of 'true' population parameters. This combination of item responses and background variables, along with the relationships between domains, is used to produce plausible values that represent accurate estimates of the distribution of population parameters. Despite the use of a single booklet design in IELS, adopting PV methodology is still highly relevant. The purpose of population surveys is to recover estimates of the true population parameter of interest (e.g. the average level of working memory). A number of studies have shown that while other approaches to generating factor scores (e.g. Expected A-Posteriori (EAP) methods) yield consistent and unbiased estimates of the point estimate, they fail to recover unbiased estimates of the variance of the parameter ((Adams, 2005[27]); (Wu, 2005[28])). Accurate recovery of the variance of the parameter of interest is important for secondary researchers whom often will want to know about the significance or relative magnitude of a relationship and these rely on not only the absolute value of a parameter (e.g. a mean) but its sampling distribution. PVs allow secondary analysis to be conducted in commonly used software such as SPSS or (Gebhardt and Berezner, 2017[29]).

IELS data is modelled in ACER ConQuest Version 5 (Adams and Wilson, 2020[25]). In IELS, the initial models specified are the most parsimonious – e.g. 1 Parameter Logistic (1PL) – with more complex models (e.g. 2PL) considered where misfit of data to the models is detected. For descriptions of the underlying models see, for dichotomous data, the Rasch model (Rasch, 1960[30]), for polytomous data, the Partial Credit Model (PCM) (Masters, 1982[31]). ACER ConQuest (Adams and Wilson, 2020[25]) fits a general model - the Multi-dimensional Random Coefficients Multinomial Logit Model (Adams, Wilson and Wang, 1997[32]) – allowing the consideration of many different approaches to parametrisation of the final models in IELS. Further discussion of resultant models (1Pl vs 2PL) is provided in 'International item calibration' section.

10.5. The multi-dimensional random coefficients multinomial logit model

IELS data is scaled using best practice guidelines suggested by Adams, Wilson and Wu (1997[33]), Glass and Verhelst (1995[34]), Mislevy and Sheehan (1987[35]), and Yamamoto and Mazzeo (1992[36]). The methods used in IELS are developed from IRT models, that have previously been shown to be effective in the analysis of large scale, nested, categorical data ((Skrondal, 2004[37]); (Von Davier, (2007).[38])). The models used in IELS are consistent with other LSAs conducted by OECD, including PISA ((OECD, 2014[39]) (OECD, 2017[14])).

In IELS the MRCMLM is constrained to fit models equivalent to the Partial credit Model (PCM). The PCM is a generalisation of the dichotomously scored 1PL model (Rasch Model).

The Rasch model estimates item difficulty parameters, which is the probability of a child responding correctly to a specific item dependent upon their location on the domain. The probability of selecting category 1 instead of 0 is modelled as:

$$P_i(\theta_n) = \frac{e^{(\theta_n - \delta_i)}}{1+e^{(\theta_n - \delta_i)}} \tag{10.1}$$

where $P_i(\theta_n)$ is the probability of person n to score 1 on item i. θ_n is the estimated latent trait of person n and δ_i the estimated location of item i on this dimension. A probability of 0.50 is the point at which the ability of a child equals the difficulty of the item.

In the case of items with more than two categories ($k > 1$) (as for example with Likert-type items) this model can be generalised to the PCM which takes the form of:

$$P_{xi}(\theta_n) = \frac{e^{(\sum_{k=0}^{x}(\theta_n - \delta_i + \tau_{ij})}}{\sum_{h=0}^{m_i} e^{\sum_{k=0}^{h}(\theta_n - \delta_i + \tau_{ik})}} \tag{10.2}$$

where $P_{xi}(\theta_n)$ denotes the probability of person n to score x on item i out of the m_i possible scores on the item. θ_n denotes the person's latent ability, the item parameter δ_i gives the location of the item on the latent continuum and τ_{ij} denotes an additional step parameter.

To apply this model of the probability of responding to a given item and to estimate the parameters (theta, delta, tau), the following is estimated using the MRCMLM. Note that the MRCMLM is a multi-dimensional extension of the random coefficients multinomial logit model (RCMLM), and thus the simplest description of this model is where there is one dimension, D, and therefore the list of dimensions is a scalar, rather than a vector.

For the RCMLM, item parameters are a fixed set of unknown parameters, $\boldsymbol{\xi}$, and latent achievement, $\boldsymbol{\theta}$, is a random effect.

I items are indexed $i = 1,\ldots,I$ with Ki 1 response categories indexed k 0,1,…,Ki. The random variable (vector-valued) $\boldsymbol{X}i = (Xi1, Xi2, \ldots, Xik_i)'$, where:

$$X_{ij} = \begin{cases} 1 \text{ if response to item } i \text{ is in category } j \\ 0 \quad \text{otherwise} \end{cases} \tag{10.3}$$

indicates the $Ki + 1$ possible responses to item i.

The zero vector (vector filled with zeroes) is used as the arbitrary reference category. The response vector (or pattern) is $\boldsymbol{X}' = (\boldsymbol{X}_1', \boldsymbol{X}_2', \ldots, \boldsymbol{X}_I')$. Each instance of these random variables are indicated by: $\boldsymbol{x}$, $\boldsymbol{x}_i$ and x_{ik}.

A vector of the items with p parameters is expressed by $\boldsymbol{\xi}' = (\xi_1, \xi_2, \ldots, \xi_p)$, where linear combinations are used in the response model, helping to explain the empirical characteristics of each item's response categories. A design matrix can be created from a set of design vectors $\boldsymbol{a}_{ij}, (i = 1, \ldots, I;\ k = 1, \ldots, K_i)$, each of length p, in the form $\boldsymbol{A} = (\boldsymbol{a}_{11}, \boldsymbol{a}_{12}, \ldots, \boldsymbol{a}_{1K_1}, \boldsymbol{a}_{21}, \ldots, \boldsymbol{a}_{2K_2}, \ldots, \boldsymbol{a}_{1K_n})$. This describes the relationships between items and model parameters.

A scoring function is then added to the RCMLM to allow for the specification of proficiency for each item response. A response score b_{ik} represents the proficiency for in each category k of item i, which can then be collected to form a vector,

$\boldsymbol{b}' = (b_{11}, b_{12}, \ldots, b_{1K_1}, b_{21}, \ldots, b_{2K_2}, \ldots, b_{1K_n})$.

For a latent variable θ, the item response probability for the RCMLM is

$$P(\boldsymbol{X}_{ik} = 1; \boldsymbol{A}, \boldsymbol{b}, \boldsymbol{\xi} | \theta) = \frac{e^{(b_{ik}\theta + \boldsymbol{a}_{ik}'\boldsymbol{\xi})}}{\sum_{k=1}^{K_i} e^{(b_{ik}\theta + \boldsymbol{a}_{ik}'\boldsymbol{\xi})}} \tag{10.4}$$

and a response vector probability model as

$$P(\boldsymbol{X} = \boldsymbol{x} | \theta) = \boldsymbol{\Psi}(\theta, \boldsymbol{\xi}) e^{[\boldsymbol{x}'(\boldsymbol{b}\theta + \boldsymbol{A}\boldsymbol{\xi})]} \tag{10.5}$$

with

$$\boldsymbol{\Psi}(\theta, \boldsymbol{\xi}) = \left\{ \sum_{z \in \boldsymbol{\Omega}} e^{[\boldsymbol{z}'(\boldsymbol{b}\theta + \boldsymbol{A}\boldsymbol{\xi})]} \right\}^{-1} \tag{10.6}$$

where $\boldsymbol{\Omega}$ is the set of all possible response vectors.

The MRCMLM, a multi-dimensional extension of the RCMLM, assumes that a set of D latent traits underlie the children's responses. Children's positions in the D dimensional latent space can be represented by vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_D)$. In the RCMLM, the scoring function of k (response category) in i (item) corresponded to a scalar, where in the MRCMLM, it corresponds to a D x 1 column vector. Each response in k on dimension $d(d = 1, \ldots, D)$ of i is then scored as b_{ikd}. All scores across D dimensions can then form a column vector $\boldsymbol{b}_{ik} = (b_{ik1}, b_{ik2}, \ldots, b_{ikD})$. These column vectors can then form a scoring submatrix for each i, $\boldsymbol{B}_i = (\boldsymbol{b}_{i1}, \boldsymbol{b}_{i2}, \ldots, \boldsymbol{b}_{iD})'$, and subsequently a scoring matrix $\boldsymbol{B} = (\boldsymbol{B}_1', \boldsymbol{B}_2', \ldots, \boldsymbol{B}_n')'$ for the entire assessment. The probability of a response in k of i is:

$$P(\boldsymbol{X}_{ik} = 1; \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\xi} | \boldsymbol{\theta}) = \frac{e^{(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}_{ik}'\boldsymbol{\xi})}}{\sum_{k=1}^{K_i} e^{(\boldsymbol{b}_{ik}\boldsymbol{\theta} + \boldsymbol{a}_{ik}'\boldsymbol{\xi})}} \tag{10.7}$$

for response vector

$$f(\boldsymbol{x};\boldsymbol{\xi}|\boldsymbol{\theta}) = \boldsymbol{\Psi}(\boldsymbol{\theta},\boldsymbol{\xi})\boldsymbol{e}^{[\boldsymbol{x}'(\boldsymbol{B\theta}+\boldsymbol{A\xi})]} \tag{10.8}$$

with

$$\boldsymbol{\Psi}(\boldsymbol{\theta},\boldsymbol{\xi}) = \left\{\sum_{z\in\Omega} e^{[\boldsymbol{z}'(\boldsymbol{B\theta}+\boldsymbol{A\xi})]}\right\}^{-\mathbf{1}}. \tag{10.9}$$

10.6. The population model

The item response model is a conditional model. To fully specify the model, therefore, the distribution of the latent trait, $\boldsymbol{\theta}$ must be defined. The density function for the latent variable, $\boldsymbol{\theta}$ is specified as $f_\theta(\boldsymbol{\theta};\ \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is the set of parameters that represents the distribution of $\boldsymbol{\theta}$**.** Under simple, unidimensional, models it is common to assume that children are sampled from a normal population with mean μ and variance σ^2, where:

$$f_\theta(\boldsymbol{\theta};\ \boldsymbol{\alpha}) \equiv f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\, e^{[-\frac{(\theta-\mu)^2}{2\theta^2}]} \tag{10.10}$$

or equivalently

$$\theta = \mu + E \tag{10.11}$$

where $E \sim N(0,\sigma^2)$.

In LSAs, there is often a rich variety of background and conditional variables available and it is possible to replace μ with a regression model, $\boldsymbol{Y}_n^T\beta$ (Adams, Wilson and Wu, 1997[33]). Where $\boldsymbol{Y}_n$ is a vector of u values for child n (e.g. gender, socio-economic status). β is the vector of regression coefficients. The latent regression (population) model for child n becomes:

$$\theta_n = \boldsymbol{Y}_n^T\boldsymbol{\beta} + E_n \tag{10.12}$$

with the assumption that E_n is *iid* with mean zero and variance θ_n which is equivalent to:

$$f_\theta(\theta_n;\boldsymbol{Y}_n,b,\sigma^2) = (2\pi\sigma^2)^{-1/2}\, e^{[-\frac{1}{2\sigma^2}(\theta_n-\boldsymbol{Y}_n^T\boldsymbol{\beta})^T(\theta_n-\boldsymbol{Y}_n^T\boldsymbol{\beta})]}. \tag{10.13}$$

Extending this to a multi-dimensional model (e.g. now drawing a sample form a conditional (on the regression parameters) multivariate normal distribution) results in the multivariate population model:

$$f_\theta(\boldsymbol{\theta}_n;\boldsymbol{w}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma}) = (2\pi)^{-d/2}\,|\boldsymbol{\Sigma}|^{-1/2}\, e^{[-\frac{1}{2}(\boldsymbol{\theta}_n-\boldsymbol{\gamma w}_n)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_n-\boldsymbol{\gamma w}_n)]} \tag{10.14}$$

where $\boldsymbol{\gamma}$ is a $u \times D$ matrix of regression coefficients, $\boldsymbol{\Sigma}$ is a $D \times D$ variance-covariance matrix, and $\boldsymbol{W}_n$ is a $u \times 1$ vector of fixed variables. In IELS, the $\boldsymbol{W}_n$ variables are referred to as conditioning variables.

10.7. Combined model

The conditional item response model described in 'The multi-dimensional random coefficients multinomial logit model' (Figure 10.3) and latent regression (population) model in 'Population model' (10.14) are now combined to produce the unconditional item response model:

$$f_x(\boldsymbol{x};\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}) = \int_{\boldsymbol{\theta}} f_x(\boldsymbol{x};\boldsymbol{\xi}|\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta};\boldsymbol{\gamma},\boldsymbol{\Sigma}) d\boldsymbol{\theta} \tag{10.15}$$

where the parameters of the model are $\boldsymbol{\gamma}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\xi}$.

The estimation procedures used in IELS follow the guidelines recommended by Adams, Wilson and Wu (1997[33]), Adams, Wilson and Wang (1997[32]), and Wu, Adams and Wilson (1997[40]).

Although the locations of individuals on the latent variables cannot be estimated under this model, it is possible to specify a posterior distribution for the latent variable, given by:

$$h_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{w}_n,\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma}|\boldsymbol{x}_n) = \frac{f_x(\boldsymbol{x}_n;\boldsymbol{\xi}|\boldsymbol{\theta}_n) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{w}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})}{f_x(x_n;\boldsymbol{w}_n,\boldsymbol{\xi},\boldsymbol{\gamma},\boldsymbol{\Sigma})}$$
$$= \frac{f_x(\boldsymbol{x}_n;\boldsymbol{\xi}|\boldsymbol{\theta}_n) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{w}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma})}{\int_{\boldsymbol{\theta}_n} f_x(\boldsymbol{x}_n;\boldsymbol{\xi}|\boldsymbol{\theta}_n) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{w}_n,\boldsymbol{\gamma},\boldsymbol{\Sigma}) d\boldsymbol{\theta}} \tag{10.16}$$

10.8. Generating plausible values

In IRT scaling, measures of individual performance are not observed, instead treated as missing data that can be inferred from item responses. IELS uses imputation methods to compute plausible values (PVs). PVs are approximations of child proficiency. For each domain in IELS, five plausible values for each child are computed.

Item parameters are anchored at the values estimated in the calibration step. Plausible values are random draws from the posterior distribution of each child's latent distribution, in line with guidelines specified by Mislevy (1991[41]) and Mislevy et al. (1992[42]).

In IELS, M vector-valued random deviates, $\{\boldsymbol{\varphi}_{mn}\}_{m=1}^{M}$ are drawn from the multivariate normal distribution defined in the population model, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}_n;\boldsymbol{W}_n\boldsymbol{\gamma},\boldsymbol{\Sigma})$, for each case n. Following this, they are used to approximate the integral in the denominator of the posterior distribution of the latent variable using the Monte-Carlo integration:

$$\int_{\theta} f_x(\boldsymbol{x};\xi|\theta) f_{\theta}(\theta;\gamma,\Sigma,\boldsymbol{W})\, d\theta \approx \frac{1}{M}\sum_{m=1}^{M} f_x(\boldsymbol{x};\xi,\boldsymbol{\varphi}_{mn}) \equiv \mathfrak{J} \tag{10.17}$$

The values

$$p_{mn} = f_x(x_n;\boldsymbol{\xi}|\boldsymbol{\varphi}_{mn}) f_{\theta}(\boldsymbol{\varphi}_{mn};\boldsymbol{W},\boldsymbol{\gamma},\boldsymbol{\Sigma}) \tag{10.18}$$

are calculated, obtaining the set of pairs $\left(\boldsymbol{\varphi}_{mn}, {}^{p_{mn}}/_{\mathfrak{Z}}\right)_{m=1}^{M}$ (i.e. an approximation of the posterior density). The probability that φ_{nj} is drawn from this density is:

$$q_{nj} = \frac{p_{mn}}{\sum_{m=1}^{M} p_{mn}} \tag{10.19}$$

For each plausible vector, L uniformly distributed random numbers $\{\eta_i\}_{i=1}^{L}$are produced. For each random draw, the vector, φ_{ni_0}, is selected as a plausible vector assuming that the following condition is satisfied:

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i < \sum_{s=1}^{i_0} q_{sn} \tag{10.20}$$

10.9. Analysis of data with plausible values

In the previous section, a brief outline of the method used to generate plausible values was outlined. This section describes the process of computing plausible values in more detail and the application of how they should be used in secondary analysis to yield unbiased parameter estimates.

It has previously been stated that plausible values should not be viewed as individual proficiency scores. PVs are random draws from the posterior distribution of the latent variable (equation (10.20), contain random error variance components and are therefore not well suited to be viewed as scores for individuals. PVs are typically used in population-level statistics. The approach developed by Mislevy and Sheehan (1987[35]), based on Rubin's (1987[43]) imputation theory, produces adequate estimates of population parameters.

In IELS, computation of five plausible values was undertaken. Any population-level analyses is subsequently done five times, one for each PV, then aggregated with significance tests adjusting for variation.

$r(\boldsymbol{\theta}, \boldsymbol{Y})$ is a statistic that with $(\boldsymbol{\theta}, \boldsymbol{Y}) = (\theta_1, y_1, \theta_2, y_2, \ldots, \theta_N, y_N)$ where (θ_n, y_n) are the values of the latent variable and the other observed characteristic for child n. Although θ_n is not observed, we observe the item responses, x_n, and construct for each child n, the marginal posterior $h_\theta(\boldsymbol{\theta}_n; y_n, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}|\boldsymbol{x}_n)$. If $h_\theta(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}|\boldsymbol{X})$ is the joint marginal posterior for $n = 1, \ldots, N$ then:

$$\begin{aligned} r^*(\boldsymbol{X}, \boldsymbol{Y}) &= E[r^*(\boldsymbol{\theta}, \boldsymbol{Y})|\boldsymbol{X}, \boldsymbol{Y}] \\ &= \int_{\boldsymbol{\theta}} r(\boldsymbol{\theta}, \boldsymbol{Y}) h_\theta(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}|\boldsymbol{X}) d\boldsymbol{\theta} \end{aligned} \tag{10.21}$$

This integral can be computed using the Monte-Carlo method. If M random vectors $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$ are drawn then the integral is approximated by:

$$\begin{aligned} r^*(\boldsymbol{X}, \boldsymbol{Y}) &\approx \frac{1}{M} \sum_{m=1}^{M} r(\boldsymbol{\theta}_m, \boldsymbol{Y}) \\ &= \frac{1}{M} \sum_{m=1}^{M} \hat{r}_m \end{aligned} \tag{10.22}$$

where $\hat{r}_m$ is the estimate of r which is computed using the m-th set of PVs.

Therefore, the final estimate of r is the average of the estimates from the computation using each vector that is randomly drawn. Suppose that U_m is the sampling variance for $\hat{r}_m$. The sampling variance of r^* is then:

(10.23)

$$V = U^* + (1 + M^{-1})B_M$$

where $U^* = \frac{1}{M}\sum_{m=1}^{M} U_m$ and $B_M = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{r}_m - r^*)^2$.

An α-% confidence interval for r^* is $r^* \pm t_v\left({}^{(1-\alpha)}/_{2}\right) v^{1/2}$ where $t_v(s)$ is the s-percentile of the t-distribution with v degrees of freedom, with $v = \left[\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d}\right]^{-1}$, $where\ f_M = (1 + M^{-1})B_M/V$ and d is the degree of freedom applied had θ_n been observed. In IELS, d will vary by country and have a maximum possible value of 92.

10.10. Application of the IRT model to IELS

IELS assesses children's responses to items that are theoretically hypothesised to belong to ten domains. In this way, these sub-domains can be thought of as ten different dimensions to be included in modelling. Each dimension could be modelled using a series of separate unidimensional models or one MRCMLM. In IELS, the MRCMLM was used in two steps: national and international item calibrations, and national conditioning and production of plausible values. International item calibration was undertaken using the conditional item response model (equation (10.7) in conjunction with the population model (equation (10.18) but conditioning variables were not used. That is, it is assumed that children have been sampled from a multivariate normal distribution

Following this, an assessment of item fit and Differential Item Functioning was undertaken to consider the quality of the item parameter estimates as international anchors in the subsequent scaling/conditioning step.

Then, a multi-dimensional model was run in each country, with item parameters anchored to values estimated in the calibration step (less any item parameters declared to be unsuitable for international anchoring) and a full population model specified to yield plausible values.

Finally, model checking is undertaken to assess the quality of estimates, including multiple runs to confirm parameter estimate stability as well as estimation of the model using alternate methods to confirm the optimal model solution. Bayesian estimation using Markov Chain Monte-Carlo (MCMC) was undertaken and the results compared with traditional MML using Monte-Carlo results. Results using both estimators were obtained using ConQuest Software (Adams and Wilson, 2020[25]). To do this, a number of comparisons were run, including comparisons of the estimated item parameters, stability of the parameters of the model over iterations, the unconditional correlation matrices, and correlations amongst the ability estimates generated. More details of the outcomes are provided in Chapter 11.

10.11. National item calibration

National calibrations were performed separately, country by country, using unweighted data. The results of these analyses were used to monitor the quality of the data and to

make decisions regarding national item treatment. When reviewing the national calibrations, particular attention was paid to the fit of the items to the scaling model, item discrimination and item-by-country interactions.

Classical test theory summary statistics were computed for all items within all sub-domains for each country. All descriptive statistics were initially provided for all observed responses as well as the various missing response codes and were shared with participating countries and the OECD. Not only is information relating to correct and incorrect scores examined, statistics for all response categories are used to evaluate how each item operates across subgroups (e.g. country, language, gender, age). The following descriptive statistics for observed and missing responses were computed;

- item difficulties (proportion of correct responses)
- frequency of scores (all response categories for each item and missing responses)
- index of discrimination (point-biserial correlations, item-total and item-rest correlation)

The statistics described above were computed for each participating country and internationally. This process allowed for the identification of potentially problematic items (e.g. items where very high proportions of participants get items right/wrong, items that are not positively correlated with the underlying trait (item-total and item-rest correlation), and response categories that do not discriminate between incorrect and correct (point-biserial correlation). Although specific cut-off values were not strictly adhered to for item exclusion, guiding values of <0.2 for item-total and item-rest correlation were applied, as were checks for point-biserial correlations in the "correct" direction (e.g. generally negative and significant for the zero or incorrect category and positive for the most-correct (in the case of partial credit items) response category to identify items that may cause problems.

Table 10.5 provides an example output (direct from ConQuest (Adams and Wilson, 2020[25])) used for the examination of item response distributions from the emergent numeracy domain. The example question asked children to undertake a simple mathematical operation (ten subtract two) expressed as a three-sentence word problem.

Table 10.5: Example of descriptive statistics for a numeracy item

Item 7								
item:7 (N000674)								
Cases for this item 6415 Item-Rest Cor. 0.49 Item-Total Cor. 0.57								
Item Threshold(s): 0.39 Weighted MNSQ 0.93								
Item Delta(s): 0.39								
Label	Score	Count	% of tot	Pt Bis	t	sig	PV1Avg:1	PV1 SD:1
1	0	792	12.35	-0.09	-7.27	0.000	-0.061	1.122
2	1	3157	49.21	0.49	45.58	0.000	0.952	1.025
3	0	682	10.63	-0.07	-5.76	0.000	-0.037	0.904
4	0	1026	15.99	-0.26	-21.96	0.000	-0.413	0.933
5	0	303	4.72	-0.17	-13.74	0.000	-0.570	0.927
6	0	455	7.09	-0.24	-20.14	0.000	-0.634	0.964

Note. N (missing) = 133.

This table provides information that allows for a classical test theory (CTT) approach to assessing item difficulties, including frequency and percentages of each response category, item-rest, item-total and point-biserial correlations. Other information provided in the table allows for an evaluation of item difficulties based on item IRT, including item threshold and delta, and mean PV for children in each response category.

6415 children have responded to this item across all countries participating in IELS. The international delta (difficulty) is 0.39. Item fit statistic (weighted MNSQ) is equal to 0.93 which indicates good fit to the model. Item-rest correlation is a correlation between the question score (e.g. 0 or 1 for multiple choice) and the test score after removing the item score from the total test score. For multiple-choice items, this index will be the usual point-biserial index of discrimination. The item-total correlation is a correlation between the item score and the overall assessment score. It is expected that if a participant gets a question correct they should, in general, have higher overall assessment scores than participants who get a question wrong. Item-rest correlation and item-total correlation are 0.49 and 0.57 respectively for this item.

The original response categories 7, 8, 9, 10, 11, and 12 are labelled 1-6 respectively in the first column in this table. The second column shows the score assigned to each response category. The correct response is given a value of 'one' while incorrect scores are given 'zero'.

The third and fourth columns in the table list the number and percentage of children in each category. For this item, 3157 children (49%) gave the correct response.

The point-biserial correlations for each category are presented in column five. This is the correlation between a response category coded as a dummy variable (a score of 1 for children that responded with the current code and a score of 0 for children in other response categories) and test score after removing the item score from the total test score. Correct responses should have positive point-biserials correlations, incorrect responses should have negative point-biserials correlations. In this case all of the incorrect responses have negative point-biserials and the correct response has a positive point-biserial of 0.49.

The two last columns show the average ability of children responding in each category and the associated standard deviation. The average ability is calculated by domain. If an item is functioning well the group of children that gave the correct response should have

a higher mean ability than the groups of children that provided each of the incorrect responses. This is true for this item.

10.12. International item calibration

All scales were calibrated using a ten-dimensional MRCMLM model without using conditioning variables**.** In total, data from 6,921 children was available for the calibration step. An approach to employing this method of analysis is via the estimation of fit statistics, predominantly Mean Square (MNSQ) (Wu, 1997[44]). Weighted MNSQ values approaching one indicate adequate fit, with values above one suggesting an under discriminating item, and values below one suggesting an over discriminating item. These fit statistics are used to ensure that items adequately fit the model and have equal discrimination.

Initially, 1PL models were estimated and preliminary fit statistics, item functioning information and DIF analyses of calibrated scales presented to the Technical Advisory Group for evaluation. After the TAG endorsed either deleting or freeing slope parameters (by country) of a selection of items, calibration models were re-run as 1PL models. As there was no significant misfit of retained items estimation of more complex 2PL models were not required.

10.13. Handling of item-by-country/language/gender interactions (DIF analysis)

IELS used a number of techniques to ensure that the IRT model adequately fit the data. An examination of Differential Item Functioning was undertaken to determine if items were easier or harder for children belonging to different groups (country, gender, and language (Estonian vs Russian)) of similar ability.

The translation of IELS questionnaires from English into two other languages (Estonian and Russian) allowed for the potential threat of cross-language validity. Certain items may be more difficult to translate into a particular language than others, thereby altering the way the item functions across country-by-language groups. To overcome this, a Differential Item Function analysis was undertaken to assess how each item operates (the same or differently) within each country-by-language group.

In order to achieve unbiased proficiency estimates for each domain across these sub-groups, it needed to be confirmed that there was no DIF for any item. Ultimately, the underlying latent trait for each domain should be the same relative difficulty across groups. Should particular item parameters be deemed inappropriate for certain groups (country, gender, or language (Estonian vs Russian)), common parameters will not be used for those cases, instead allowing for item parameters to be freely estimated. The estimation of unique item parameters for specific items and specific groups results in the reduction of measurement error.

There are multiple ways to assess whether there is DIF occurring for an item. Firstly, there is the magnitude of DIF. This magnitude is expressed in the response model as an interaction between the item and the grouping variable of interest. Generally, an estimate of interaction that is between -0.2 and 0.2 is non-concerning. A test of statistical significance is also important. However, with the large sample sizes in IELS, nearly all estimates that have a magnitude of DIF that warrant further investigation are statistically significant. Another factor that helps with the judgement about how to treat items that exhibit DIF is whether that DIF is confined to specific groupings or appears in many groupings. Another aid to interpretation and judgement of DIF is the homoscedasticity of bivariate DIF plots. Given that DIF is acceptable to a small degree in international

studies, the plotting of bivariate DIF estimates can be revealing. Consider the following two bivariate plots (see Figure 10.2), both of which indicate the presence of DIF. The plot in Figure 10.1 is somewhat concerning, whereas the heteroscedasticity of the plot in Figure 11.2 is clear evidence of extreme DIF.

Figure 10.2: Bivariate plot of item difficulty by gender for numeracy items

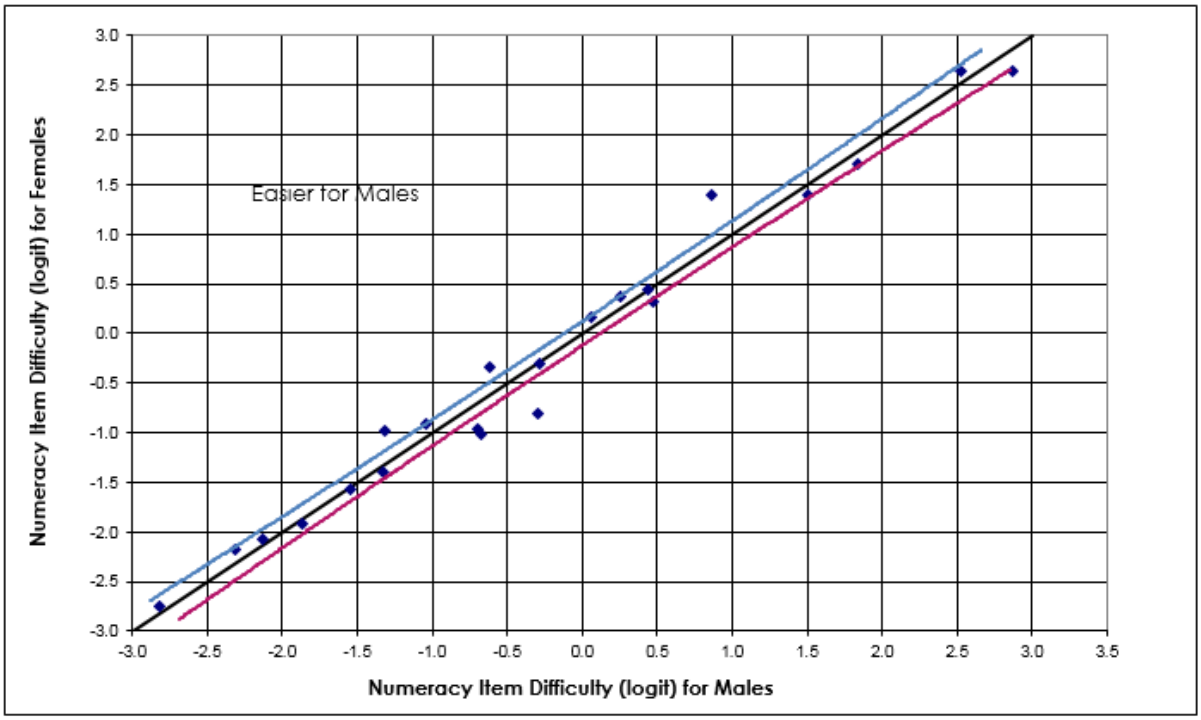
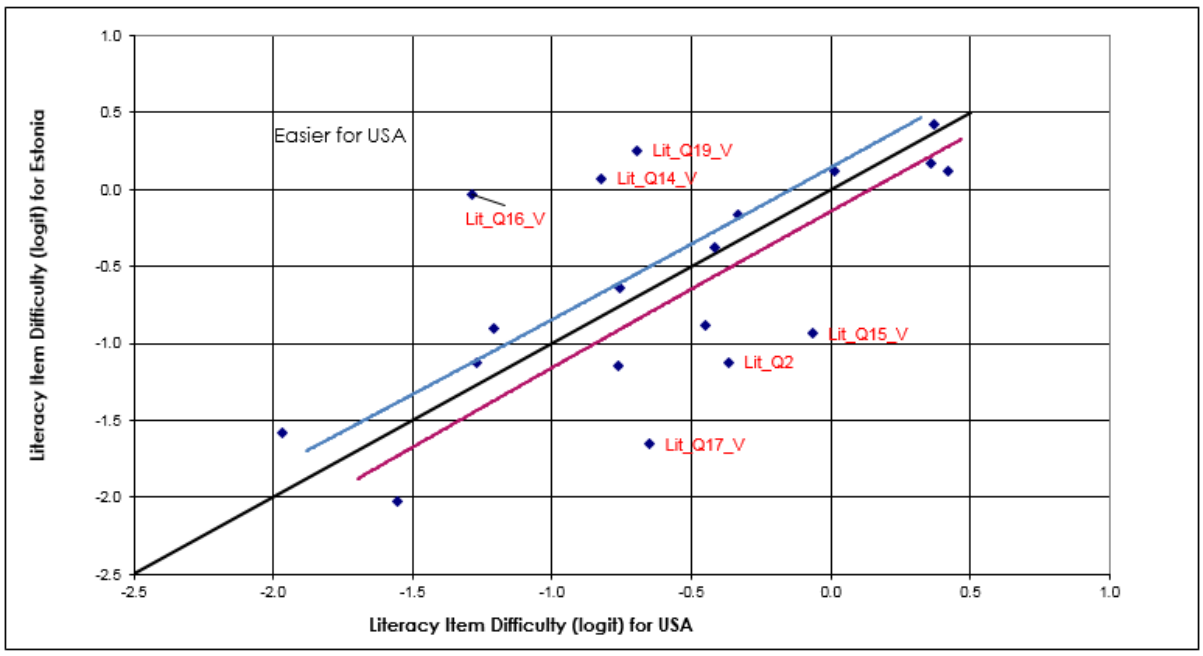
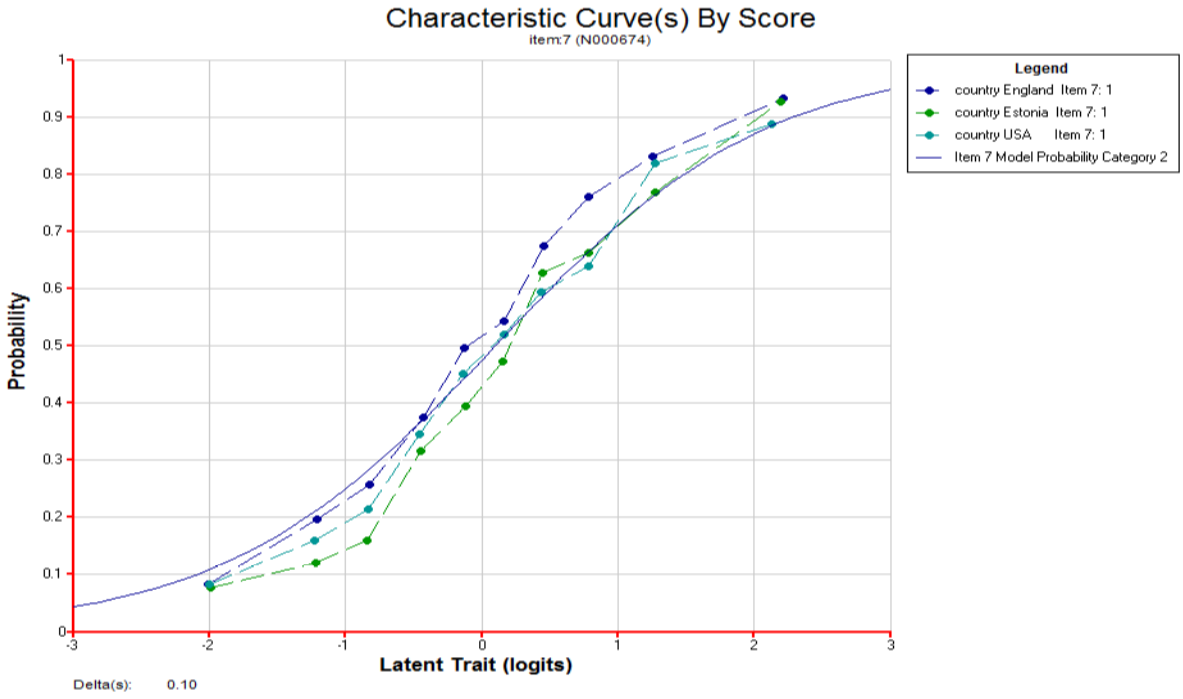


Figure 10.3: Bivariate plot of item difficulty by country for literacy items



Estimated Item Characteristic Curves (ICC) for each item and country (and by language) were viewed to evaluate whether they differed from the expected ICC based on the total sample. If the estimated ICC of a particular group differs from the expected ICC of the total sample then it provides evidence of DIF. An example of a well-fitting item can be seen in Figure 10.3.

Figure 10.4: ICC by country for a numeracy item



It is worth noting the analytical constraints faced by IELS in this first round of data collection.

Firstly, with only three participating countries, DIF by country is more difficult to isolate as each country contributes a third of the variance (i.e. the data are senate weighted for the analysis so each country contributes equally regardless of their population size). DIF is demonstrated when item parameters show deviance from the international average, therefore, small deviances can tend to make all the countries appear to function poorly: as one country may truly deviate from the international mean, so too must the other countries be arranged to compensate (as there are a limited way of arranging 3 countries to show a deviation). That is with only three groups, we often see one group alongside the international mean (computed from two countries data). In a study where 10 or more countries participate, DIF by country can more clearly be identified as an outstanding issue relating to one country when compared to an international mean to which 9 other countries contributed and more closely reflect the international mean.

Similar constraints are present when considering language DIF with only three languages in IELS. Pairwise comparisons of DIF are appropriate in this context but these often show sharp DIF for no obvious reason.

The interaction between country and language in terms of DIF was considered. However, it is impossible to disentangle a country effect from a language effect. This type of interaction (language by country) is not identified as all Estonian and Russian speakers are only situated within Estonia and all English speakers within England and the United States. This leaves five levels of the interaction/combinations (language by country) unidentified. These are English by Estonia, Russian by England, Russian by the United States, Estonian by England, and Estonian by the United States.

A-priori, removal of an item can reduce reliability of the overall scale and can reduce the range of abilities covered (if the item removed was the only one covering a certain level of difficulty). Freeing item parameters for groups can also reduce reliability, although, if the items are measuring different things across groups, the freeing of item parameters may have little effect on reliability.

Depending on the severity of the DIF, items were either retained using international item parameters, retained with item parameters freed across language, or removed completely. A more detailed account of the item treatments are provided in Chapter 11.

10.14. Handling processing speed in the self-regulation developmental area

The cognitive skill of processing speed is a related but distinct skill compared to self-regulation. Children with high processing speed are hypothesised to have higher self-regulation skills and (therefore) complete more items (due to the missing-by-design structure of the assessments). To account for this, a proxy measure of processing speed (average reaction time to non-scored (i.e. non-switching) inhibition items) is included as a regressor in the population model.

10.15. Multi-dimensional scaling of 10 latent dimensions (domains) in IELS and generating plausible values

Each scale was constructed as theorised in the assessment framework. However, analysis was undertaken where there was some evidence of more complex dimensionality. The only case where a theoretically proposed dimension was separated into multiple dimensions as a result of the dimensionality analysis, was empathy (emotion identification and emotion attribution). This is consistent with the underlying theorisation of the dimensions, which identifies skills of recognising and labelling emotions, as well as the skill of feeling (and attributing) an emotional response. Item fit statistics for the 10D model are provided in Annex E. MNSQ values outside the range 0.8-1.2 indicate potentially ill-fitting items.

Once scales were developed, the population model (IRT calibration model and latent regression model combined) was applied to draw five plausible values for each child (across each scale).

Population models were estimated separately for each country-by-language group. Due to the multiple-group scaling approach, item parameter files for each country-by-language were created and used separately during the population modelling process, thus accounting for between-group interactions.

Applying population modelling independently for each country-by-language group ensures that the latent regression model uses data specific to that group (i.e. assessment and contextual information specific to those groups). This one-to-one relationship between the IRT and latent regression models reduced bias. However, the majority of

country-by-language specific item parameters were common across groups, ensuring that cross-country comparison of plausible values can occur.

In order to generate plausible values for the ten domains of interest, multi-dimensional scaling of ten latent dimensions (domains) was conducted.

Statistical reporting for each country-by-language group, for each domain, is done using the five plausible values assigned to each child. Although the five different sets of plausible values are equally appropriate to estimate population-level statistics, multiple imputations improve the estimation of the population variance and therefore yield (under the assumption of an appropriately specific population model) more accurate population estimates including conditional means (e.g. mean difference between boys and girls) or other user-specified secondary analysis (Von Davier, 2009[45]) (Wu, 2005[28])).

The item calibration step (described in more detail in the above section titled 'The IRT models for scaling') provided estimates for item parameters based on children's responses to assessment items. The scales that resulted from the calibration process and subsequent population modelling allowed for comparisons to be made between country-by-language groups.

10.15.1. Constructing conditioning variables

The IELS conditioning variables are prepared using procedures based on those used in PISA. In the population model there was an attempt to include all of the sampling and contextual information into the model. Contextual variables included information captured from parents' responses to the parent questionnaire and educators' responses to the staff questionnaire. Parents answered questions relating to their child's gender, date of birth, family background (i.e. home composition, country of birth, language spoken, income, employment), level of current capacity in literacy and numeracy, ICT use, rate of developmental progression (i.e. social and emotional skills, motor skills, self-regulation, language skills, numeracy skills), prior difficulties (i.e. low birth weight or premature birth, hearing, vision, mobility, learning), behaviour (i.e. prosocial, disruptive), activities outside the home (i.e. formal and informal ECEC experience), activities inside the home, and access to books and ICT resources. Educators answered questions relating to their own gender, age, level of education and qualification, experience, and about each child's parental involvement, services or support received, level of current capacity (i.e. literacy, numeracy), rate of developmental progression (i.e. social and emotional skills, motor skills, self-regulation, language skills, numeracy skills) and behaviour (i.e. prosocial, disruptive). More details are provided in Chapter 4.

All available contextual information was used either as direct or indirect regressors in the conditional model. The preparation of the variables for the conditioning proceeded as follows.

A combination of variables relating to gender, age, school adjusted measures of the ten IELS domains, parent-ratings of their children's behaviour and trust, processing speed, and stratification and were included as direct regressors. More specifically, PVs relating to parents responses produced from an 8 dimensional model (three parent dimensions (social-emotional), three educator dimensions (social-emotional) and two student dimensions (empathy)) were included in the conditioning model. This allows secondary analysts to include parents' responses relating to social-emotional items as predictors in regression models (i.e. one of the ten outcome domains regressed on parents' responses to social-emotional items), without introducing bias.

All other categorical variables were dummy coded (Cohen and Cohen, 1983[46]). These dummy variables and all numeric variables (e.g. processing speed) were analysed in a principal component analysis. Principal component analyses were conducted for each country separately, to ensure that unique relationships between contextual factors for each country was accounted for. Once 90% of the variance was accounted for the principal component extraction process was stopped to avoid over-parameterisation of the model. England, Estonia (Estonian), Estonia (Russian) and United States included 167, 157, 157 and 174 factors extracted from PCAs as indirect regressors, respectively.

10.16. Establishing an IELS scale for the purpose of future reporting of trends

As this is the first IELS study, the scale established will be linked to in any future study in order to establish trends. In following the tradition of the other OECD studies, the outcome domains were scaled and transformed such that they have an international mean 500 and a standard deviation 100.

10.17. Transforming

A scale for each of the ten outcome domains established in IELS underwent a transformation where the sets of five PVs from the final combined model were stacked together to form an international database. This resulted in a full set of transformed PVs, using senate weights, with weighted means of 500 and weighted standard deviations of 100. Also, negative values for transformed PVs were replaced by the lowest positive value.

To establish scales transformation equations, overall mean and standard deviation were calculated for each of ten domains based on five plausible values and using equally weighted countries' data. These were used to standardise each domain plausible values sets to have an overall weighted mean of 500 and weighted standard deviation of 100.

For each domain, each set of plausible values in logits was transformed to an IELS scale score as follows:

$$Scaled\ Score\ PV_{d,i} = \frac{Logit\ PV_{d,i} - Mean\ PV_d}{Std\ Dev\ PV_d} \times 100 + 500$$

Where subscript $d = 1, \ldots, 10$ is a dimension and $i = 1, \ldots, 5$ is a plausible value

The transformations required for each of the domains are as follows:

Literacy scaled score = ((L - 0.396) / 0.81) x 100 + 500

Numeracy scaled score = ((L - 0.22) / 1.18) x 100 + 500

Inhibition scaled score = ((L + 0.117) / 0.49) x 100 + 500

Mental flexibility scaled score = ((L - 0.416) / 0.49) x 100 + 500

Working memory scaled score = ((L + 0.615) / 1.17) x 100 + 500

Emotion identification scaled score = ((L – 1.119) / 0.61) x 100 + 500

Emotion attribution scaled score = ((L - 0.008) / 0.4) x 100 + 500

Prosocial behaviour scaled score = ((L – 1.825) / 1.78) x 100 + 500

Disruptive behaviour scaled score = ((L – 2.118) / 1.72) x 100 + 500

Trust scaled score = ((L – 0.491) / 0.97) x 100 + 500

Where *L* is the logit scale outcome.

One of the benefits of using plausible values it allows us to account for uncertainty (measurement error) which is reflected, for example, in the calculation of standard errors of the mean for each country. In order to account for both sampling error and measurement error the following formulae are used in the calculation of standard error of a statistic (e.g. a mean). The following approach to calculating standard errors using PVs is consistent with the approach outlined by Little and Rubin (1987[47]).

For *m* PVs (*m*=5 in IELS) on a domain we calculate the mean $\hat{\theta}_d$ and the variance $Var(\hat{\theta}_d)$, where *d* {1, 2, …, *m*}

The point estimate of the mean $\overline{\theta}$ is the mean of the *m* values of $\hat{\theta}_d$:

(10.24)

$$\overline{\theta} = \frac{1}{m}\sum_{d=1}^{m}\hat{\theta}_d$$

The variance of $\overline{\theta}$, is given by the total variance:

$$Var(\overline{\theta}) = Var\_w(\overline{\theta}) + (1 + \tfrac{1}{m})Var\_b(\overline{\theta}) \quad (10.25)$$

where:

(10.26)

$$Var\_w(\overline{\theta}) = \frac{1}{m}\sum_{d=1}^{m}Var(\hat{\theta}_d)$$

and:

(10.27)

$$Var\_b(\overline{\theta}) = \frac{1}{m-1}\sum_{d=1}^{m}(\hat{\theta}_d - \overline{\theta})^2$$

The standard error of the mean $\overline{\theta}$, is given by:

(10.28)

$$Standard\ error\ of\ the\ mean\ (\overline{\theta}) = \frac{\sqrt{Var(\overline{\theta})}}{\sqrt{n}}$$

where *n* is the number of children.

Worked examples that show the usefulness of using PVs and calculating standard errors of the mean in order to account for measurement error can be seen in von Davier, Gonzalez and Mislevy (2009[45]).

Notes

[8] Note that for inhibition, working memory and mental flexibility, the design of the assessment did not allow children who were not understanding the task at a particular level to continue to more complex levels. Thus, completion rates for these domains are lower than for other domains.

Chapter 11. Scaling outcomes

11.1. Targeting by dimension

The IELS sample data used in the calibration scaling step was examined for targeting of the assessment for each domain. Item plots are used (Figure 11.1 to Figure 11.10) as a way of evaluating the targeting of each assessment and consist of two panels. The first panel shows the distribution (Rasch scaled) of children's skills or behaviour within the domain, with those children who are placed at the top end of the figure having higher levels of skills or behaviour compared with those on the lower end. The last panel shows the distribution of item difficulty parameter estimates. Presenting them both in this way (on the same scale) shows the match between item difficulty and children's levels of skills or behaviour. It must be noted that where a child is placed on the scale at the same position as an item, they have a 50% probability of responding correctly to that item. A test is well targeted if the average of item parameter estimates is about the same as the average of the children's level of proficiency and the item parameter estimates are evenly spread across the skills distribution.

Figure 11.1: Item plot for literacy

Figure 11.2: Item plot for numeracy

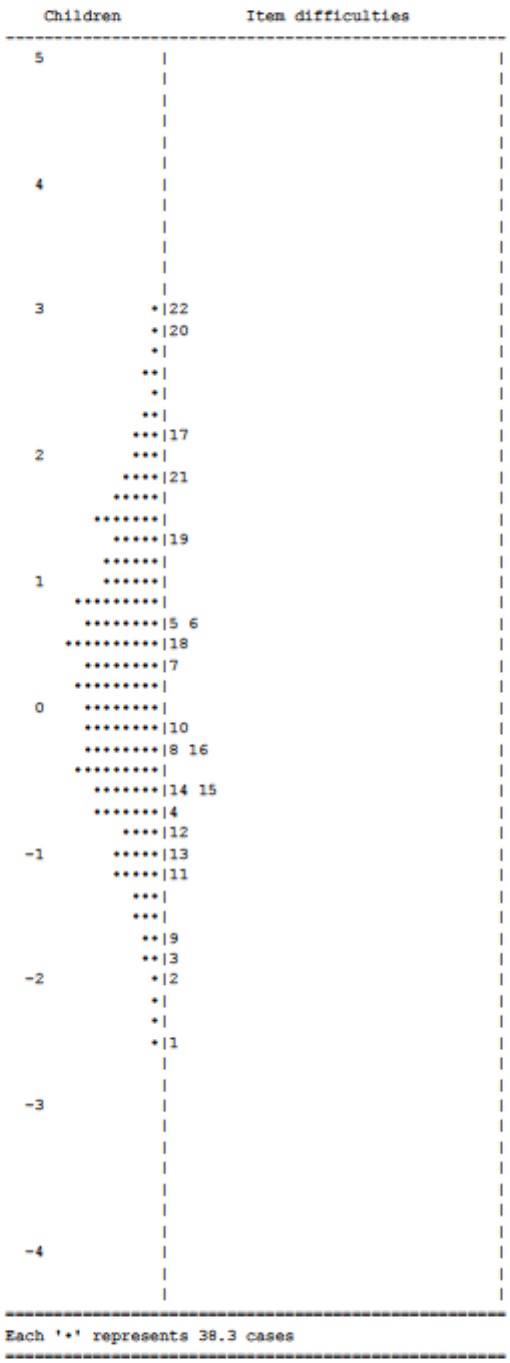


Figure 11.3: Item plot for inhibition

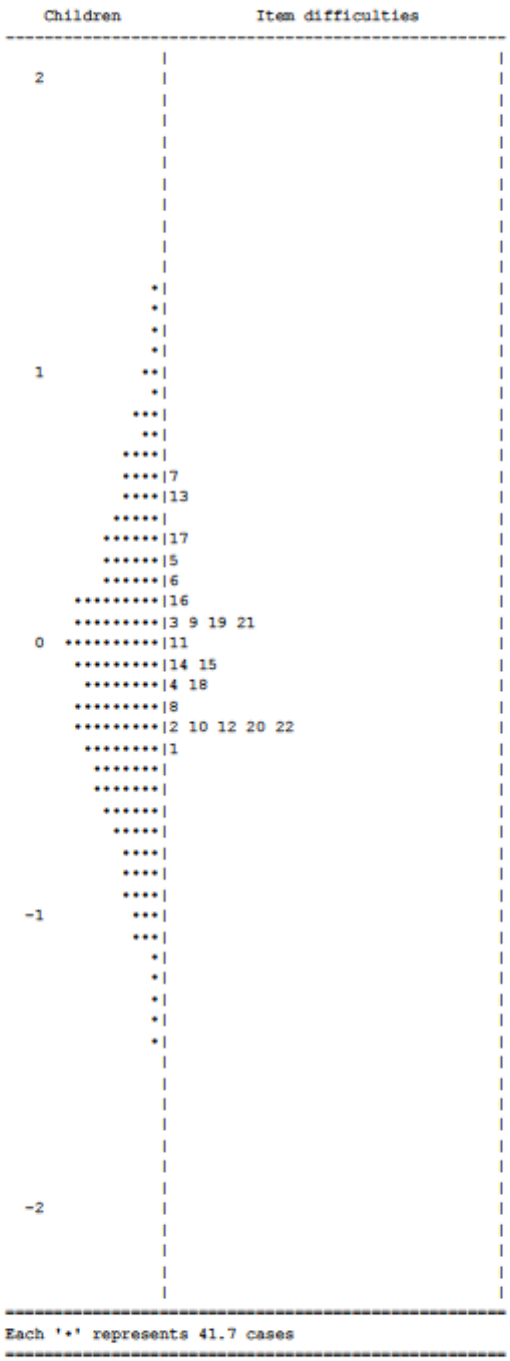


Figure 11.4: Item plot for mental flexibility

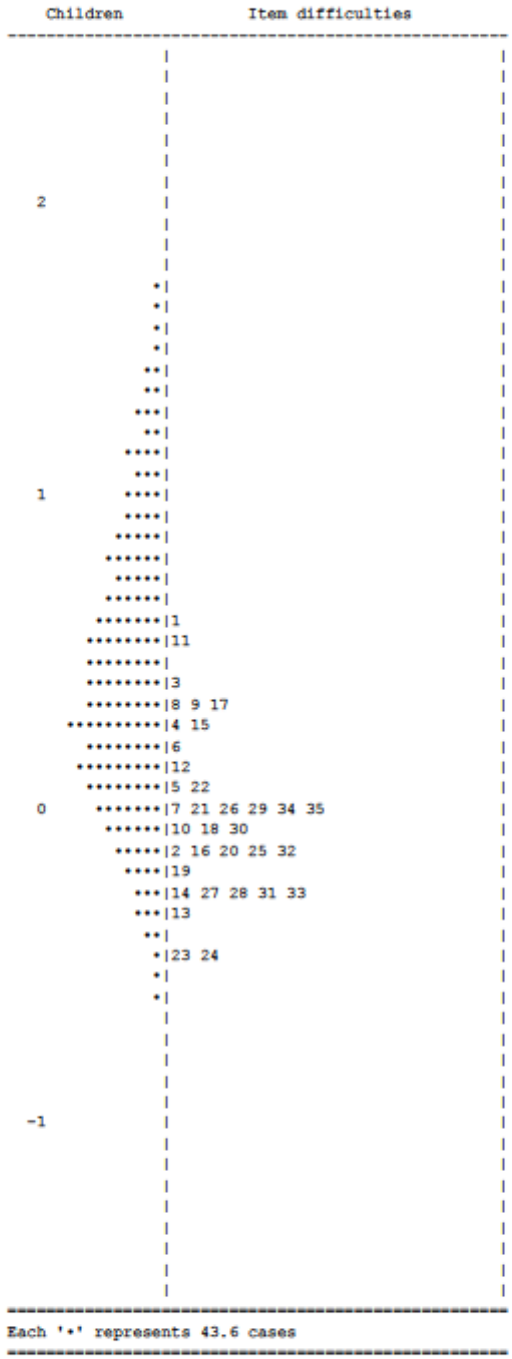


Figure 11.5: Item plot for working memory

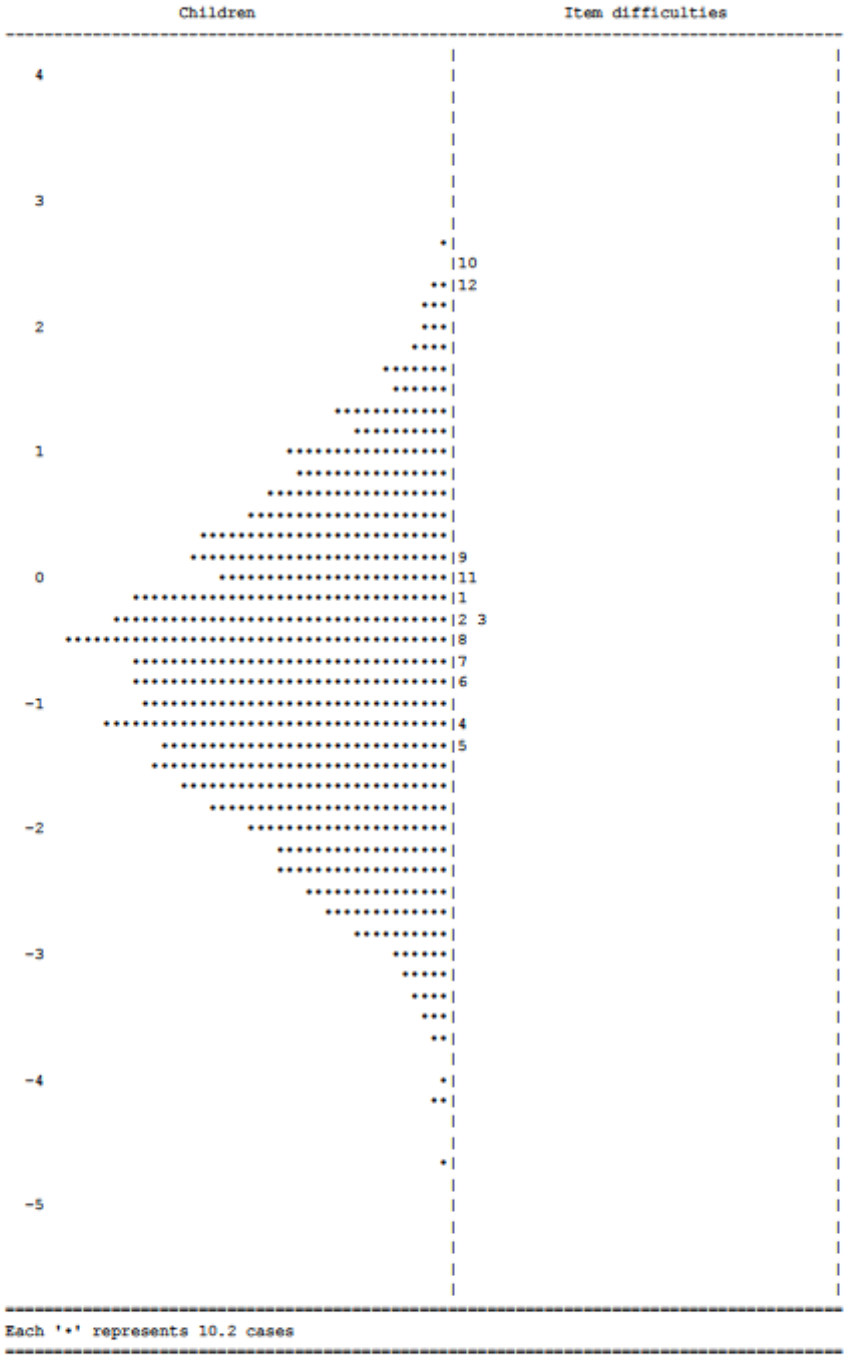


Figure 11.6: Item plot for emotion identification

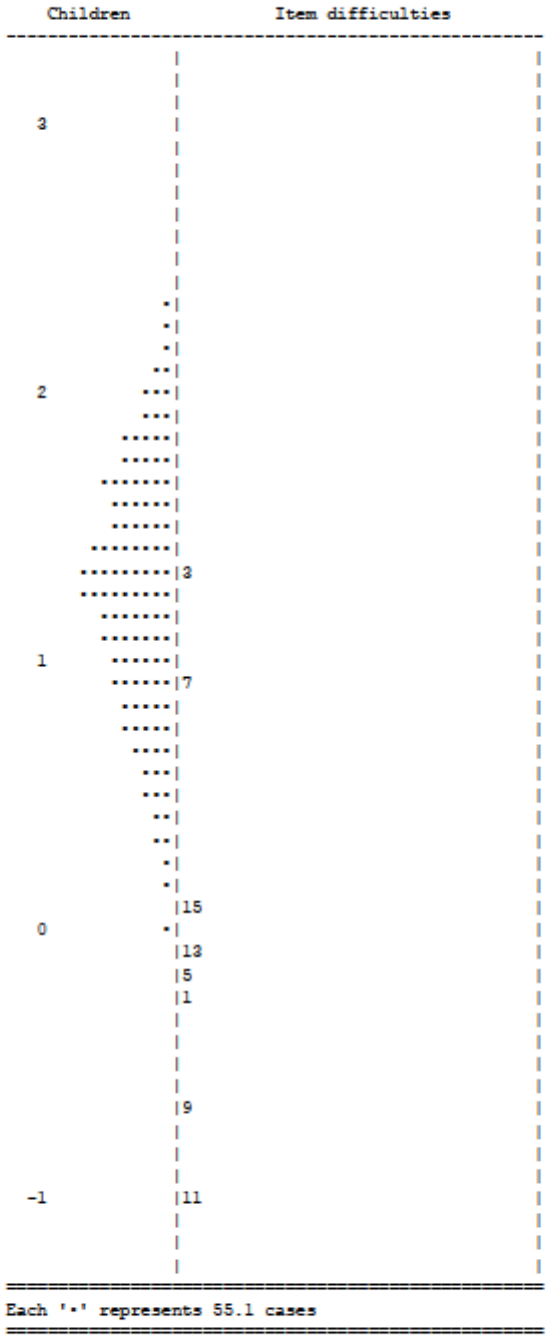


Figure 11.7: Item plot for emotion attribution

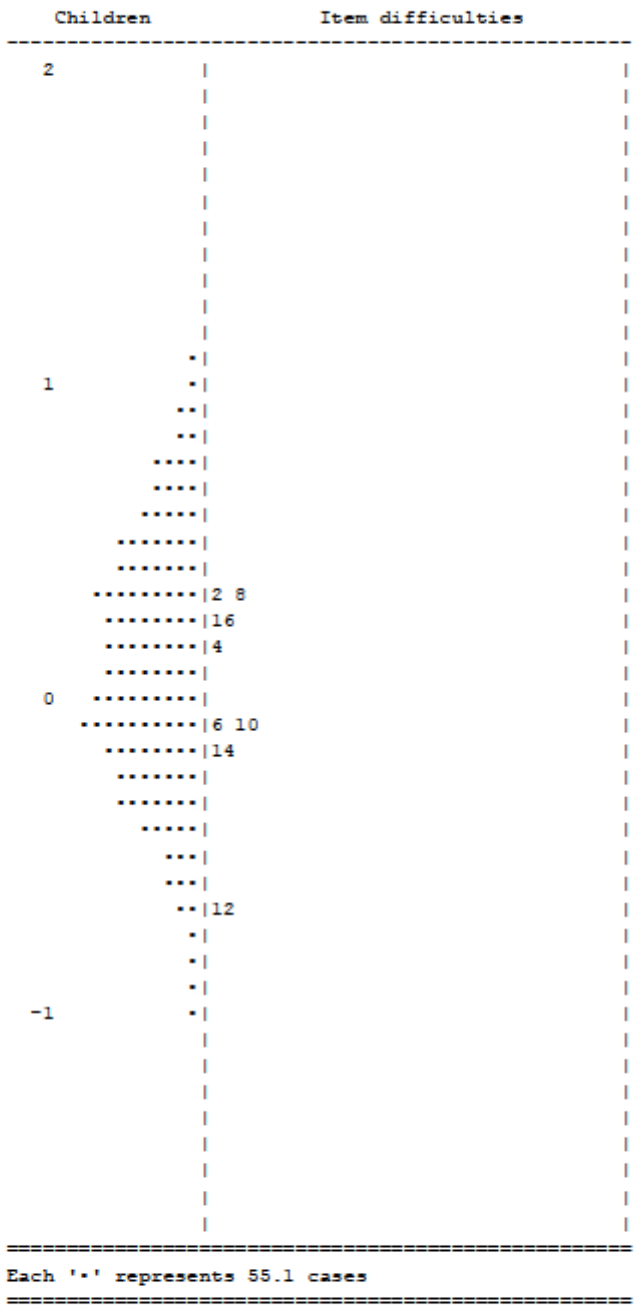


Figure 11.8: Item plot for prosocial behaviour

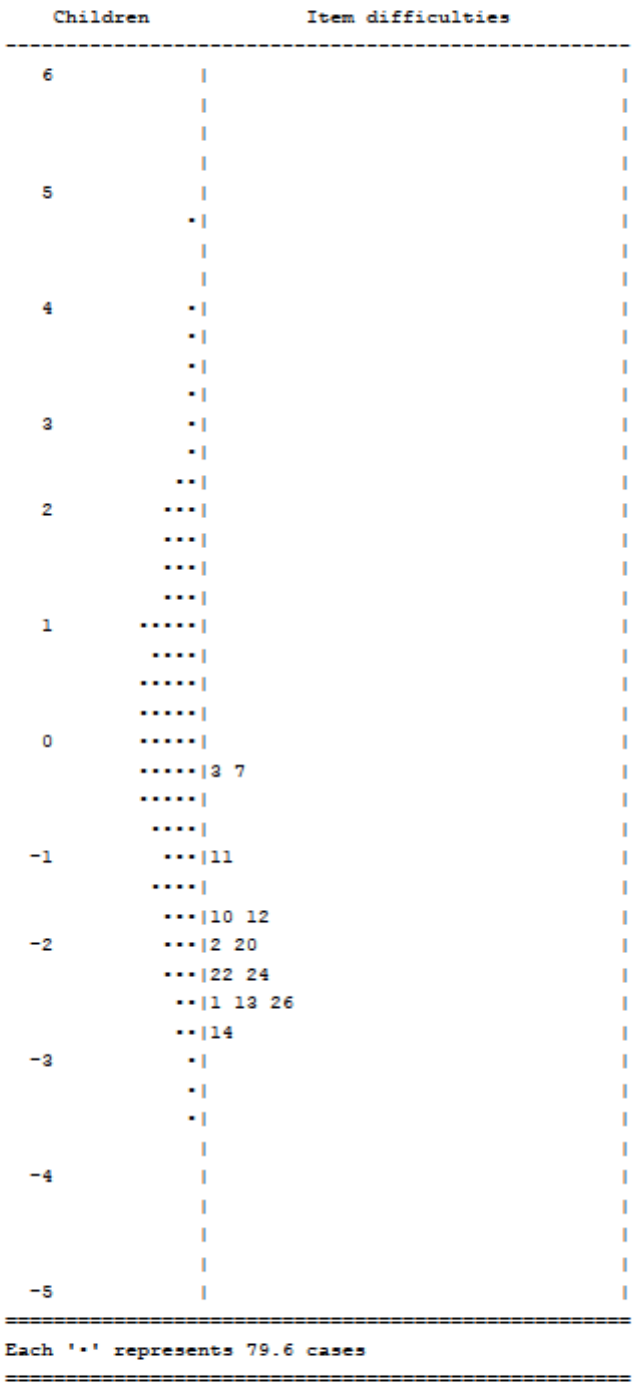


Figure 11.9: Item plot for disruptive behaviour

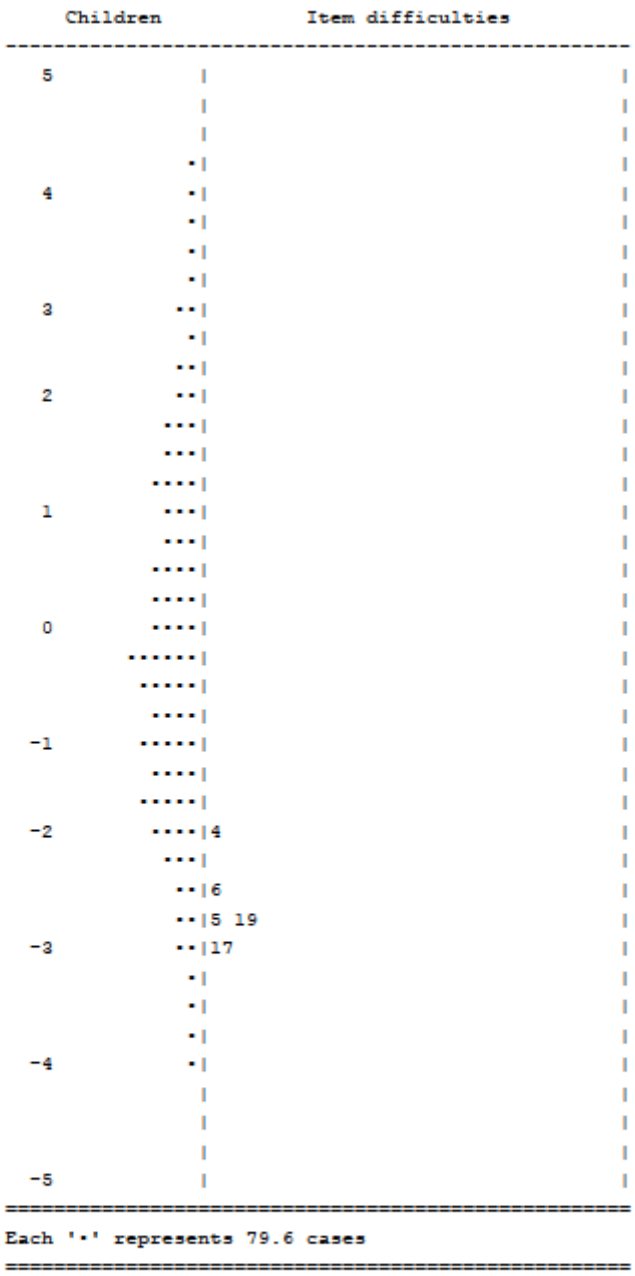
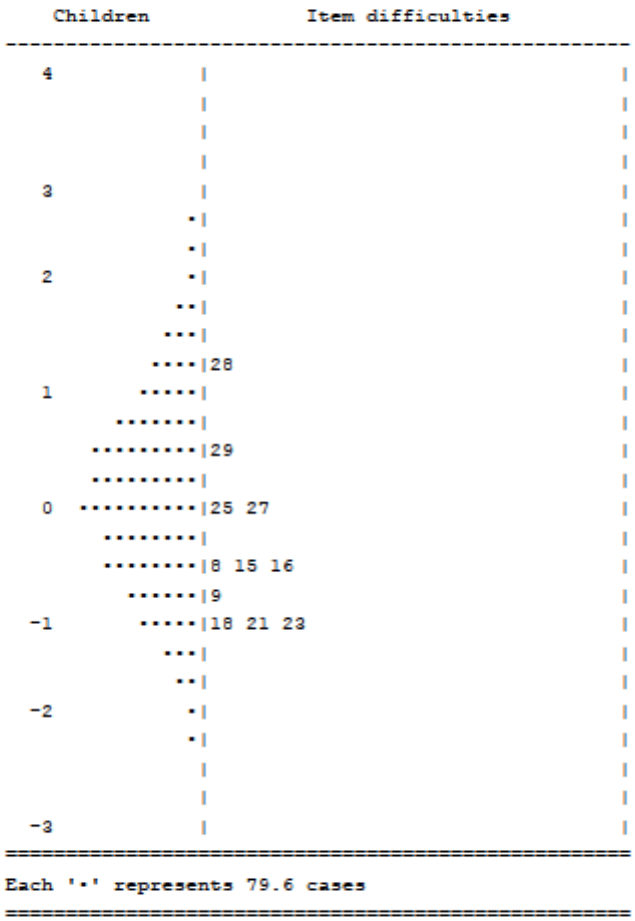


Figure 11.10: Item plot for trust



11.1.1. Inter-correlations between domains

Correlations between domains in IELS were estimated using the 'mcr_SE_cor_2PV.sps' macro in SPSS consistent with the method outlined in the PISA Data Analysis Manual: SPSS® Second Edition (OECD, 2009[48]). This method calculates correlation coefficients between domains using the five PVs for each domain and takes into account sampling weights as well as the complex sample design. The correlations between domains are quite varied, ranging from 0.07-0.80. There is solid evidence of convergent validity (e.g. correlation within the pre-academic and cognitive domains) and divergent validity (lower correlations between the socio-emotional and cognitive domains). The standard error of the correlations (not shown) are produced using the method outlined in the 'Transforming' section of Chapter 10. .

11.2. Reliability by dimension

Reliabilities in the context of large scale assessments related to the degree to which the combined, marginal IRT model reduces uncertainty in the estimation of abilities.

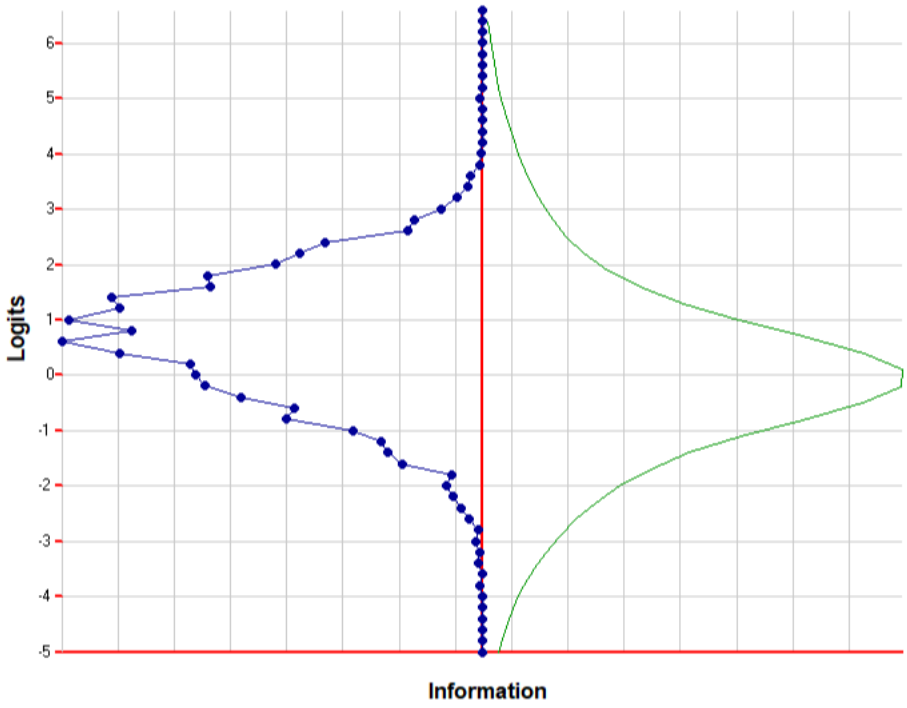
In order to assess the reliability of each domain, expected a-posteriori/plausible value (EAP/PV) (Adams, 2005[27]) reliability estimates are produced for each domain and for each country-by-language group (see Table 11.1). These estimates are produced from the combined model (item response and latent regression (population) model), using all sampled students, while computing plausible values. In this way, EAP/PV reliability estimates are a function, not just of items, but also of the population that responds to them. This process allows for all information in the model, including correlations between latent variables (dimensions) and regressors, to be utilised in the calculations. This results in each child having their own reliability estimate meaning that the EAP/PV measures the degree to which we have reduced the uncertainty of the estimation of proficiencies (Adams, 2005[27]). The EAP/PV is a more robust measure of reliability in this context when compared to classical test theory reliability estimates (i.e. Cronbach's alpha), when we are concerned with issues including the targeting of item difficulty to children's ability and the effect of missing data.

Table 11.1: EAP/PV reliability estimates for each domain and for each country-by-language group

Domain	Estonia		England	United States
	Estonian	Russian		
Prosocial behaviour	0.95	0.95	0.92	0.93
Disruptive behaviour	0.87	0.87	0.84	0.87
Trust	0.90	0.97	0.88	0.90
Literacy	0.85	0.93	0.86	0.83
Numeracy	0.88	0.96	0.89	0.90
Emotion identification	0.78	0.96	0.79	0.75
Emotion attribution	0.71	0.89	0.67	0.70
Mental flexibility	0.86	0.86	0.84	0.78
Inhibition	0.86	0.93	0.91	0.80
Working memory	0.83	0.95	0.79	0.79

In addition to this, test information function plots were used to assess both targeting (also see 'Targeting by dimension' section above) and the coverage of the population proficiencies. An example of a test information function plot for the numeracy domain using England data is in Figure 11.11. The left side panel describes the distribution of test information, while the right side panel is the latent distribution, both placed on a logit scale.

Figure 11.11: Test information function plot for the numeracy domain using England data



The information provided in Figure 11.11 is an example of a latent distribution for one of the latent constructs measured in IELS that provides good coverage of the latent proficiencies of children.

The standard error of the test information is the inverse of the information function. Therefore, at the maximum point of the function the standard error is at its lowest and at the minimum point of the function the standard error is at its lowest.

11.3. Item treatments

In order to ensure that the item response models fit the data, an examination of DIF was undertaken. A detailed description of the process that was used in IELS was provided in the *Handling of item-by-country/language interactions (DIF analysis)* section. The examination consisted of an assessment of estimated ICCs for each item and country-by-language group and the fit of the data using MNSQ values. If DIF was present, one of the following treatments were applied:

0. Note DIF and undertake no further treatment.
1. Remove the item from the assessment pool.
2. Keep the item in the assessment pool but free the parameters around the grouping variable indicating DIF. This essentially removes the item from the international pool and replicates the item over groups of interest.

At the completion of the conditioning step, one item (from the prosocial behaviour domain) was excluded from the study and nine items (eight from the literacy and one from the trust domains) were freed across language due to DIF. No significant DIF was present for the

inhibition, mental flexibility, emotion attribution or emotion identification domains. Although there was some DIF present for numeracy, working memory and disruptive behaviour domains, all items were included in population modelling and estimated using international item parameters. In the literacy domain, the vocabulary items (SL001169, SL001157, SL000963, SL000957, SL000953, SL006352, SL006351 and SL006356) showed substantive DIF by language. The vocabulary item parameters were freed across language and we treated these as separate items for each language (i.e. one English set administered to England and the United States, one non-English set administered in Estonia). This way, the items contribute to estimates of literacy within language, but not across languages. In the trust domain, model specification was checked for several items, with one item (IBRG0616) being freed by language. In the prosocial behaviour domain, model specification was checked for several items, with one item (IBRG0614) being removed from the final conditioning model. A summary of the items included in IELS, including the total number, percentage invariant (across countries) and % dropped (following an examination of item fit and DIF analyses) are provided in Table 11.2. Details of the common and unique item parameters that were used in the final conditioning models across all country-by-language groups can be seen in Annex F.

Table 11.2: Summary of items used in IELS and item treatments

Domains	No. of items	% international items	% dropped
Literacy	22	64	0
Numeracy	22	100	0
Inhibition	22	100	0
Mental flexibility	35	100	0
Working memory	12	100	0
Emotion identification	8	100	0
Emotion attribution	8	100	0
Prosocial behaviour	14	100	7
Disruptive behaviour	5	100	0
Trust	9	89	0

In Table 11.2, it can be seen that 64% of literacy items were categorised as international, meaning 36% were considered national items, predominantly due to the vocabulary items within the domain functioning differently across languages. Although these items are not invariant they are useful. For example, the Estonian language vocabulary items were used in the estimation of proficiencies for children who speak Estonian.

Chapter 12. Data management processes

This chapter describes the procedures followed during the verification and management of the IELS main study database, implemented by the IEA, in collaboration with the ISC at ACER and the National Study Centres of the participating countries.

Managing and processing the IELS national databases and ensuring their integrity, in order to create the international databases, was a complex enterprise that required close collaboration between the IEA, the ISC, and the National Study Centres of participating countries. Once the countries' data files were created and submitted to the IEA, an exhaustive process of verification and editing known as 'data cleaning' began.

Data cleaning is the process of checking data for inconsistencies and formatting the data accordingly, in order to create a Standardised output. The overriding goals of data cleaning processes in the context of IELS were to ensure the following;

- that all information in the databases conformed to the internationally defined data structure
- that codebooks and national adaptation documentation appropriately reflected national adaptations to questionnaires
- that all variables used for international comparisons were comparable across countries.

Control measures were applied by all collaborating partners throughout the data cleaning process in order to assure the quality and accuracy of the IELS data.

12.1. Data sources

As detailed in Chapter 3. , the IELS direct assessment of early learning consisted of a tablet-administered data collection, while the indirect assessment of children's cognitive and socio-emotional skills consisted of both online- and paper-administered questionnaires for staff and parents of participating children. Contextual information was also gathered by the online- and paper-administered questionnaires.

Another goal of data cleaning was to limit potential sources of error originating from the use of two parallel questionnaire modes to an absolute minimum to ensure uniform and comparable conditions across modes and countries. To these ends, IELS questionnaires in both modes were self-administered, had identical contents (to the greatest extent possible), comparable layout and appearance, and took place over the same period.

As detailed in Chapter 7. , countries used IEA WinW3S to track participation of children, staff and parents. The IEA WinW3S databases were also used during data cleaning as a source of information about participation status, data availability, ID linkages, and other comments made by the National Study Centres relevant to data cleaning.

12.2. Computer-based child assessment

For England and Estonia, the direct assessment data were provided to the IEA by the ISC in the form of .csv files. ACER communicated the expected structure to the United States National Study Centre, who provided their own data files corresponding to the other two countries' format. These included files for each domain, as well as timing data files.

12.3. Online data collection of staff and parent questionnaires

As documented in Chapter 2. of this report, IELS offered online collection of Staff and parent questionnaire data as an international option conducted according to a mixed-mode design. Participating countries adopted the online option as a default data-collection mode for some or all respondents (that is, staff and parents of sampled children). National Study Centres had to ensure and document that individual respondents who refused to participate in the online mode, did not have access to the required infrastructure for online participation, or simply preferred to participate via paper-based assessment were provided with a paper questionnaire, thereby ruling out unit non-response.

National Study Centres provided every respondent with individual login information along with information on how to access the online questionnaire. This login information contained the ID and checksum provided from the IEA WinW3S, meaning that the identity validation step occurring at the National Study Centres was the same as for the paper-based questionnaires.

As respondents completed their online questionnaires, their data were automatically stored in one central international server and, therefore, no manual data entry was needed. Data for each country-language combination were stored in a separate table on the server. The different language versions within countries were then merged (at the IEA) with the data from the paper-based questionnaires and with data collected as part of the within-school sampling process.

As for the direct assessment data, for England and Estonia, the indirect assessment data were provided to the IEA by the ISC in the form of .csv files. ACER communicated the expected structure to the United States National Study Centre, who provided their own data files corresponding to the other two countries' format. These included files for each domain, as well as timing data files. The United States did not provide timing data for the staff questionnaire, as agreed upon with the ISC.

12.4. Data entry and verification of paper questionnaires

Each National Study Centre was responsible for transcribing the information from any paper-based assessment booklets and questionnaires into computer data files using the IEA DME software. The DME is a software system developed by the IEA that facilitates data entry and incorporates validation checks to identify inconsistencies. As a general principle, National Study Centres were instructed to enter data for any booklet or questionnaire that contained at least one valid response, discarding unused or empty instruments.

National Study Centres entered responses from the paper instruments into data files created from an internationally predefined codebook template. The codebook contained information about the names, lengths, labels, valid ranges for continuous measures or counts or valid values for nominal or ordinal questions, and missing codes for each variable.

National Study Centres were responsible for adapting the international codebook templates before entering data, reflecting any approved adaptations made to the national questionnaire versions (e.g. an added national response category).

To ensure consistency across participating countries, the basic rule for data entry in the DME required national staff to enter data "as is" without any interpretation, correction, truncation, imputation or cleaning. Resolution of any inconsistencies remaining after this data entry stage was subsequently undertaken during data cleaning (see the section later in this chapter on "Confirming the integrity of the international and national databases").

The guiding principles for data entry included the following:

- Responses to categorical questions to be generally coded as "1" if the first option was used, "2" if the second option was marked, and so on.
- Responses to "check-all-that-apply" questions to be coded as either "1" (checked) or "9" (not checked/omitted).
- Non-responses, ambiguous responses, responses given outside of the expected format or conflicting responses (e.g. selection of two options in a multiple-choice question) to be coded as "omitted".
- Misprinted questions or items to be entered as "not administered".

Data entered with the DME were automatically validated. As each respondent ID number was entered, it was checked by the DME software for alignment with a five-digit checksum generated by IEA WinW3S. A mistype in either the ID or the checksum resulted in an error message prompting the data entry person to check the entered values. The data-verification module of DME also checked for a range of other issues such as inconsistencies in identification codes and out-of-range or otherwise invalid codes. When such issues were flagged by the software, the individuals entering the data were prompted to resolve or to confirm the inconsistencies before resuming data entry.

12.5. Double data entry

To check data entry reliability in participating countries, National Study Centres were required to enter a random sample of 5% or 100 non-blank questionnaires of each type twice by two different data entry persons. The IEA recommended that countries begin the double data entry process as early as possible during the data capture period in order to identify possible systematic incidental misunderstandings or mishandlings of data entry rules and to initiate appropriate remedial actions (for example, retraining staff). Those entering the data were required to resolve identified discrepancies between the first and second data entries by consulting the original instruments and applying the international rules in a uniform way.

While it was desirable that each and every discrepancy be resolved before submission of the complete dataset, the acceptable level of disagreement between the originally entered and double-entered data was established at one percent or less for questionnaire data. Values above this level required a complete re-entry of data. This restriction guaranteed that the margin of error observed for processed data remained well below the required threshold.

The level of disagreement between the originally entered and double-entered data was evaluated by both National Study Centres and the IEA, and it was found that in general the margin of error observed for processed data was well below the required threshold.

12.6. Confirming the integrity of the international and national databases

12.6.1. Overview

As described above, either the National Study Centres or ACER submitted the data from the online assessment and questionnaires (depending on the country) to the IEA. National IELS paper-based questionnaire data were submitted, via the appropriate data files, to the IEA from the National Study Centres. Staff at the IEA then subjected these data to a comprehensive process of checking and editing. To facilitate the data cleaning process, the IEA asked the National Study Centres to provide them with detailed documentation of their data together with their national data files. The data documentation included exported IEA WinW3S databases, copies of all original survey tracking forms, the national versions of questionnaires, as well as information from the Survey Activities Questionnaire (see details in Chapter 8.). National Study Centres also submitted their final translation, adaptation and verification monitoring workbooks in order to provide and confirm complete documentation on all national adaptations. In addition, National Study Centres were asked to provide documentation on all changes or edits applied to the data prior to submission, as well as verify any anomalies that remained (this could be for example a violation of the rotation rule when the country explained a child was absent on day 1, participated on day 2, but received the assessment from day 1). Some countries also submitted requests for edits to the data to be made during data cleaning (for example, changes to ID linkages).

In order to ensure the integrity of the international database, a uniform data cleaning process was followed, involving regular consultation between the IEA and the National Study Centres. After each country had submitted its data and required documentation, the IEA, in collaboration with the National Study Centres, conducted a four-step cleaning procedure upon the submitted data and documentation:

1. Documentation and structure check
2. Identification variable (ID) and linkage cleaning
3. Background cleaning (resolving inconsistencies in questionnaire data)
4. Valid range checks

The cleaning process was an iterative process. Numerous iterations of the four-step cleaning procedure were completed on each national data set. This repetition ensured that all data were properly cleaned and that any new errors that could have been introduced during the data cleaning were rectified. The cleaning process was repeated as many times as necessary until all data were made consistent and comparable. Any inconsistencies detected during the cleaning process were resolved in collaboration with National Study Centres, and all corrections made during the cleaning process were documented in a series of cleaning reports produced for each country.

The following four cleaning reports were prepared and shared with countries for clarification/verification:

1. Duplicates (not applicable for United States)
2. Data availability vs. participation status discrepancies
3. Day 1 vs. day 2 participation discrepancies
4. Linkage findings (child data without linked parent or staff questionnaire data and vice-versa; staff questionnaire Part A data without staff questionnaire Part B data and vice-versa).

Based on the National Study Centres' response to these cleaning queries, modifications were made to the data.

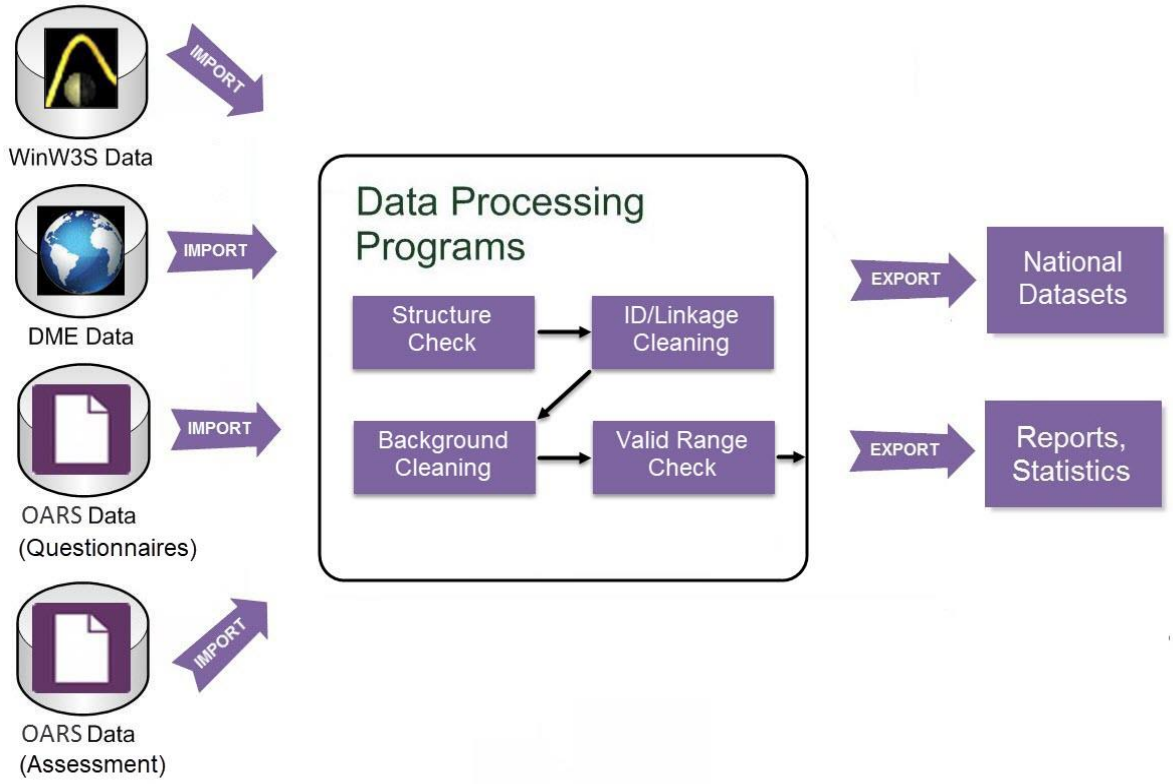
After the final cleaning iteration, the databases with information about child participation and exclusion status were passed to the IEA's sampling unit, which used this information to calculate child participation rates, exclusion rates and child sampling weights (see Chapter 9. for details). Afterwards, the data, including sampling weights, were sent to the ISC so that IRT scaling could be performed to derive estimates of children's outcomes and skills (see Chapter 11. for details). The National Study Centres were provided with interim data products to review at different points in the process.

12.7. Preparing national data files for analysis

The main objective of the data cleaning process was to ensure that the data adhered to international formats, that child, staff and parent information could be linked across different study files, and that the data reflected the information collected within each country in an accurate and consistent manner.

As shown in the graphic below and outlined above, the program-based data cleaning consisted of a set of activities explained in the following subsections. The IEA carried out all of these activities in close communication with the National Study Centres.

Figure 12.1: Data cleaning workflow



12.8. Checking documentation, import and structure of the data

For each country, data cleaning began with an exploratory review of all files in terms of structure and accompanying data documentation, including a review of the TAVMs, tracking forms and the Survey Activities Questionnaires.

The IEA first merged the tracking information and sampling information captured in the IEA WinW3S databases with the child-level databases containing the corresponding direct assessment data. During this step, IEA staff also merged the data from the staff and parent questionnaires for both the online and paper modes of administration. At this stage, data from the different sources were transformed and imported into different SPSS databases so that all necessary information would be available during all subsequent data processing stages.

The first checks identified differences between the international and the national file structure. Some countries made adaptations to their questionnaires (structural adaptations to the direct assessment were not allowed), such as adding national variables or omitting or modifying international variables. The extent and nature of such changes differed across countries. One National Study Centre administered the questionnaires without any modifications (apart from translations and necessary adaptations relating to cultural or language-specific terms), whereas the other two countries inserted or removed response categories within existing international variables or added national variables.

To keep track of national adaptations, National Study Centres were tasked with completing the TAVMs while they were adapting and translating the international version of the survey questionnaires. When necessary, the IEA modified the structure of the national data files to ensure that the resulting data remained comparable across countries.

The IEA then discarded variables created purely for verification purposes during data entry, and made provisions for adding new variables necessary for analysis and reporting, including reporting variables, derived variables, sampling weights and IRT scale scores.

Once each data file matched the international format, as specified in the international codebooks, a series of standard data cleaning rules for further processing of the national data files were applied.

Processing at this stage employed programs developed by the IEA that could identify and correct inconsistencies in the data. All actions taken by the cleaning program or by IEA staff with respect to data inconsistencies were recorded for later review.

The IEA reported problems that could not be solved automatically via the cleaning programs to the responsible NPM, so that National Study Centre staff could check the original data-collection instruments and tracking forms to trace the source of these errors. Wherever possible, the IEA suggested a remedy and asked the National Study Centres to either accept it or propose an alternative. If a National Study Centre could not solve issues through verification of the instruments or forms, the IEA applied a general cleaning rule to the files to rectify this error. When all automatic updates had been applied, SPSS recoding scripts were applied to implement any remaining corrections to the data files.

12.9. Cleaning identification and linkage variables

Each record in a data file needs to have a unique identification number. The existence of records with duplicate ID numbers in a file implies an error. All National Study Centres administered the staff and parent questionnaire on line in addition to the paper mode. There was a possibility that a respondent completed both the paper and the online versions of the

questionnaire. In addition, some National Study Centres provided direct assessment records with duplicate IDs.

If two records in an IELS database shared the same ID number and contained exactly the same data, one of the duplicate records was deleted. If two records with the same ID contained different data and it was not possible to identify which record contained the "true data," National Study Centres were asked to confirm which record should be kept. If no party (IEA, National Study Centre or the ISC) could ascertain which data to keep, records very rarely had to be deleted as there was no way to confirm which records were accurate.

Although the ID cleaning covered all data from all instruments, it focused mainly on the child-level data. In addition to checking the unique child ID, it was crucial to check variables pertaining to child participation, exclusion status and age, as well as dates of testing, in order to calculate the age of the child at the time of testing.

As data on children, staff and parents appeared across different data files, a process of *linkage cleaning* was implemented to ensure that the data files would correctly link together. The linking of the data files followed a hierarchical system of identification codes that included centre, staff and child components. These codes linked the children with their centre, and linked staff with their centre. Parents were identified using the same ID as their children. Using IEA WinW3S, National Study Centres had to indicate linkages between staff and children.

Linkage cleaning consisted of a number of checks to verify that child entries matched between all file types. At this stage, checks were conducted to ensure that staff and child records linked correctly with their corresponding centres. The Child Tracking Forms and staff tracking forms were crucial in resolving any anomalies. The IEA also liaised with NPMs about any problematic cases, and the National Study Centres were provided with cleaning reports listing all linkage inconsistencies identified within the data.

After verifying that there were no missing data or linkage issues that could be rectified within the databases, staff and parent questionnaire data without linked child data were removed from the databases.

In addition to linkage cleaning, tracking information (i.e. the participation status recorded in the tracking forms) was compared against actual data availability. For the questionnaires, the registered mode was also compared against the type of data available. However, in general, National Study Centres did not record the individual mode of questionnaire completion and so this was set to match the type of data available (either online or paper).

Moreover, some National Study Centres sometimes did not complete the tracking forms for the questionnaires. In such cases, the IEA investigated the data availability and set each National Study Centres questionnaire return status accordingly.

For the child-level records, participation between day 1 and day 2 was compared to identify logical inconsistencies (for example, if a child was listed as having participated 1 day, but having left the centre permanently the next). Inconsistencies were shared with National Study Centres for verification and clarification.

In general, if a child had valid data in at least two direct assessment domains, was not marked as having no parental permission, and was within the valid age range, their participation status was set to "participating" (regardless of whether it came from day 1, day 2 or a combination thereof) and their data were included in the final datasets.

12.10. Resolving inconsistencies between tracking information and questionnaire data

Two different sets of IELS data indicated the age of children. The first set was the tracking information provided by the centre co-ordinator throughout the within-school sampling process. The second set comprised the actual responses given by parents in the parent questionnaires. In some cases, data across these two sets did not match and resolution was needed. If discrepancies were found between existing tracking and questionnaire age data for children, the IEA brought this to the attention of the National Study Centres, who investigated and indicated which source of information was correct.

12.11. Handling of missing data

Two types of entries were possible during the IELS data capture: valid data values and missing data values. Missing data can be assigned a value of omitted/invalid or not administered during data entry:

- *Omitted or invalid*: used when the respondent had a chance to answer the question but did not do so, and the corresponding question or item was thus left blank; or for responses that were not interpretable (e.g. when respondents ticked more than one box in a multiple-choice question).
- *Not administered:* this signified that the item or question was not administered to the respondent, which meant that the respondent could not read and answer the question. The not administered missing code was used for those child direct assessment items that were not in the set of assessment blocks administered to a child deliberately (for example, due to stop rules in some of the assessment domains). This missing code was also used for those records that were included in the international database but did not contain a single response to one of the assigned questionnaires. In addition, the not administered code was used for individual questionnaire items that a National Study Centre decided not to include in the country-specific version of the questionnaire. Furthermore, if a particular question or item (or a whole page) was misprinted or for other reasons not available to the respondent, then the corresponding variable(s) was coded as not administered.

12.12. Data cleaning quality control

Because IELS was a complex study with very high standards for data quality, maintaining these standards required an extensive set of interrelated data cleaning procedures. To ensure that all procedures were conducted in the correct sequence, that no special requirements were overlooked and that the cleaning process was implemented independently of the persons in charge, the data quality control included the following steps:

- *Thorough testing of all data cleaning programs*: before applying the programs to real datasets, the IEA applied them to simulation datasets containing all possible problems and inconsistencies.
- *Registering all incoming data and documents in a specific database*: the IEA recorded the date of arrival, as well as specific issues requiring attention.
- *Carrying out data cleaning according to strict rules*: deviations from the cleaning sequence were not possible, and the scope for involuntary changes to the cleaning procedures was minimal.

- Documenting all systematic data recoding that applied to all countries: the IEA recorded these in the data processing programs.
- Logging every "manual" correction to a country's data files in a recoding script: logging these changes allowed IEA staff to undo changes or to redo the whole manual cleaning process at any later stage of the data cleaning process.
- Repeating, on completion of data cleaning for a country, all cleaning steps from the beginning: this step allowed the IEA to detect any problems that might have been inadvertently introduced during the data cleaning process.
- Working closely with National Study Centres at different steps of the cleaning process: the IEA provided National Study Centres with the processed data files and accompanying documentation and statistics so that Centre staff could thoroughly review and correct any identified inconsistencies.

The IEA compared national adaptations recorded in the documentation for the national datasets (i.e. the TAVMs) against the structure of the submitted national data files. Whenever possible, the IEA recoded national deviations to ensure consistency with the international data structure (in consultation with the ISC).

Chapter 13. Sampling outcomes

This chapter focuses on the sampling outcomes of IELS, presenting indicators of the quality of the achieved samples. In the first section, the achieved centre/school sample sizes are presented for each participating country, as well as the numbers of participating children, parents and staff members. Estimates are provided for the total target population size and the exclusion rates in each participating country are given. Weighted and unweighted participation rates have been calculated at centre/school level and for all respondents. Finally, design effects and effective sample sizes are provided for key variables in each participating country.

13.1. Overview of samples

An initial overview of the samples is provided in, Table 13.2 and Table 13.3, which show the intended and achieved sample sizes of centres/schools for each explicit stratum per country including the use of replacement centres/schools.

Table 13.1: School participation in England

Explicit stratum (school type – free school meal (FSM) eligibility)	Total sampled schools	Ineligible schools	Participating sampled schools	Participating first replacements	Participating second replacements	Non-participating schools
Maintained – lowest 20% FSM	26	-	22	2	-	2
Maintained – 2nd lowest 20% FSM	30	-	26	1	1	2
Maintained – middle 20% FSM	32	-	29	2	1	-
Maintained – 2nd highest 20% FSM	32	-	21	4	2	5
Maintained – highest 20% FSM	24	-	20	4	-	-
Academy – lowest 20% FSM	8	-	7	1	-	-
Academy – 2nd lowest 20% FSM	8	-	7	-	-	1
Academy – middle 20% FSM	10	-	7	2	1	-
Academy – 2nd highest 20% FSM	12	-	9	2	-	1
Academy – highest 20% FSM	12	-	12	-	-	-
Independent – unknown FSM	8	-	4	4	-	-
Total	202	-	164	22	5	11

Table 13.2: Centre participation in Estonia

Explicit stratum (language – location)	Total sampled centres	Ineligible centres	Participating sampled centres	Participating first replacements	Participating second replacements	Non-participating centres
Estonian – urban	110	-	90	8	1	11
Estonian – rural	78	1	57	12	2	6
Russian	32	-	26	-	-	6
Total	220	1	173	20	3	23

Table 13.3: School participation in the United States

Explicit stratum (school type – free lunch – region)	Total sampled schools	Ineligible schools	Participating sampled schools	Participating first replacements	Participating second replacements	Non-participating schools
Public – less than 60% free lunch – Northeast	12	-	7	-	-	5
Public – less than 60% free lunch – Midwest	20	-	10	4	3	3
Public – less than 60% free lunch – South	22	-	15	1	1	5
Public – less than 60% free lunch – West	18	-	8	3	2	5
Public – at least 60% free lunch – Northeast	8	-	4	-	1	3
Public – at least 60% free lunch – Midwest	10	-	4	2	1	3
Public – at least 60% free lunch – South	28	2	18	2	3	3
Public – at least 60% free lunch – West	20	-	11	3	2	4
Public – unknown free lunch – Northeast	8	-	3	-	-	5
Public – unknown free lunch – Midwest	8	1	3	2	-	2
Public – unknown free lunch – South	8	-	4	3	-	1
Public – unknown free lunch – West	8	2	2	1	1	2
Private – Northeast	8	-	3	-	-	5
Private – Midwest	8	1	4	1	1	1
Private – South	8	3	2	-	-	3
Private – West	8	1	1	1	1	4
Total	202	10	99	23	16	54

The participation of children within participating centres/schools is summarised in Table 13.4.

Table 13.4: Child participation

Child participation status	England	Estonia	United States
Participated	2 577	2 110	2 234
Participated with no special accommodation	2 575	2 104	2 176
Participated with special accommodation	2	6	58
Absent	52	105	43
No parental permission	89	289	139
Did not participate due to special education needs	32	16	66
Left centre/school permanently	47	17	62
Out of age range	5	147	-
Total sampled children	2 802	2 684	2 544

Parents were asked to complete the parent questionnaire. Parents who completed the parent questionnaire, but without a corresponding participating child, were not included in the analysis. Table 13.5 gives the numbers of sampled and participating children and the numbers of participating parents.

Table 13.5: Parent participation

Country	England	Estonia	United States
Sampled children	2 802	2 684	2 544
Participating children	2 577	2 110	2 234
Participating parents	1 734	1 821	1 597

For each participating child, the staff member who knew the child best was asked to complete a questionnaire about the child. The numbers of completed questionnaires are given in Table 13.6. Note that most staff members were anticipated to complete questionnaires for multiple children.

Table 13.6: Staff participation

Country	England	Estonia	United States
Sampled children	2 802	2 684	2 544
Participating children	2 577	2 110	2 234
Completed staff questionnaires	2 313	2 001	2 134

13.2. Estimated population sizes and exclusion rates

Using sample weights, the size of the whole target population can be estimated based on sample data. From a sampling perspective, the target population of IELS can be divided into four parts:

- children who belonged to centres/schools that could not be sampled because these centres/schools were excluded before centre/school sample selection
- children who were expected to participate if sampled and available, without a special accommodation
- children who were expected to participate if sampled and available, with a special accommodation

- children who were not expected to participate due to their special education needs and who count as exclusions.

The final exclusion rate $Rate_{excl}$ combines exclusions at centre/school level prior to the sample selection and within-sample exclusions of children. It has been calculated as:

$$Rate_{excl} = 1 - \left(\frac{E_1}{TP} \times \frac{W^P}{W^P + W^{E_2}}\right) \quad (13.1)$$

E_1 denotes the number of children in centres/schools excluded prior to the centre/school sample selection, as indicated by the participating country. TP is the expected number of children belonging to the target population prior to the sample selection, including the exclusions. W^P is the sum of sample weights of participating children, and W^{E_2} is the sum of sample weights of excluded children in participating centres/schools.

Table 13.7 presents the estimated population sizes and numbers of excluded children. The values come from different sources: the exclusions before centre/school sample selection originate from statistical data of the participating countries; the counts of participating children, children participating with special accommodation and children not participating due to special education needs are weighted estimates based on the IELS data.

Table 13.7: Estimated population sizes and exclusion rates

Country	Population exclusions before centre/school sample selection	Population represented by participating children, without special accommodation	Population represented by participating children, with special accommodation	Population represented by children who did not participate due to special education needs	Total population	Exclusion rate
England	6 722	603 370	641	7 740	618 473	2.2%
Estonia	55	11 399	29	115	11 598	1.4%
United States	314 863	2 773 863	85 305	85 829	3 259 860	8.6%

13.3. Participation rates

In order to ensure the highest data quality and to guarantee comparability between the participating countries, rigorous sample implementation was required in IELS. To keep non-response bias to a minimum, participating countries were encouraged to ensure the highest possible participation rates. For participation rate requirements, please refer to Chapter 14. .

Participation rates were calculated at centre/school level and for all groups of respondents: children, parents and staff. Both unweighted and weighted participation rates are available. Unweighted participation rates are useful when analysing the success of convincing sampled centres/schools and children to participate in general. The weighted participation rates indicate the proportion of the population represented by the sample, prior to making

non-response adjustments. Hence, they are useful when checking for possible effects of non-response that might impact on the reliability of data. For parents and staff members, two different types of participation rates have been calculated: one based on participating children only, and one based on all sampled children. The response rate referring to participating children gives information about the completeness of the data (i.e. the existence of parent or staff member data for each participating child). On the other hand, the response rate referring to sampled children provides information about the potential for non-response bias (i.e. the level of non-response adding up from two sources: non-participating children and non-participating parents/staff of participating children).

In the following formulae, the notation introduced in Chapter 9. is retained (please refer to Table 9.1).

13.4. Unweighted centre/school response rate before replacement

The unweighted centre/school response rate before replacement gives the proportion of participating originally sampled centres/schools, based on all eligible sampled centres/schools. It is calculated by:

(13.2)

$$\frac{\sum_{i\in Y} 1}{\sum_{i\in(Y\cup R\cup N)} 1}$$

Y denotes the set of participating originally sampled centres/schools, R denotes the set of participating replacement centres/schools and N denotes the set of non-participating eligible originally sampled centres/schools that were not replaced.

Ineligible centres/schools, such as those that were closed or did not have children at the age of 5, are not considered in the calculation of participation rates.

13.5. Weighted centre/school response rate before replacement

The weighted centre/school response rate before replacement is calculated by:

(13.3)

$$\frac{\sum_{i\in Y}\sum_{j} w_{1i}\times w_{2ij}\times f_{2ij}}{\sum_{i\in(Y\cup R)}\sum_{j} w_{1i}\times f_{1i}\times w_{2ij}\times f_{2ij}}$$

In the numerator, only the originally sampled participating centres/schools are considered, multiplying the design weights computed at the centre/school level w_{1i} and at the child level w_{2ij} by the non-response adjustment at child level f_{2ij}. In the denominator, the centre/school non-response adjustment is included as a multiplication factor f_{1i}.

13.6. Unweighted centre/school response rate after replacement

For the unweighted centre/school response rate after replacement, the replacement centres/schools are added into the nominator:

(13.4)

$$\frac{\sum_{i\in(Y\cup R)} 1}{\sum_{i\in(Y\cup R\cup N)} 1}$$

13.7. Weighted centre/school response rate after replacement

The same applies for the weighted centre/school response rate after replacement:

(13.5)

$$\frac{\sum_{i\in(Y\cup R)}\sum_j w_{1i}\times w_{2i}\times f_{2i}}{\sum_{i\in(Y\cup R)}\sum_j w_{1i}\times f_{1i}\times w_{2i}\times f_{2i}}$$

13.8. Unweighted child response rate

The unweighted child response rate is the proportion of children who participate in the study, given that the centre/school participates. For the calculation, the number of participating children P (status "participated" or "participated with special accommodation") in participating centres/schools is divided by the sum of participating children and sampled children A who were not present or declined participation (status "absent") or whose parents did not allow them to participate ("no parental permission"):

(13.6)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon P}1}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)}1}$$

Children who turned out to be out of scope ("out of age range"), who "did not participate due to special education needs" or who had "left centre/school permanently" were not considered in the calculation.

13.9. Weighted child response rate

Similar to the structure of equations (13.3) and (13.5), for the weighted child response rate, the child non-response adjustment f_{2i} is considered in the denominator, but not in the numerator:

(13.7)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)} w_{1i}\times w_{2i}}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)} w_{1i}\times w_{2i}\times f_{2i}}$$

13.10. Unweighted parent/staff response rate, based on participating children

Response rates were also calculated for the populations of parents and staff. The response rate for staff measures for how many children staff questionnaires were completed. Note that it does not measure how many staff members completed questionnaires. This difference is important as the number of questionnaires completed by each staff member varied.

For the unweighted parent/staff response rate based on participating children, the number of completed questionnaires Q was divided by the number of participating children P:

(13.8)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon Q}1}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon P}1}$$

The same formula is used for both parents and staff members.

13.11. Weighted parent/staff response rate, based on participating children

For the weighted parent/staff response rate based on participating children, the weights (without child non-response adjustment) for children with completed parent/staff questionnaires are divided by the same weights, but for all participating children:

(13.9)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon Q} w_{1i}\times f_{1i}\times w_{2i}}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon P} w_{1i}\times f_{1i}\times w_{2i}}$$

13.12. Unweighted parent/staff response rate, based on sampled children

For the unweighted parent/staff response rate based on all sampled children, the number of completed questionnaires Q was divided by the number of eligible children:

(13.10)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon Q} 1}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)} 1}$$

13.13. Weighted parent/staff response rate, based on sampled children

For the weighted parent/staff response rate based on all sampled children, the formula for the weighted response rate based on participating children is combined with the weighted child response rate:

(13.11)

$$\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon Q} w_{1i}\times f_{1i}\times w_{2i}}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon P} w_{1i}\times f_{1i}\times w_{2i}}\times\frac{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)} w_{1i}\times w_{2i}}{\sum_{i\in(Y\cup R)}\sum_{j\epsilon(P\cup A)} w_{1i}\times w_{2i}\times f_{2i}}$$

13.14. Participation rates achieved in IELS

Table 13.8, Table 13.9, Table 13.10 and Table 13.11 show the participation rates that were achieved in IELS.

Table 13.8: Centre/school participation rates

Country	Weighting status	Participation before replacement	Participation after replacement
England	unweighted	81.2%	94.6%
	weighted	81.3%	94.5%
Estonia	unweighted	79.0%	89.5%
	weighted	78.7%	89.6%
United States	unweighted	51.6%	71.9%
	weighted	53.6%	75.2%

Table 13.9: Child participation rates

Country	Weighting status	Participation within participating centres/schools	Combined participation
England	unweighted	94.8%	89.6%
	weighted	94.9%	89.7%
Estonia	unweighted	84.3%	75.4%
	weighted	84.1%	75.4%
United States	unweighted	92.3%	66.3%
	weighted	92.7%	69.8%

Table 13.10: Parent participation rates

Country	Weighting status	Parent participation based on sampled children	Parent participation based on participating children
England	unweighted	63.0%	67.6%
	weighted	64.1%	67.5%
Estonia	unweighted	71.6%	87.2%
	weighted	72.4%	86.0%
United States	unweighted	64.2%	72.5%
	weighted	66.0%	71.2%

Table 13.11: Staff participation rates

Country	Weighting status	Staff participation based on sampled children	Staff participation based on participating children
England	unweighted	84.2%	90.3%
	weighted	85.1%	89.7%
Estonia	unweighted	78.6%	95.6%
	weighted	79.2%	94.1%
United States	unweighted	85.9%	96.9%
	weighted	89.4%	96.4%

13.15. Design effects and effective sample sizes

Due to the complex sampling design used in IELS, which included clustering effects, stratification and unequal weights, the realised samples have been less efficient than simple random samples of the same size would have been. The major driver of this reduced efficiency is the cluster effect that occurs if sampling units are not independent from each other. In IELS, centres/schools have been selected first, and multiple children have been sampled within these centres/schools. Children within a centre/school are more similar to each other as they share the same environment, staff, and tend to come from similar socio-economic backgrounds. Therefore, a random sample of, for example, one child from each of 3 000 centres/schools would much better cover the diversity in the population than a random sample of 15 children from each of 200 centres/schools. This latter cluster sample is much less efficient. Appropriately designed stratification of the sample, however, helps to improve the efficiency of a cluster design.

The design effect $deff$ measures the ratio of the sampling variance $\hat{V}$ of an estimated parameter $\hat{\theta}$ under the given sampling plan, estimated with Balanced Repeated Replication (BRR), to the variance of the same parameter under a simple random sampling (SRS) approach (Kish, 1965[13]). In IELS, the design effect was calculated as:

$$deff(\hat{\theta}, BRR) = \frac{\hat{V}_{BRR}(\hat{\theta})}{\hat{V}_{SRS}(\hat{\theta})} \tag{13.12}$$

The design effect measures how many times larger the sample needs to be for a given complex design when compared to a simple random sample, if the precision of both samples should be equal.

Alternatively, an effective sample size $n_{effective}$ can be calculated as the ratio of the nominal sample size n and the design effect:

$$n_{effective} = \frac{n}{deff} \tag{13.13}$$

The effective sample size estimates the equivalent size of a simple random sample with the same precision. Design effects and effective sample sizes are useful when developing designs and sample requirements for new, similar studies. For example, to determine the required nominal sample size for a study with a similar target population and sampling design, the required effective sample size can be multiplied with the design effect.

Design effects and effective sample sizes are variable-dependent. For this report, they were calculated for key variables available in the datasets; results are shown in Table 13.12. Note that the presented design effects refer exclusively to sampling variance.

Table 13.12: Design effects and effective sample sizes for key variables by participating country

Key variable	England		Estonia		United States	
	Design effect	Effective sample size	Design effect	Effective sample size	Design effect	Effective sample size
Emergent literacy	1.67	1 541	3.30	639	2.62	852
Emergent numeracy	1.81	1 425	2.62	805	3.01	742
Emotion identification	2.18	1 184	2.79	755	1.79	1 251
Emotion attribution	1.94	1 330	2.52	836	1.53	1 458
Inhibition	1.33	1 935	2.03	1 042	1.68	1 330
Working memory	1.55	1 665	2.70	781	3.04	734
Mental flexibility	1.85	1 390	2.08	1 012	2.78	804
Educator – prosocial behaviour	3.13	824	2.67	790	3.26	686
Educator – disruptive behaviour	3.53	729	2.12	995	2.80	797
Educator – trust	3.63	710	3.89	542	2.81	796

When determining the sample sizes for IELS, a design effect of 4 was assumed to achieve an effective sample size of 400. The results show that for these key variables, the design effect is smaller than assumed and the effective sample sizes are actually higher than 400, leading to smaller standard errors and increased precision of estimates.

Chapter 14. Data adjudication

This chapter describes the process used to adjudicate the quality of each participating country's data. Adherence to the IELS Technical Standards was used to adjudicate the quality of the data, particularly those standards around:

- sample exclusions and participation rates
- adaptation, translation and verification of assessment and questionnaire materials
- adherence to specified operational procedures.

The outcomes of each country's adjudication process are presented following the description of the process.

14.1. Data adjudication process

For IELS, a series of quality assurance procedures were designed and implemented. The documentation of the implementation of quality assurance procedures form an initial basis for the adjudication process. Quality assurance in IELS included:

- Technical Standards
- procedures within operations manuals
- adaptation, translation and verification guidelines

The degree to which quality assurance procedures were followed is ascertained through a variety of documentation. These include:

- sampling forms
- reports of centre/school, parent and child participation rates
- non-response bias analysis (where appropriate)
- Translation, Adaptation and Verification and Monitoring workbooks
- National Quality Assurance Monitoring reports
- International Quality Assurance Monitoring reports

The adjudication process involves each institutional member of the Consortium reviewing and reflecting on the quality assurance documentation within each relative area of responsibility. IEA for example was primarily responsible for determining the degree to which sampling standards and prescribed operational procedures had been followed. cApStAn were primarily responsible for determining the degree to which the procedures around adaptation, translation and verification of assessment and contextual questionnaire materials were followed. Both IEA and cApStAn consulted with ACER as the International Study Centre to judge the impact of any deviations from the quality assurance procedures on the overall quality of the resultant data.

The data adjudication process concluded with a series of recommendations from the Consortium to the OECD about the publication of data from each participating country. Recommendations can fall into three general categories:

1. Publication of data is recommended, along with documentation that all standards have been met and there are no concerns about the comparability of the data.
2. Publication of data is recommended, along with documentation that some standards have been not been met or that there are concerns about the comparability of the data.
3. Publication of data is not recommended, along with documentation that some standards have not been met or that there are concerns about the comparability of the data.

From the outset, it should be noted that all recommendations arising from the adjudication process for each of the three countries in IELS were for publication of the data. However, some documentation of quality assurance issues, including non-adherence to Technical Standards are documented as well.

14.2. Data adjudication outcomes

14.2.1. Population exclusion

The IELS Technical Standards specify that no more than 5% of the 5 year-old population should be excluded from the national target population, to form the national study population.

England met the standard with a 2.2% overall exclusion rate.

Estonia met the standard with a 1.4% overall exclusion rate.

The United States did not meet the standard, with 5.9% of schools excluded prior to sampling and an 8.6% overall exclusion rate.

It is noted that the United States excluded all children from educational institutions without kindergartens from the national study population. This category of exclusion alone accounted for 3.8% of the population of 5 year-olds registered in schools. Not meeting this exclusion standard means that there is a risk that the United States national study population is not as representative of its national target population, as other participating countries. This risk was known in advance of the IELS implementation and was judged acceptable provided the risk was publicly documented, as it is here.

14.3. Participation

Sample participation rates are adjudicated on the basis of the IELS Technical Standards. Table 14.1 presents the adjudication ratings of participation scenarios.

Table 14.1: Adjudication ratings of participation scenarios

Centre/school participation in IELS		Child participation after centre/school replacements	Rating
Before replacement	**After replacement**		
At least 75%	At least 75%	At least 75%	Acceptable
		At least 50%, but less than 75%	Intermediate[1]
At least 50%, but less than 75%	At least 75%	At least 75%	Intermediate[2]
		At least 50%, but less than 75%	Limited[3]
Less than 50%	At least 75%	At least 75%	Limited[2]
		At least 50%, but less than 75%	Insufficient
Less than 75%	Less than 75%	At least 75%	Insufficient
		At least 50%, but less than 75%	Insufficient
Regardless of centre/school participation		Less than 50%	Insufficient

Notes: 1 – non-response bias analysis (NRBA) needed; 2 – centres/schools NRBA needed; 3 – children & centres/schools NRBA needed.

The participation outcomes and adjudication ratings of each country are presented in Table 14.2.

Table 14.2: Adjudication ratings of participating countries

	Weighted centre/school participation in IELS		Weighted child participation after centre/school replacements	Rating
	Before replacement	**After replacement**		
England	81.3%	94.5%	94.9%	Acceptable
Estonia	78.7%	89.6%	84.1%	Acceptable
United States	53.6%	75.2%	92.7%	Intermediate

As seen in Table 14.2, both England and Estonia comfortably met the participation standards and were judged as having 'Acceptable' participation rates.

The United States did not meet the threshold of 75% centre/school participation before replacement, but did so after replacement and had a high child participation rate after this. This gave the United States an initial 'Intermediate' participation rating. Because the non-participation issue within the United States was only at the school level, the United States was subsequently required to undertake a non-response bias analysis at the school level, to provide the Consortium with more information on which to judge the likely quality of the achieved sample. On examining the results of the non-response bias analysis, Consortium experts endorsed the rating of an intermediate outcome.

14.4. Linguistic quality assurance

The linguistic quality assurance process is documented in Chapter 6. of this report. Linguistic quality assurance procedures included: translatability assessment, translation and adaptation guidelines, and documentation of the process in the Translation, Adaptation and Verification Monitoring workbooks, verification, and final content checks of materials (audio, computer-based and paper-based text).

The linguistic quality assurance process revealed some problematic issues including:

- agreed adaptations not being implemented by translators / National Study Centre staff
- updates to the source version from field test to main study not being implemented
- minor linguistic defects and register issues in translations
- some quality issues (pacing and rhythm) with the voice-over of the audio components.

None of the issues mentioned were found to be systemic. All of the issues identified along the way were rectified either by cApStAn or the National Study Centre.

There were no recommendations to exclude any data from any of the three participating countries due to linguistic quality concerns.

14.5. Data processing

Data management and the quality assurance processes incorporated in these activities are detailed in Chapter 12. of this report. For data adjudication purposes, only issues that might affect the quality of the reporting were taken into consideration.

There were several issues identified that could potentially impact on the quality of the reporting. Issues included:

- tracking forms not being completed
- duplicate records
- incorrect IDs
- incomplete data entry
- processing of data before submission
- International Standard Classification of Occupations (ISCO) coding reliability.

All deviations were investigated and where possible rectified. After this process was completed, none of the remaining issues were judged to be of serious enough risk to warrant further investigation, nor were they judged to seriously affect the overall quality of the data.

There were no recommendations to exclude any data from any of the three participating countries due to data processing concerns.

14.6. Field operations

The IELS field operations are detailed in Chapter 7. of this report. Quality assurance monitoring at the national and international levels (see Chapter 8.) was the primary quality assurance mechanism for field operations.

As with the other quality assurance processes, the International Quality Assurance Monitoring program determined that there were some issues which could potentially have an impact on the quality of the study. These included some cases of;

- altering of operational manual content
- centre co-ordinators appearing to be untrained or unfamiliar with the operations of the study
- mis-application of language experience codes
- standardised forms not used to gather critical information (alternative forms used)

In the case of quality assurance monitoring, a major goal is to document issues for the future improvement of the project. Two issues that arose across contexts were;

- the importance of high-quality training of centre/school co-ordinators and Study Administrators
- the importance of ensuring that information is collected in a standardised manner.

None of the issues above was judged serious enough to represent a threat to the overall quality of the data, either individually or collectively. After examining the field operations issues during data adjudication, there were no recommendations to exclude any data from any of the three participating countries due to operational concerns.

Chapter 15. Procedures and construct validation of context questionnaire data

There are three types of indices, all derived from the IELS questionnaire data that are discussed in this chapter:

- Simple indices that were constructed through arithmetical transformation or recoding;
- Scale indices that were derived through the scaling of items - this was achieved by using item response modelling with two or more categories; and
- A composite index of socio-economic status based on data from parental education, parental occupation and household income – this was achieved using regression analysis.

The first part of this chapter outlines the procedures used for the scaling of questionnaire data from IELS data. The second part lists the simple indices that were derived from IELS data and describes how they were created. Finally, the third part describes the scaled indices with statistical information on the factor structure of related item sets, scale reliabilities and parameters used for the IRT scaling.

15.1. Computation of simple indices

15.1.1. Parent questionnaire

Early childhood education and care attendance

Question 8 of the parent questionnaire asked parents to indicate whether their child had regularly attended different types of early childhood education and care settings prior to the current setting. Each country was asked to list the types of care and educational settings later classified as *supervision and care*, *ISCED level 01* (early childhood educational development) or *ISCED Level 02* (pre-primary education). In England and the United States the allocation of centres to the latter two categories depended on the child's age at the time of the survey. Responses to this question were used to derive indices of *attendance to a supervision and care setting* (ELPAQ08_SVC) and *attendance at an Early Childhood Education centre* (ELPAQ08_ISCED) prior to the current setting. Early childhood education and care settings were chosen from nationally adapted lists of available programmes prior to their current early childhood education and care arrangement. The nationally adapted lists for each participating country were then coded accordingly into internationally comparable categories.

Early childhood education and care: age of and intensity of attendance

In question 9 of the parent questionnaire, respondents were presented the same early childhood education and care settings that they indicated their child attended in the previous question. For each of these settings, they were asked to indicate at what age the child had attended each of the early childhood and care programmes (before age 1, age 1, age 2, age

3, age 4, and age 5) and for how long (either more than 20 hours or less than 20 hours). These data were used to establish indices of attendance for each of the three types of care settings (Supervision and care, ISCED 01, ISCED 02) for the six different age groups (before Age 1, Age 1, Age 2, Age 3, Age 4, Age 5). This totalled 18 different indices of *early childhood education and care setting attendance and intensity* (ELPAQ09SVC00 to ELPAQ09SVC05, ELPAQ090100 to ELPAQ09010, ELPAQ090200 to ELPAQ090205).

Maternal education

A tertiary education index was created for the mother (EDU_MOT) of the child. If the parent was the mother (as indicated in ELPAQ1301=1), their completed tertiary education based on their responses to their highest level of formal education (ELPAQ2001 and ELPAQ2002) was recoded as follows:

Mother completed tertiary education (ISCED level 6, 7 or 8) (EDU_MOT)

(0) Mother completed education up to ISCED level 5

(1) Mother completed tertiary education (ISCED level 6, 7 or 8)

Household composition

In question 15 of the parent questionnaire, respondents were asked to indicate how many parents, grandparents and others (including aunts, uncles, cousins, friends etc.) usually live at home with them. The items ELPAQ15701 and ELPAQ1502 were recoded into the following categories to form the *dichotomous household composition* index (SING_PARENT):

(0) Two-parent household (a maximum of two mothers or two fathers or one mother and one father)

(1) Single-parent household (a maximum of one mother or one father resided with the child)

Households where no mothers or fathers resided with the child were assigned a missing value for this index.

Immigration background

Immigration background is defined in IELS as having two parents born in a country other than that in which the child participated (or one in the case of a single-parent household). The IELS database contains three country-specific variables relating to the child's country of birth, parent/guardian 1 and parent/guardian 2's country of birth. The items ELPAQ1701, ELPAQ1702 and ELPAQ1703 were recoded into the following categories to form the *dichotomous immigration background – parent(s) born abroad* index (IMMIG):

(0) No

(1) Yes

Language most spoken at home

Question 18 of the IELS parent questionnaire asked which language the child as well as parent/guardian 1 and parent/guardian 2 spoke at home most of the time, with each country providing a list of language options that were most relevant in their national context. This information was used to derive two indices reflecting the use of the assessment language at home.

The item ELPAQ1801 was used to derive the child language (STUD_LANG) index in which responses were grouped into two categories:

(0) The language spoken at home most of the time differed from the language of assessment;

(1)The language spoken at home most of the time was the language of assessment.

The parent language index (PARENT_LANG) was derived based on the definition that at least one parent primarily spoke a language other than the assessment language. Item ELPAQ1802 and ELPAQ1803 were used to derive the parent language (PARENT_LANG) index and the responses were grouped into two categories:

(0) Parent(s) primarily speak assessment language;

(1) Parent(s) primarily speaks other language.

Where there is a missing value for either ELPAQ1802 and ELPAQ1803 items and the non-missing item is a 1 (parent primarily speaks assessment language) and it is a single-parent household (as indicated using the single-parent index), these cases were assigned a 0.

Highest occupational status of parents

Occupational data for both of the child's parents were obtained by asking open-ended questions. The responses were coded to four-digit ISCO codes (ILO, 2007) and then mapped to the international socio-economic index of occupational status (*ISEI*) (Ganzeboom, 2010[49]). Three indices were derived based on this information: *Parent 1's occupational status* (P1SEI); *Parent 2's occupational status* (P2SEI); and the *highest occupational status of parents* (HISEI) which corresponds to the higher ISEI score of either parent or to the only available parent's ISEI score. For all three indices, higher ISEI scores indicate higher levels of occupational status.

Educational level of parents

Parental education is a second family background variable that is often used in the analysis of educational outcomes. The main difficulties with measuring educational background of parents relate to international comparability (education systems differ widely between countries and within countries over time) and, especially with increasing migration, difficulties in the national mapping of parental qualifications attained elsewhere.

Parent-responses regarding parental education were classified using ISCED (UNESCO, 2012[50]). Indices on parental education were constructed by recoding educational qualifications into the following categories: (0) ISCED 1 (primary education) or less, (1) ISCED Level 2 (lower secondary education), (2) ISCED Level 3 (upper secondary education), (3) ISCED Level 4 (non-tertiary post-secondary) or ISCED Level 5 (short-cycle tertiary), (4) ISCED 6 (bachelor's level tertiary) or ISCED 7 (master's level tertiary) or ISCED 8 (doctorate level tertiary). Indices with these categories were provided for the child's two parents (P1SCED and P2SCED). In addition, the index on the highest educational level of parents (HISCED) corresponds to the higher ISCED level of either parent. The index for highest educational level of parents was also recoded into estimated number of years of schooling (PARED).

15.2. Staff questionnaire

15.2.1. Parental involvement in their child's schooling

Question 2 of section B of the staff questionnaire asked teachers of participating children to rate the level of involvement of the child's parents/guardians in activities taking place at their current setting. A dichotomous index was created to capture *teachers' perceptions of parental involvement in their child's* schooling (INVOLVE) which was coded as:

(1) Parents are slightly or not involved (where ELCHQ0201 equals 3 'Slightly involved' and 4 'Not involved at all')

(2) Parents are strongly or moderately involved (where ELCHQ0201 equals 1 'Strongly involved' and 2 'Moderately involved').

15.3. Scaling analysis and methodology

15.3.1. Construct validation

The development of comparable measures of the child's development is a major goal of IELS. Confirming the expected dimensionality of item sets to measure these constructs is of particular importance as measures derived from questionnaires are often used to predict differences in children's development and are, thus, potential sources of policy-relevant information about ways of improving educational systems. There are different methodological approaches for validating questionnaire constructs, each with their advantages and limitations. For IELS, item dimensionality was first confirmed using structural equation modelling (SEM) based on the conceptual underpinning of the questions in the assessment framework and from the identified structure from field test data. Once this was established, the internal consistency of the proposed scales were then established. Finally, IRT scaling methodology (the one-parameter Rasch model (Rasch, 1960[51]) was used to derive estimates of each scale for each child[9]. Each step in this process is described below.

Structural equation modelling (SEM) (Kaplan, 2009[52]) provides a tool for modelling and confirming theoretically expected dimensions measured with sets of children, teacher, or school questionnaire data. At the field trial stage, it can also be used to re-specify originally expected dimensional structures.[10] When using confirmatory factor analysis, researchers acknowledge the need to employ a theoretical model of item dimensionality that can be tested via the collected data. Within the SEM framework, latent variables link to observable variables via measurement equations. An observed variable x is thus modelled as:

$$\chi = \Lambda_x \xi + \delta \qquad (15.1)$$

where Λ_x is a q x k matrix of factor loadings, ξ denotes the latent variable(s), and δ is a q x 1 vector of unique error variables. The expected covariance matrix is fitted according to the theoretical factor structure.

When conducting the confirmatory factor analyses for IELS questionnaire data, selected model-fit indices provided measures of the extent to which a particular model with an assumed a-priori structure had a good fit with regard to the observed data. For the IELS analysis, the assessment of model fit was primarily conducted through reviews of the root-mean square error of approximation (RMSEA), the comparative fit index (CFI), and the non-normed fit index (NNFI), all of which are less affected than other indices by sample size and model complexity (Bollen and Long, 1993[53]).

We interpreted RMSEA values indicating model fit as unacceptable with values over 0.10, as marginally satisfactory with values between 0.08 and 0.10, as satisfactory between 0.05 and 0.08, and as a close fit with values lower than 0.05 (MacCallum, Browne and Sugawara, 1996[54]). As additional fit indices, CFI and NNFI are bound between 0 and 1. Values below 0.90 indicate a non-satisfactory model fit whereas values greater than 0.95 were interpreted as suggesting a close model fit, ((Bentler and Bonett, 1980[55]); (Hu, 1999[56])).

In addition to these fit indices, reviews of standardised factor loadings and the corresponding residual item variances provide further evidence of model fit for questionnaire data. Standardised factor loadings l ' can be interpreted in the same way as Standardised regression coefficients by assuming that indicator variables are regressed on an underlying latent factor. The loadings reflect the extent to which each indicator measures the underlying construct. Squared Standardised factor loadings indicate how much variance in an indicator variable can be explained by the latent factor and are related to the (Standardised) residual variance estimate d' (reflecting the estimated proportion of unexplained variance) as:

$$\delta' = (1 - l'^2) \tag{15.2}$$

The use of multi-dimensional models also allows an assessment of the estimated correlation(s) between latent factors, which provide(s) information on the similarity of the different dimensions measured by related item sets.

Generally, maximum likelihood estimation and covariance matrices are not always appropriate for analyses of (categorical) questionnaire items as they may not provide robust estimates. Therefore, the analyses of IELS relied on robust weighted least squares estimation (WLSMV) ((Flora, 2004[57]); (Muthén and du Toit, 1997[58])) to estimate the confirmatory factor models. The software package used for the estimations was MPLUS 7 (Muthén, 2012[59]).

Confirmatory factor analyses were carried out for sets of conceptually related questionnaire items that measured one or more different dimensions. This approach allowed an assessment of the measurement model as well as of the associations between related latent factors. National samples received weights that ensured equal representations of countries in the analyses.

Once the dimensionality of the construct was confirmed, Cronbach's alpha was used to check internal consistency of each scaled index within the countries and to compare it between the countries.

Indices related to questions to staff members on the study child's development were then scaled using IRT scaling methodology, using a consistent approach to the scaling and construct validation in PISA (the following text in this section has been adapted from the PISA 2012 technical report (OECD, 2014[39]).

With the One-Parameter (Rasch) model (Rasch, 1960[51]) for dichotomous items, the probability of selecting category 1 instead of 0 is modelled as:

$$P_i(q) = \frac{\exp(q_n - d_i)}{1 + \exp(q_n - d_i)} \tag{15.3}$$

where $P_i(q_n)$ is the probability of person n to score 1 on item i. q_n is the estimated latent trait of person n and d_i the estimated location of item i on this dimension. For each item, item responses are modelled as a function of the latent trait q_n.

In the case of items with more than two categories (as for example with Likert-type items) this model can be generalised to the Partial credit model (Masters and Wright, 1997[60]), which takes the form of:

$$P_{x_i}(\theta_n) = \frac{exp\sum_{j=0}^{x}(\theta_n - \delta_i + \tau_{ij})}{\sum_{h=0}^{m_i} exp\sum_{j=0}^{h}(\theta_n - \delta_i + \tau_{ij})} x_i = 0,1,\ldots,m_i \quad (15.4)$$

where $P_{xi}(\theta_n)$ denotes the probability of person n to score x on item i out of the m_i possible scores on the item. θ_n denotes the person's latent trait, the item parameter δ_i gives the location of the item on the latent continuum and τ_{ij} denotes an additional parameter for each step j.

International item parameters were obtained using the ConQuest software (Adams and Wilson, 2015[61]). Weighted likelihood estimation (WLE) (Warm, 1989[62]) was used to obtain individual participant scores. The WLEs were derived using the ConQuest software (Adams and Wilson, 2015[61])with pre-calibrated item parameters.

WLEs were transformed to an international metric with an average of zero and a standard deviation of one. The transformation was achieved by applying the formula:

$$\theta'_n = \frac{\theta_n - \mu_{\theta(IELS18)}}{\sigma_{\theta(IELS18)}} \quad (15.5)$$

where θ'_n are the scores in the IELS metric, θ_n are the original weighted likelihood estimates in logits, and $\mu_{\theta(IELS18)}$ is the IELS mean of logit scores with three equally weighted country subsamples. $\sigma_{\theta(IELS18)}$ is the corresponding IELS standard deviation of the original weighted likelihood estimates.

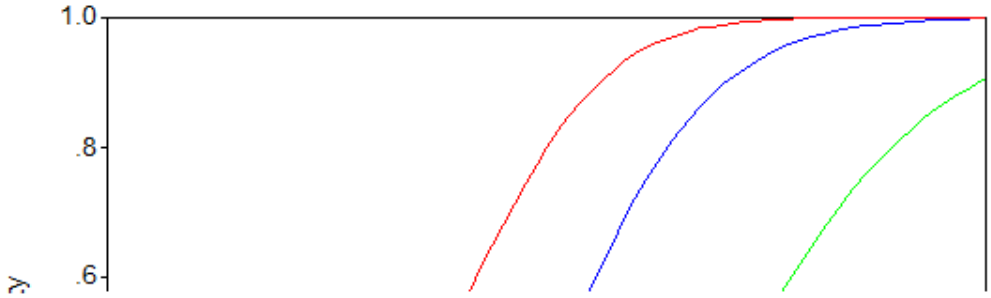
15.4. Describing questionnaire scale indices

As in PISA, IELS categorical items from the context questionnaires were scaled using IRT modelling. WLEs (logits) for the latent dimensions were transformed to scales with an average of 0 and a standard deviation of 1 (with equally weighted samples). It is possible to interpret these scores by comparing individual scores or group average scores to the mean, but the individual scores do not reveal anything about the actual item responses and it is impossible to determine from scale score values to what extent respondents endorsed the items used for the measurement of the latent variable. However, the scaling model used to derive individual scores allows descriptions of these scales by mapping scale scores to (expected) item responses.

Item characteristics can be described using the parameters of the partial credit model by summing for each category its probability of being chosen with the probabilities of all higher categories. This is equivalent to computing the odds of scoring higher than a particular category.

The results of plotting these cumulative probabilities against scale scores for a fictitious item are displayed in Figure 15.1. The three vertical lines denote those points on the latent continuum where it becomes more likely to score >0, >1 or >2. These locations, Γ_k , are Thurstonian thresholds that can be obtained through an iterative procedure that calculates summed probabilities for each category at each (decimal) point on the latent variable.

Figure 15.1: Summed category probabilities for fictitious item



Summed probabilities are not identical with expected item scores and have to be understood in terms of the probability to score at least a particular category. Other ways of describing the item characteristics based on the partial credit model are item characteristic curves (by plotting the individual category probabilities) and expected item score curves (for a more detailed description see (Masters and Wright, 1997[60]).

Thurstonian thresholds can be used to indicate those points on a scale for each item category, at which respondents have a .5 probability to score this category or higher. For example, in the case of Likert-type items with categories "Strongly disagree" (SD), "Disagree" (D), "Agree" (A) and "Strongly agree" (SA) it is possible to determine at what point of a scale a respondent has a 50 per cent chance to agree with the item.

15.5. Computation of scaled indices

15.5.1. Staff questionnaire

Staff ratings of child's global development

Staff members in Section B, were asked to indicate whether the child has capacity for nine different developmental skills (selecting from "Yes", "No" and "Not sure"). The nine items in the question were used to derive a scale of *staff reports of child's global development* (GLOBAL)[11] ("Not sure" was treated as missing for the purposes of scaling). Higher scale