



Education
Endowment
Foundation

Update of the EEF's Early Years Measures Database

Final report

September 2026

Sara Bashir Malik, Lydia Marshall, Anushay
Mazhar, Astrid Pickenpack, Rachel Outhred, and
Robert Wishart





The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, colleges, and early years settings to improve teaching and learning for 2- to 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by The Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were reendowed with an additional £137m from government, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:



Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP



0207 802 1653



info@eefoundation.org.uk



www.educationendowmentfoundation.org.uk



Department
for Education



Table of contents

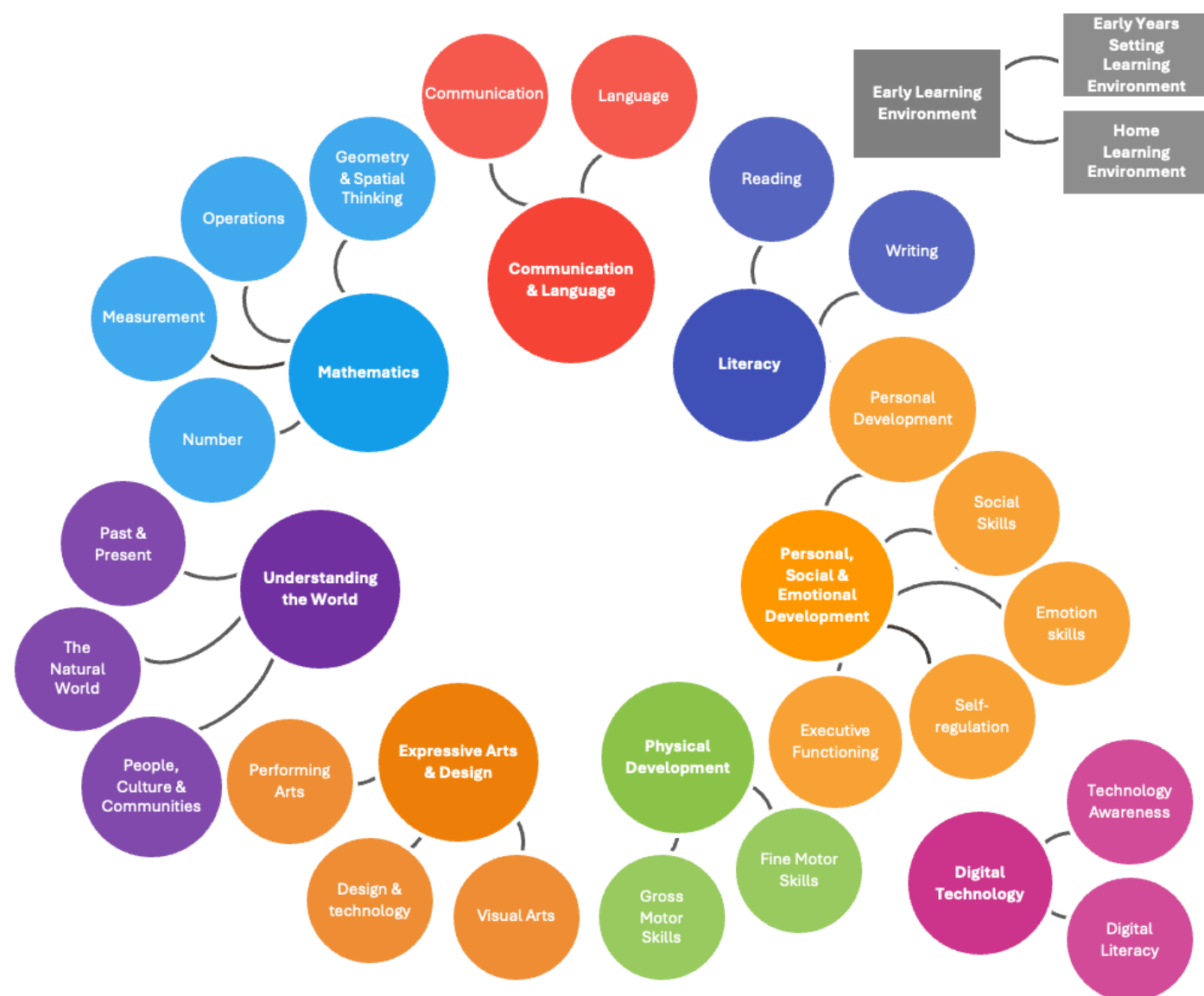
Executive summary	3
Background and review rationale	7
Methodology	9
Search results.....	18
Results of review.....	19
Findings.....	27
Implications	29
Project team	32
References	33
Appendix A: Taxonomy	35
Appendix B: Search terms.....	42
Appendix C: Changes and decisions following protocol publication	44
Appendix D: Extraction template	47
Appendix E: Proficiencies coverage by domain	58

Executive summary

This report presents an updated review of the Early Years Measures Database (EYMD) to strengthen its coverage, usability, and relevance for evaluators, researchers, and practitioners working in early childhood. Building on the original EYMD, we sought to identify, appraise, and clearly present measures assessing developmental outcomes for children aged from birth to six years.¹

We underpin the review with a revised EYMD taxonomy (Figure 1), which provides a structured way to organise measures by domain, sub-domain, and proficiency. The domains and sub-domains included in the EYMD are depicted below. The largest shapes represent domains with sub-domains emanating from them. Digital Technology and the Early Learning Environment are presented separately as underpinning learning and development in other domains.

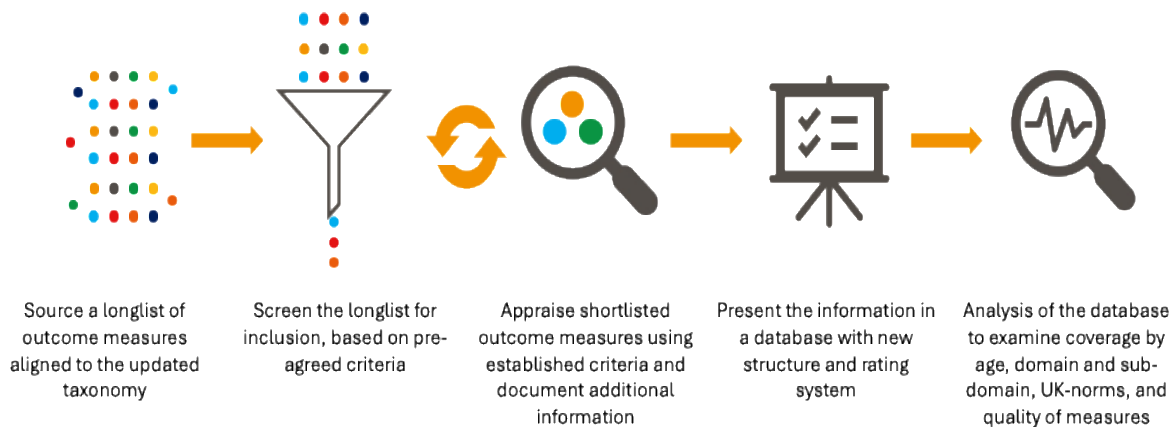
Figure 1: Revised EYMD taxonomy



¹ The Early Years Foundation Stage (EYFS) Statutory Framework in England focuses on children aged from birth to five years and has largely shaped the domains in the EYMD. Internationally, however, early childhood is typically defined as ages spanning from birth to six years, which is the age range adopted in the EYMD.

We used the following process for updating the EYMD (Figure 2).

Figure 2: Process for updating the EYMD

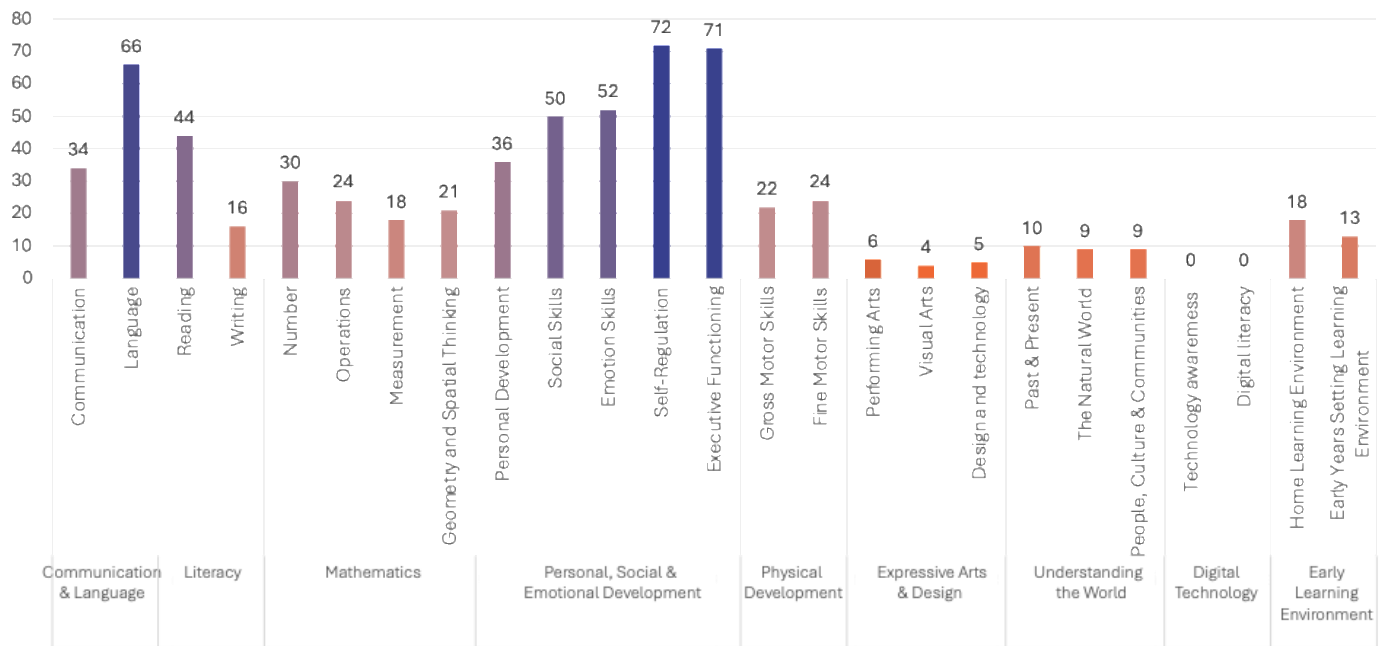


To source measures for the EYMD, we conducted a comprehensive identification process drawing on multiple sources, including existing EYMD entries, measures used in previous education trials, assessment publisher websites, the World Bank Early Childhood Development Measurement Inventory (World Bank, 2017), expert recommendations, and academic databases. Following screening and de-duplication, we shortlisted 253 measures for appraisal. We reviewed each shortlisted measure to assess the strength of psychometric evidence, including the size and characteristics of the standardisation sample, validity, and reliability. We also coded additional details such as sub-domains and proficiencies assessed, age range, administration format, time requirements, and training required for administration.

Alongside the review, we developed a set of recommendations to improve the usability, transparency, and interpretability of the EYMD. These recommendations focus on clarifying what measures assess, improving the presentation of psychometric evidence, and supporting users to navigate more intuitively between high-level domain overviews and detailed measure-level information.

Our findings show that coverage across domains is uneven. Measures are most prevalent in the prime areas of Communication and Language, and Personal, Social, and Emotional Development (PSED), and in the specific areas of Literacy and Mathematics, reflecting both the relative ease of standardised measurement in these areas and their prioritisation within academic and evaluation contexts. In contrast, we identified fewer measures for Expressive Arts and Design and Understanding the World. These domains encompass broad, context-dependent skills that are less suited to standardised assessment, particularly for children aged from birth to six years. Their development in these areas is typically observed during play, exploration, and creative activities rather than through structured assessment tasks with clearly defined responses. Measurement in these areas is also often less prioritised than in core academic domains like Literacy and Mathematics. The nature of available tools also reflects established measurement traditions. Literacy and Mathematics measures are largely task-based and child-completed, Early Learning Environment tools rely heavily on observation, and PSED measures frequently draw on adult report alongside child-completed measures. We identified no eligible measures for the Digital Technology domain.

Figure 3: Number of measures in the EYMD, by domain and subdomain



Turning to the availability of measures for different age groups, we found gaps for infants and toddlers, where available measures cluster within a narrow set of domains, primarily Communication and Language and Physical Development. While in some ways this reflects the importance of different skills at different stages of development, this concentration may also shape which outcomes evaluators assess in infancy and toddlerhood, limiting the ability to capture change across a broader range of developmental areas.

In terms of the quality of tools, evidence for construct validity was generally strong, indicating that most tools are grounded in clear theoretical frameworks. However, criterion validity was more uneven, particularly in Expressive Arts and Design and in Early Learning Environment, where external benchmarks are harder to establish. Reliability evidence showed a mixed pattern. Internal consistency was often strong when reported, but test-retest and inter-rater reliability were frequently missing. Early Learning Environment measures focus on contextual rather than child-level constructs. Their psychometric evidence is more variable and less consistently reported due to their multidimensional and observational nature.

Our findings highlight:

- domains where the EYMD can support evaluators and commissioners to make confident decisions around selecting measures that align with intervention aims and target age groups, while being aware of the practical constraints; and
- the need to apply psychometric standards proportionately across domains and exercise greater caution when interpreting ratings for outcome measures.

We identify the following priorities for future tool development:

- address gaps in domains, particularly in Expressive Arts and Design, Understanding the World, and Digital Technology;
- expand coverage of measures targeting historically under-measured proficiencies, and the adaptation of existing tools to capture more advanced concepts at younger ages, particularly to mitigate the risk of ceiling effects when such proficiencies are omitted;
- prioritise multi-respondent approaches to strengthen validity across contexts;

- age-appropriate measures for a wider range of proficiencies and domains for children aged from birth to two years;
- United Kingdom norming and validation beyond core academic and language domains, especially in PSED; and
- psychometrically robust measures of the Early Learning Environment.

For evaluators and commissioners of early years trials, including the Education Endowment Foundation (EEF), the results reinforce:

- the importance of aligning outcome selection with intervention intent;
- the use of bespoke tools for under-measured areas;
- the potential of playing a market-shaping role in stimulating new measure development where gaps persist;
- the need for piloting and robust assessor training where observational measures with weak evidence of reliability are used;
- the value in making technical information used to justify outcome selection more visible within trial reports to support cumulative learning across the field; and
- triangulating across respondent types to provide a more balanced and robust picture of impact than relying on a single assessment mode.

Overall, the updated EYMD enables more informed and transparent selection of outcome measures for early years research, evaluation, and practice.

Background and review rationale

Background

In 2014, the Education Endowment Foundation (EEF) commissioned the Institute of Education at University College London (UCL) to review measures designed to assess children aged from birth to six years (Dockrell *et al.*, 2017). This resulted in the launch of two key resources in 2017: i) the **Early Years Measures Database (EYMD)**, which catalogues tools for assessing developmental domains including language and communication, social and emotional development, literacy, and numeracy; and ii) an **accompanying document**, which summarises the findings of the review and how the EYMD was developed, and offers some guidance on what to consider when choosing measures to evaluate children's abilities. This review was an initial attempt to identify measures that would be most relevant to the EEF trials in the early years and was not intended to be comprehensive in its coverage of how outcomes for early years should be measured.

In 2025, the EEF carried out an internal review to explore whether there was a need to update the original EYMD and what this update would look like. Feedback from different stakeholders in the internal review highlighted gaps in coverage, usability challenges, and the need to expand its scope, showing that the database in its current form did not meet research needs or support practice effectively. These findings underlined the need for an update to ensure the EYMD remains a relevant, effective, and comprehensive resource.

In 2022, the EEF received a reendowment from the Department for Education (DfE), which extended its early years remit both in terms of the age range of children its work could support (two-year-olds and older) and the type of early years settings it aimed to work with (school-based, private, voluntary, and independent [PVI], and childminder settings). As the original EYMD had limited coverage for two-year-olds and was primarily focused on school-based contexts, this expansion created a clear need for additional measures that could capture outcomes for younger children and across a wider range of early years settings. This extension, along with a collaboration with the DfE Stronger Practice Hubs to support education recovery following the Covid-19 pandemic, notably increased the number of early years evaluations commissioned by the EEF.² This work was the start of a shift in the organisation's strategic priorities to focus more strongly on the early years. The EEF's early years strategy includes a focus on building expertise across the EEF's evaluator panel in conducting early years evaluations and aims to facilitate high-quality research through the development of better evaluation infrastructure.

The United Kingdom (UK) Government's role and emphasis on the early years have also shifted in recent years. There have been several updates to the DfE's **Early Years Foundation Stage (EYFS) framework** in 2021 and more recently in September 2025 (DfE, 2021; DfE, 2025a). The EYFS framework sets out the statutory requirements for teaching and learning in the early years, structured across seven areas of learning.³ It culminates in the Early Years Foundation Stage Profile (EYFSP), a teacher-completed assessment at the end of the Reception year, which assesses children against the Early Learning Goals (ELGs). The ELGs define the expected level of development for children at the end of the Reception year across each area of learning, and determine whether they have reached a 'good level of development'⁴ based on expected attainment across key areas of learning. Wider government investment and policy initiatives have further underlined the centrality of early years (DfE, 2025b). The '**Best Start in Life**' strategy has placed emphasis on improving early development and ensuring more children reach a good level of development by age five in line with the UK Government's Opportunity Mission, which aims to increase the proportion of children achieving this to 75% by 2028.

Alongside these policy shifts, a growing body of evidence underscores the importance of the earliest years of life. Development in the first three years of life is widely recognised as critical for later outcomes (e.g. see Tierney and Nelson, 2009). Early development is driven by rapid neural connections shaped by experience, environment, nutrition, safety, and responsive stimulation. These foundations influence lifelong learning and well-being (Tierney and Nelson, 2009), as well as

² At the time of writing, the EEF has completed 36 evaluations in the early years space, with a further 17 evaluations ongoing.

³ The seven areas of learning include the three prime areas (Communication and Language, Physical Development, and PSED) and four specific areas (Literacy, Mathematics, Understanding the World, and Expressive Arts and Design).

⁴ Children are defined as having reached a good level of development at the end of the EYFS, if they have achieved the expected level for the ELGs in all the prime areas of learning and the specific areas of Mathematics and Literacy.

other outcomes including earnings in adulthood (Britto *et al.*, 2017). Yet despite this, valid and reliable measures for the birth to three years age group remain relatively scarce, leaving a major gap in the EYMD.

Taken together, these policy developments, funding reforms, and scientific insights strengthen the case for updating the EYMD. In particular, they point to the need both to expand coverage for the youngest children, an age group underrepresented in the original database, and to ensure alignment with the EYFS. Considering this context and these evolving priorities, the EEF has commissioned Oxford MeasurEd to update the EYMD.

Project aims and objectives

The overall aim of the project was to update the original EYMD. Specifically, the project had four objectives, including:

- **Defining domains.** Expanding the EYMD to include and clearly define additional domains and refine existing domain labels, with the aim to ensure better alignment with the EEF other early years resources such as the **Early Years Evidence Store** and national frameworks including the EYFS framework.
- **Identification and review of new and existing measures.** Including:
 - identifying a new longlist of measures to add to the database, including suitable measures to assess children aged birth to three years;
 - for a shortlist of measures for which the information was available, assessing their psychometric properties, and reporting administration characteristics, alignment with EYFS **ELGs**, and presence in any national administrative database;
 - updating the information on all existing measures in the database to align with the new structure of the database and ensure that the information is up to date;
 - examining how well different developmental domains and age groups are covered by existing measures, identifying gaps where measures are lacking; and
 - exploring the suitability of internationally developed measures where UK-specific measures are lacking.
- **Improving usability and presentation of EYMD.** Making recommendations to improve the usability of the EYMD, including the database interface, individual measures' pages, and rating systems for psychometric properties.
- **Developing guidance and capturing learning.** Producing resources to guide the EEF and its partners in effectively using the EYMD, including practical tips and insights from previously published funded trials by the EEF, and guidance for the EEF on how to capture such insights moving forward.

Methodology

The methodology for this review comprised of three integrated strands of work: i) refining the EYMD taxonomy and structure; ii) updating the EYMD content and developing guidance and resources; and iii) revising the EYMD design interface. Each strand combined desk-based review, stakeholder consultation, and iterative refinement in close collaboration with the EEF.

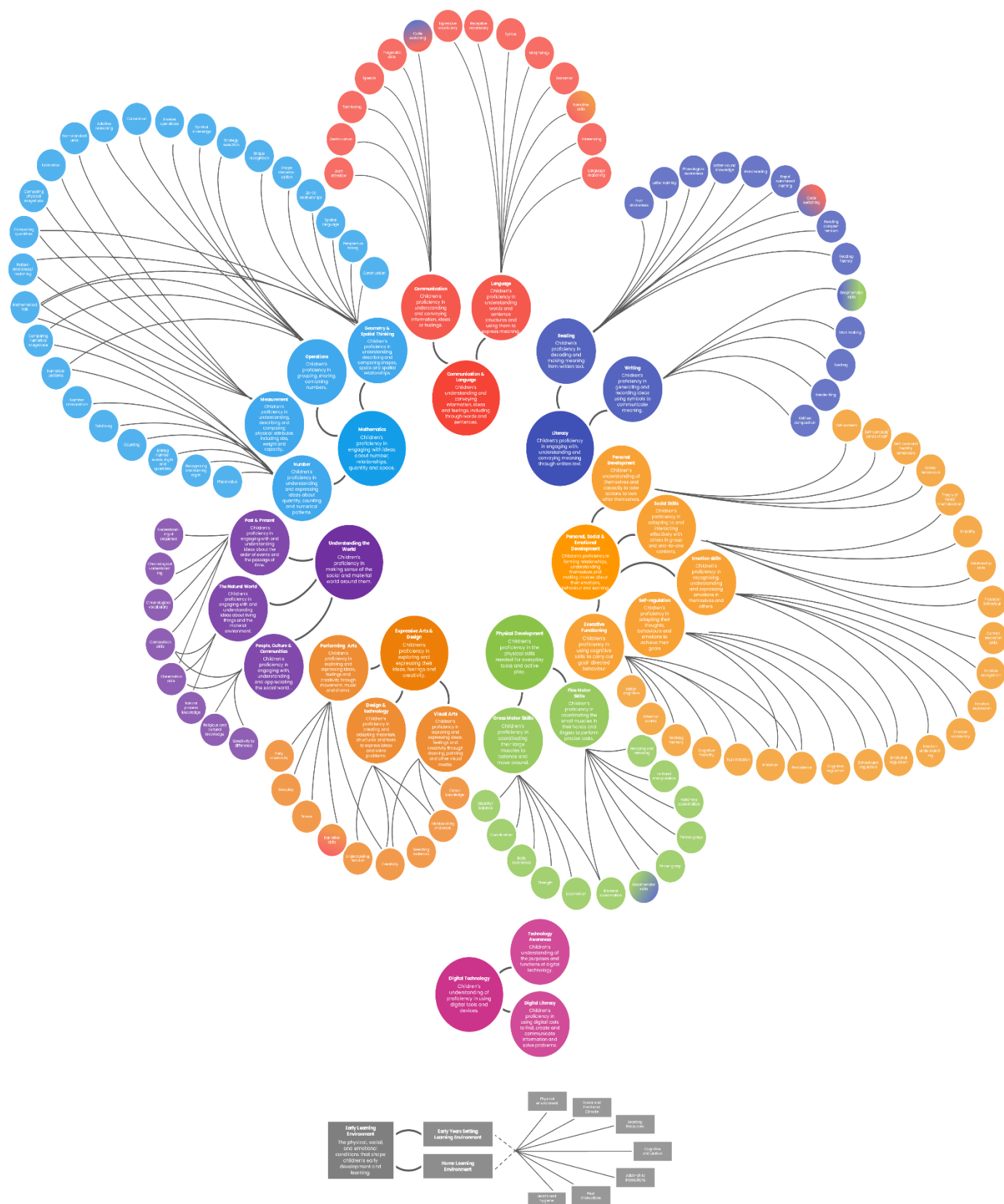
Refining the EYMD taxonomy and structure

Development of the taxonomy

This work strand strengthened the EYMD by aligning it with the EYFS framework and the related resources by the EEF, and by refining the underlying taxonomy to provide clearer and more robust domain definitions. We agreed with the EEF that alignment with the EYFS as a priority and used it as the primary structure for the database, complemented by evidence from research literature, global frameworks, and stakeholder input to ensure theoretical coherence. We cross-referenced existing EYMD domains with the EYFS and the EEF resources and refined them through a targeted desk review. This informed a more detailed taxonomy structured around domains, sub-domains, and proficiencies. Sub-domains refer to distinct, theoretically grounded components within a broader developmental area, while proficiencies represent observable behaviours or capacities within each sub-domain that might be measured through assessment tools. While the taxonomy largely reflects the EYFS structure, we also introduced proficiencies related to Digital Technology and included the Early Learning Environment as a cross-cutting domain to capture contextual factors that underpin learning and development across other domains. The taxonomy provides a consistent framework to help users to select tools that best align with intended outcomes, compare measures, and identify gaps in coverage. We validated the taxonomy through expert consultation and agreed it with the EEF. We used the taxonomy to tag all measures in the EYMD, enabling more precise searching and filtering.

The final taxonomy of domains, sub-domains, and proficiencies is set out in Figure 4, with the definitions for each domain, sub-domain, and proficiency presented in Appendix A. We explain the minor changes we made to the taxonomy following the publication of the protocol (Malik *et al.*, 2025) in Appendix C.

Figure 4: The EYMD taxonomy

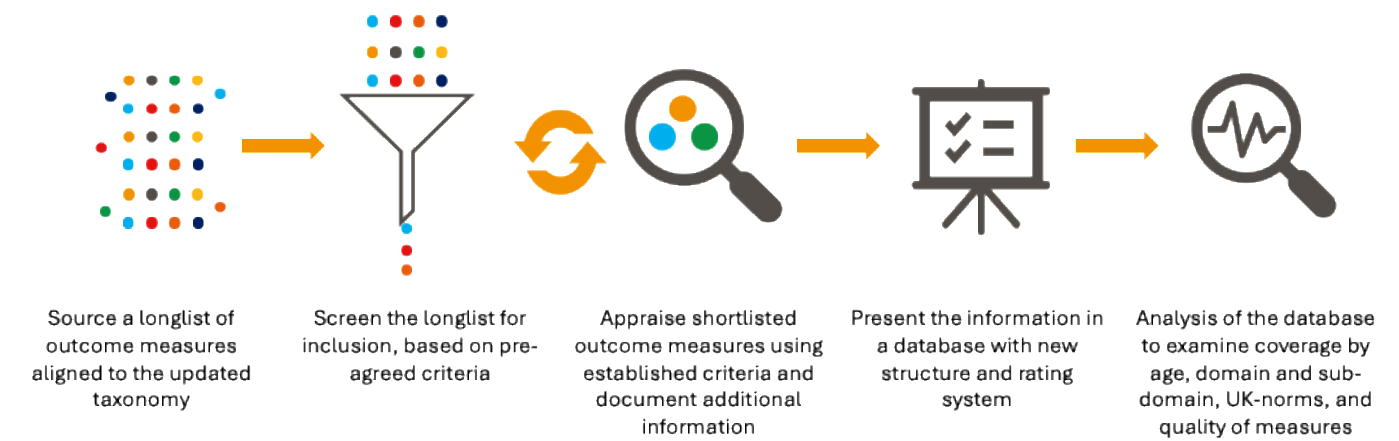


The largest shapes represent domains, with sub-domains emanating from them and proficiencies linked to the sub-domains. Some proficiencies are linked to more than one sub-domain, shown with lines going from the proficiencies to more than one sub-domain. Where a proficiency is attributed to more than one domain, we have used graduated colouring to illustrate this. Digital Technology proficiencies and the Early Learning Environment are presented as underpinning learning and development in other domains.

Updating the EYMD content

This work strand involved identifying new outcome measures for inclusion in the EYMD, systematically documenting and appraising shortlisted measures, and mapping measurement coverage. The process for updating the EYMD is given in Figure 5.

Figure 5: Process for updating the EYMD



Sourcing a longlist of outcome measures

The first stage of the review focused on developing a comprehensive longlist of available outcome measures aligned to the updated EYMD domains. We agreed broad inclusion criteria with the EEF and applied them to ensure comprehensive but purposeful coverage. Between September and October 2025, we identified measures through multiple sources, including the EEF- and DfE-funded evaluations, earlier desk-based work, targeted searches of major assessment publishers (see Appendix C for list of publishers searched), the World Bank’s Early Childhood Development Measurement Inventory (World Bank, 2017), expert recommendations, relevant academic research databases using predefined search terms (see Appendix B) and the existing EYMD, which was used to ensure continuity with earlier versions of the database. For the academic research databases, an initial pilot search was conducted using Google Scholar to identify a longlist of measures for the sub-domains of Self-Regulation and Executive Function prior to the publication of the protocol. This involved applying the predefined list of search terms and screening the first ten pages of results. While this approach proved useful for identifying some relevant measures, it became clear that relying on manual Google Scholar searches would be time-consuming and less systematic than alternative approaches. We moved to a more systematic academic database search process managed through EPPI-Reviewer (see Appendix C).

Using the predefined search terms for each domain, **searches were conducted in PsycINFO via PsycNET.** The resulting records were exported and imported into EPPI-Reviewer, where they were organised for screening. Once records from all domains had been uploaded, duplicates were identified and removed. The remaining records were screened based on titles and abstracts. During this stage, records were flagged where the name of a specific assessment measure was explicitly mentioned, as well as those that referred to the use of an assessment tool without clearly naming it. In cases where a measure was referenced but not named, the source was subsequently located using Google Scholar in order to identify the instrument and add it to the list of candidate measures for the longlist.

We consolidated measures identified from different sources into a single dataset and removed duplicate assessments to ensure that each measure was represented only once in the longlist. We described the detailed approach to identifying a longlist of outcome measures in the **protocol** (Malik *et al.*, 2025) and outline changes made to the approach following the publication of the protocol in Appendix C.

Table 1 summarises **the original criteria for inclusion of measures in the longlist and the additional exclusion criteria applied throughout the review to contain the scope of the database.** It clarifies which types of measures we included in the EYMD and which we excluded to support consistent interpretation and transparency in decision-making.

Table 1: Inclusion and exclusion criteria used during longlisting

Original inclusion criteria
• Targets one or more of the domains in the updated EYMD taxonomy • Published since the 2017 review^a • Suitable for children aged birth to six years • Available in English
Additional exclusion criteria
• Discontinued tools, e.g. Early Maths Diagnostic Assessment (EMDA), or those that are not published yet • Measures available only within a single research study without wider availability, e.g. Study of Early Education and Development (SEED) • Tools that only support the recording and analysis of behaviour in behavioural research or diagnosis but are not intended as measures of children's development, e.g. Multi-Option Observation System for Experimental Studies (MOOSES™) • Measures focused on parental beliefs and self-efficacy rather than child-level or learning environment outcomes, even where they were widely used in practice, e.g. Brief Parental Self-Efficacy Scale (BPSES) tool • Sensory processing measures used primarily for clinical or diagnostic purposes, e.g. Sensory Profile School Companion (SPSC)

^a This criterion only applies to newly identified measures. Measures that were included in the longlist from the existing EYMD were not excluded because they were published before 2017.

We included the latest available version of each assessment wherever possible. Where the latest version lacked published validation evidence, we used psychometric data from the most recent validated version. We entered measures in the database once, even if they comprised multiple subtests, to avoid duplication and misrepresentation of the evidence base.

Screening to produce a shortlist

We screened the longlisted measures to produce a shortlist for detailed appraisal based on the availability of reliability and validity evidence, administration guidance, and scoring criteria. In contrast to the 2017 review and in agreement with the EEF, we did not apply availability of UK norms as an initial filter, allowing retention of internationally developed measures with documented use in UK contexts. This was particularly important for domains such as the Early Learning Environment domain, where well-established tools are often developed outside the UK. We recorded reasons for exclusion at this stage systematically. Table 2 lists the criteria used to shortlist measures in each round of review.

Table 2: Shortlisting criteria 2017 and 2025

Criteria	Original requirement in 2017	Updated requirement for 2025 review
UK standardisation sample	Required for norm-referenced assessments	Not required. Measures were eligible if either UK-normed or validated on a documented sample (UK or international), even where formal norms were not available
Validity information	Required	Required
Reliability information	Required	Required
Administration guidelines	Required	Required
Scoring criteria	Required	Required

While the protocol (Malik et al., 2025) anticipated the need to prioritise measures if the number eligible for appraisal exceeded available resources, this was not required. Following initial screening, the shortlist comprised approximately 250 measures, which aligned with the number of measures that could be appraised with the available resources. As a result, we agreed with the EEF that no additional prioritisation criteria or further tightening of inclusion thresholds was necessary.

However, once the detailed appraisal (see below) began, **we identified a small number of measures that, despite being included in the longlist, were not particularly relevant for the scope of the EYMD.** We therefore applied the additional shortlisting criteria listed in Table 3.

Table 3: Additional shortlisting criteria

Additional shortlisting criteria
Exclude clinical, diagnostic, or therapeutic tools focused on assessing symptoms or identifying disordered or atypical behaviour rather than assessing the development of proficiencies or competencies relevant to early education and intervention, unless they directly operationalised developmental constructs relevant to the EYMD taxonomy
Exclude measures intended to screen for abuse, trauma, or severe emotional or behavioural disturbance
Exclude measures with relevant psychometric evidence only for populations outside the target age range
Exclude measures that are not available for purchase in the UK

Appraising the psychometric properties of shortlisted measures

We appraised shortlisted measures using a structured framework adapted from the European Federation of Psychologists' Associations (EFPA) test review model.⁵ We assessed key psychometric properties, including standardisation, validity, and reliability. We rated evidence for each property on a three-point scale (see Table 4⁶) and provided a brief narrative justification for the rating.

We based appraisals primarily on publicly available sources, including publisher documentation and peer-reviewed studies. We made certain exceptions to this rule, which are outlined in Appendix C. Where multiple sources existed for a measure, we prioritised the most relevant evidence, typically the most recent UK-based study, followed by the largest or most comparable sample. Where documentation was available in multiple languages, we prioritised English-language sources. We reported evidence from the EEF-funded trials separately to capture learning from use in practice, including issues such as floor or ceiling effects.

Where standard validity statistics (such as correlation coefficients) were unavailable, we used alternative but comparable forms of validity evidence where appropriate (e.g. classification accuracy relative to another validated measure). In such cases, we applied ratings conservatively and clearly justified them in the accompanying narrative.

Table 4: Criteria for psychometric appraisal

Criterion (relevant pages in EFPA, 2013 ⁷)	Factor assessed	Rating
Standardisation sample (pp. 27, 35)	UK standardisation sample size	N/A: Not standardised N/A: Standardised with a non-UK population ✓ Limited: <200 ✓✓ Moderate: 200–999 ✓✓✓ Strong: ≥1,000

⁵ We provided full details of how we adapted the EFPA review model in the **protocol** (Malik *et al.*, 2025).

⁶ We have relabelled the ratings as limited, moderate, and strong to resolve inconsistencies in the protocol (Malik *et al.*, 2025) and to use terminology that more clearly conveys differences in the strength of evidence, rather than evaluative labels such as inadequate, good, or excellent.

⁷ The EFPA Test Review Criteria were largely modelled on the form and content of the British Psychological Society's (BPS) test review criteria and criteria developed by the Dutch Committee on Tests and Testing (COTAN) of the Dutch Association of Psychologists (NIP). EFPA is grateful to the BPS and NIP for permission to build on their criteria in developing the European model. All intellectual rights in the original BPS and NIP criteria are acknowledged and remain with those bodies.

	UK standardisation sample characteristics	N/A: Not standardised N/A: Standardised with a non-UK population ✓ Limited: No attempt was made to collect a random or stratified sample. Sample may be entirely opportunistic (e.g. convenience samples) Demographic or contextual variables (e.g. age, gender, ethnicity) are not reported or show clear imbalance relative to the target population ✓✓ Moderate: Some attempt was made to recruit a balanced sample (e.g. stratification, quotas) but balance was not fully achieved. Sample may be broadly diverse but still underrepresents certain groups or demographics are not reported after random sampling model used ✓✓✓ Strong: Data were gathered by means of a random sampling model and good representativeness is established with respect to variables such as gender, age, ethnicity)
Validity (pp. 54–60)	Construct validity: How well a test measures the psychological construct it claims to assess	✓ Limited: No evidence that items and test were developed based on theory and construct definition OR statistical tests (e.g. Factor Analysis, Item-test correlations) demonstrate lack of construct validity (see EFPA guidance for further detail) ✓✓ Moderate: Evidence that items and test were developed based on theory and construct definition ✓✓✓ Strong: Evidence that items and test were developed based on theory and construct definition AND statistical tests demonstrate construct validity and confirm construct validity (see EFPA guidance for further detail)
	Criterion validity: How well test scores correlate with relevant external criteria	N/A: Not criterion referenced ✓: Limited ($r < 0.20$) correlation with the criterion measure OR small (< 100) sample size ✓✓: Moderate ($0.20 \leq r < 0.50$) correlation with the criterion measure AND adequate sample size (≥ 100) ✓✓✓: Strong ($r \geq 0.50$) correlation with the criterion measure AND adequate sample size (≥ 100); OR Good ($0.35 \leq r < 0.50$) correlation with the criterion measure shown in one large sample (> 200) or more than one adequate sized study
Reliability (pp. 43–51)	Internal consistency: How well the items work together to measure the same underlying construct	Not reported ✓: Limited ($r < 0.70$) reliability coefficient OR small (< 100) sample size ✓✓: Moderate ($0.70 \leq r < 0.80$) reliability coefficient AND adequate sample size (≥ 100) ✓✓✓: Strong ($r \geq 0.90$) reliability coefficient AND adequate sample size (≥ 100); OR Good ($0.80 \leq r < 0.90$) reliability coefficient in one large sample (> 200) or more than one adequate sized study
	Test retest reliability: How well a measure produces consistent scores when administered to the same individuals at different points in time	Not reported ✓: Limited ($r < 0.60$) reliability coefficient OR small (< 100) sample size ✓✓: Moderate ($0.60 \leq r < 0.70$) reliability coefficient AND adequate sample size (≥ 100) ✓✓✓: Strong ($r \geq 0.80$) reliability coefficient AND adequate sample size (≥ 100); OR Good ($0.70 \leq r < 0.80$) reliability coefficient in one large sample (> 200) or more than one adequate sized study
	Inter-rater reliability: How well different raters or observers produce consistent scores when assessing the same individuals or behaviours	Not reported ✓: Limited ($r < 0.60$) reliability coefficient (pairwise correlation) OR small (< 100) sample size ✓✓: Moderate ($0.60 \leq r < 0.70$) reliability coefficient AND adequate sample size (≥ 100) ✓✓✓: Strong ($r \geq 0.80$) reliability coefficient AND adequate sample size (≥ 100); OR Good ($0.70 \leq r < 0.80$) reliability coefficient in one large sample (> 200) or more than one adequate sized study

In addition to psychometric quality, we recorded key descriptive and implementation information for each measure, including alignment with the EYMD taxonomy, administration and scoring requirements, time burden, and cost, to support informed use.

We classified and assigned measures to domains and sub-domains according to the EYMD taxonomy, even where alternative domain frameworks were referenced in source documentation such as in the World Bank's Early Childhood Development Measures Inventory (World Bank, 2017). We undertook reclassification only where there was a clear construct definition, scale structure, and supporting documentary evidence, to avoid inaccurately assigning a domain or sub-domain to a measure. For example, a measure described in external documentation as assessing 'self-regulation' may be found to capture behaviours such as co-operation, sharing, peer interaction, and conflict resolution, on closer review of its construct definition and item content. In cases such as these, the measure would be reclassified under the social skills sub-domain within Personal, Social, and Emotional Development (PSED) rather than self-regulation.

Similarly, we primarily classified **the specific proficiencies measured by the tools based on whether test developers named specific proficiencies in available documentation. Where the proficiencies were not explicitly named by the test developers, we allocated proficiencies based on our judgement** after reviewing items and/or how the test developers described the constructs in available documentation. Allocated proficiencies should therefore be read as indicative rather than definitive.

The extraction template used to record this information is given in Appendix D.

Analysis of findings

Following the longlisting, shortlisting, and psychometric appraisal of outcome measures, we presented the information outlined above in an Excel database, structured and coded to be searchable by the final domains, sub-domains, and proficiencies, as well as by psychometric appraisal and key implementation characteristics such as target age group and mode and format of administration.

We then analysed the database to examine:

- the volume of measures, comparing longlisted and shortlisted tools by domain;
- age coverage, including the distribution of measures across domains and age groups;
- the availability of UK-normed measures;
- coverage across sub-domains and proficiencies, relative to the revised EYMD taxonomy; and
- the quality of measures, based on the psychometric appraisal.

The 'Results of review' section of this report presents the results of these analyses, including findings on the availability, coverage, and quality of early years assessment tools across domains, sub-domains, and age groups. In the 'Implications' section, we discuss the implications for future tool development and use.

Revising the EYMD design interface

Approach

In parallel with the content updates, we **redesigned the EYMD interface to reflect our recommended changes (see below) with the aim to improve usability, interpretability, and alignment with the needs of its users**. The design process followed the approach set out in the **protocol** (Malik *et al.*, 2025). We developed and iteratively refined mock-ups, consulted with the EEF staff, and conducted user testing with evaluators through surveys and workshops. We used feedback from these activities to identify limitations in the existing EYMD interface and to inform successive design iterations.

Key recommended changes

We recommended changes to the existing EYMD to address challenges reported by users, including difficulty navigating the database, limited transparency around what measures assess, and challenges interpreting psychometric

information. These recommendations proposed that the interface follows the revised EYMD taxonomy, enabling users to browse and filter measures by domain, sub-domain, and proficiency, and to move more intuitively between high-level domain overviews and detailed measure-level information. We recommended search and filtering functions that reflect the criteria evaluators most commonly used when selecting measures, such as age range, time required, and use in UK contexts.

We also recommended and implemented changes to the presentation of ratings and supporting information to improve clarity and consistency. Based on our recommendations, we agreed with the EEF to present psychometric evidence in a structured and transparent way, combining high-level summary judgements with narrative justifications to support interpretation. We recommended retaining information icons and hover features in a purposeful manner to provide concise definitions and guidance on technical terms with non-technical users in mind. We also recommended redesigning the individual measures pages to bring together information on construct coverage, implementation considerations, and psychometric properties in a consistent layout, addressing feedback that relevant information was previously fragmented or difficult to interpret. The recommended changes to the design are supported by a **practical user guide on how to use the EYMD** (Malik *et al.*, 2026) and the methodology and taxonomy underpinning the database.

Methodological reflections

We made methodological choices during the development of the protocol and the review itself, which involved deliberate trade-offs between breadth, depth, feasibility, and usability. We include these reflections to support appropriate interpretation of the database and to inform future maintenance and development of the EYMD as a living resource.

Strengths of the approach

- The use of a multi-source, multi-stage methodology that combined academic database searching, targeted grey literature searches, expert consultation, and systematic appraisal. This approach **reduced reliance on any single source of evidence and increased the likelihood of identifying a range of measures relevant to early years evaluation**. The use of EPPI-Reviewer to manage academic searches supported transparency and reproducibility.
- We explicitly aligned the EYMD taxonomy with the EYFS framework. Using the EYFS as the primary organising structure **ensured policy relevance and increased the practical utility of the database** for evaluators, practitioners, and policymakers working in English early years contexts. At the same time, the taxonomy was informed by developmental research and refined through expert consultation, helping to balance policy alignment with theoretical coherence.
- The systematic inclusion of live links and clearly referenced source materials for each measure wherever possible. During consultations with evaluators on the EYMD design, evaluators emphasised that access to documentation that backs the information in the database **inspires confidence in the database and supports informed decision-making**. Including direct links enables users to verify information, interrogate psychometric evidence in greater depth, and assess suitability for specific contexts rather than relying solely on summary ratings.
- **The database provides a foundation for more systematic capture of learning from the use of tools in the EEF-funded evaluations**, enabling practical insights from implementation to sit alongside formal psychometric evidence and inform future updates.

Limitations

- **The review relied primarily on publicly available information to appraise psychometric properties.** While this supports transparency and replicability, it inevitably constrains the depth of appraisal for measures where validation studies are unpublished, incomplete, or behind paywalls. Although targeted purchases of technical manuals helped address this limitation for a small number of high-priority tools, gaps in accessible evidence remain.
- **We also mapped tools to the taxonomy based only on available documentation.** In cases where access to full instruments and scoring materials, or resources for item-level review, were not available, the

proficiencies assessed by some tools may not have been captured, potentially leading to under-representation of certain proficiencies in the database.

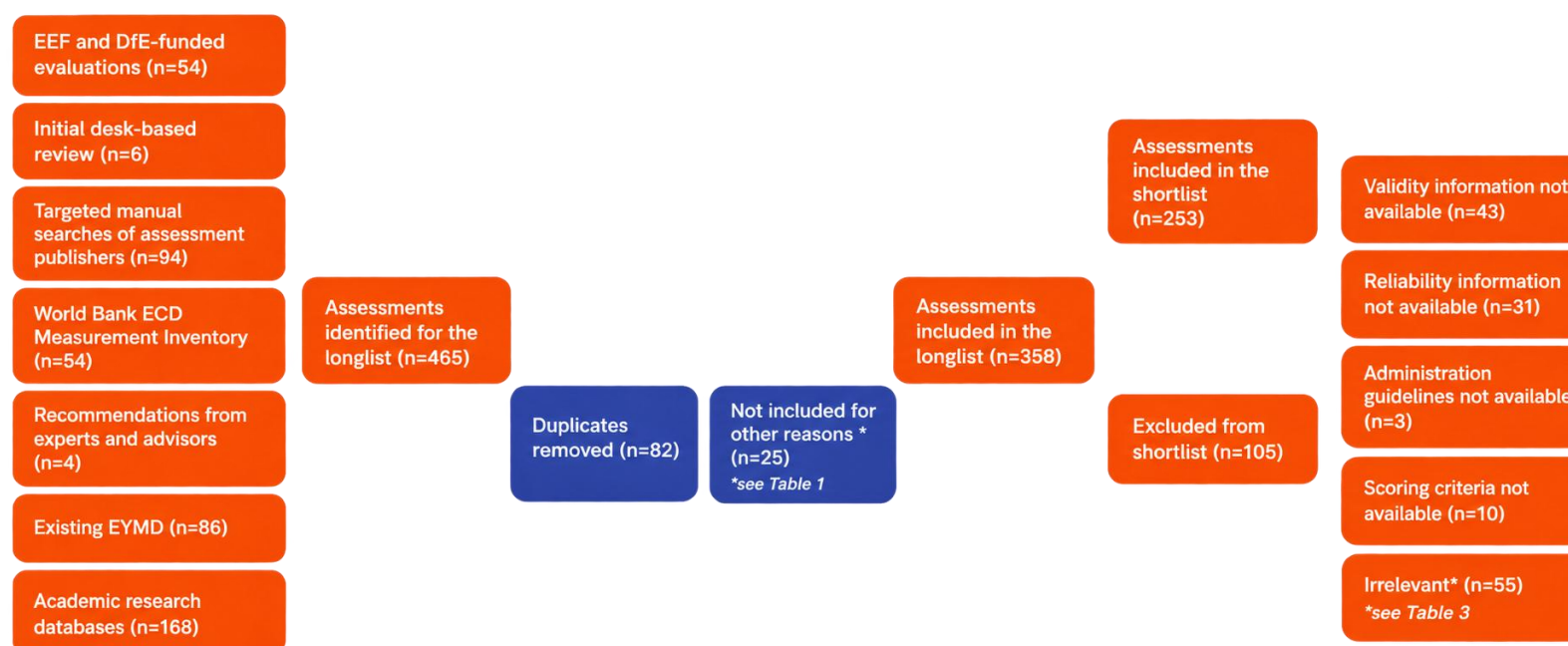
- While expert consultation played a critical role in refining the taxonomy, **expert input varied across domains**. Timing constraints and the availability of specialists meant that some areas received more detailed scrutiny than others. Input was more limited in domains such as Mathematics, Physical Development, Expressive Arts and Design, and Understanding the World, and no dedicated consultation was conducted for some of these domains. However, we did follow up with discussions with theme leads at the EEF to further refine the sub-domains, skills, and definitions within Mathematics in particular.
- The **taxonomy intentionally avoids exhaustive proficiency lists** to preserve usability. While this supports clarity and ease of navigation of the database, it also means that some developmental nuances are subsumed or collapsed, and not all theoretically relevant proficiencies are explicitly named. Users should therefore interpret the taxonomy as a classification and discovery tool rather than a comprehensive model of early childhood development.
- While search terms included were intended to include measures for children with English as an Additional Language (EAL) and bilingual learners, **evidence on the use and validity of identified measures with children who have limited or no English was limited**. Where the sample for validity or standardisation of measures included these groups, this has been noted where possible. However, reporting depended on the level of detail available in source materials and is reflected selectively to maintain conciseness. Many measures, particularly those relying on verbal instructions or responses, may be sensitive to children's language proficiency rather than the intended construct. As a result, findings should be interpreted with caution where language demands are high. This highlights an important gap in the evidence base given the linguistic diversity of early years settings in England.
- When reporting the inter-rater reliability, **we do not distinguish between instances where inter-rater reliability is not reported and where it is not applicable**. For standardised, objectively scored assessments in domains such as Literacy, Mathematics, and Understanding the World, scoring consistency is ensured through fixed marking schemes and inter-rater reliability is less meaningful. As a result, classifying these cases as 'not reported' overstates gaps in the evidence base in contexts where this metric is not applicable.
- **Using live links to external sources supports transparency but introduce ongoing maintenance challenges**. Publisher websites, academic repositories, and online reports may change location, move behind paywalls, or be removed over time. Without active monitoring, broken or outdated links risk undermining user confidence and the perceived reliability of the database.

Search results

Flow of assessments identified

Figure 6 provides an overview of the flow of measures through the identification, screening, and shortlisting stages described in the 'Methodology' section above. It summarises the volume of measures identified from each source, the number removed at each stage, and the reasons for exclusion during screening. In total, we identified 358 outcome measures⁸ for inclusion in the longlist across all domains. The final shortlist comprised 253 measures that met the criteria for detailed psychometric appraisal.

Figure 6: Flow diagram of assessments identified^a



^a The counts shown in the figure for individual reasons for exclusion do not sum to the total number of assessments excluded from the shortlist (n=166), as some assessments met more than one exclusion criterion

⁸ An additional measure was identified and added to the database after this report was finalised. Consequently, it is not included in the figures, counts, or analyses presented in this report. Its inclusion does not alter the overall findings or conclusions of the report.

Results of review

Description of database

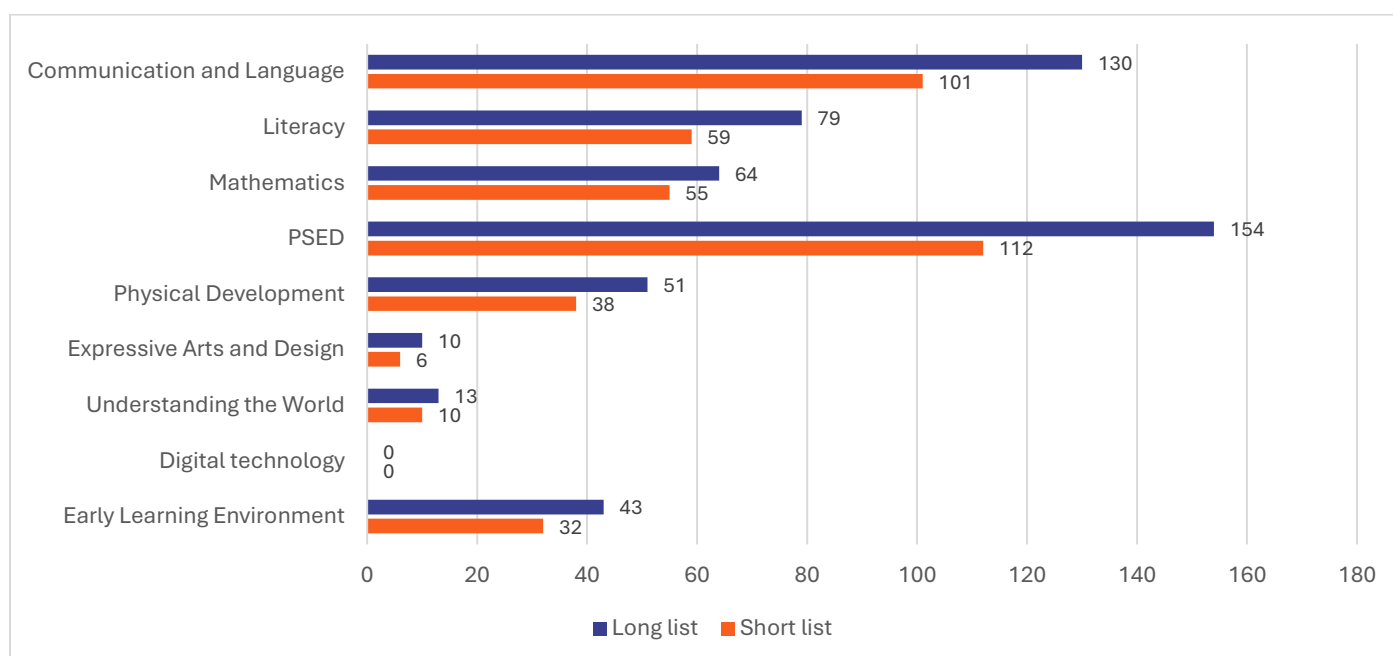
Volume of measures: Longlist and shortlist by domain

Across all domains, we identified substantial numbers of measures at the longlisting stage. Figure 7 shows that PSED, Communication and Language, and Literacy had the largest number of measures, reflecting the centrality of these domains in early years research and evaluation. Mathematics and Physical Development were also well represented, while Expressive Arts and Design and Understanding the World had comparatively fewer measures. We identified a moderate number of Early Learning Environment measures, indicating growing but still limited availability of tools in this area.

We did not identify measures for Digital Technology at either stage, highlighting a clear gap in the measurement landscape. We did not include this domain in the initial iterations of the taxonomy, but expert consultations highlighted that Digital Technology increasingly underpins learning and development across other domains. This interpretation is consistent with wider sector evidence. Screen use is now nearly universal in early childhood, and the UK Government has recently announced new guidance on children's screen use, with strong calls from the sector for this to go beyond the home and recognise the role of early years settings, as well as to address digital literacy and online safety rather than screen time alone (Early Years Alliance, 2026). It may also reflect our positioning of the domain as a means of supporting outcomes in other domains (e.g. computer games used to support self-regulation or e-readers used to facilitate interactive reading), rather than being treated as a developmental outcome in its own right. Therefore, we retained the Digital Technology domain to reflect its increasing importance in early years education and emerging national policy and kept it positioned as underpinning learning across domains rather than as a standalone, easily measured area. The absence of explicit measures therefore reflects current measurement practice, not a lack of relevance, and retaining the domain helps future-proof the taxonomy as policy and practice continue to evolve.

The proportion of measures that were shortlisted was broadly consistent across domains, suggesting that screening criteria did not disproportionately filter out measures from particular domains.

Figure 7: Number of unique measures, by domain^a



^a The counts shown in the figure do not sum to the total number of assessments, as some assessments cover more than one domain.

Coverage by sub-domains and proficiencies in the shortlist

In the shortlist, at the sub-domain and proficiency level, we observed the highest coverage in areas traditionally prioritised in early assessment: Language (expressive and receptive); Reading (focusing on phonological awareness); Number (counting and recognising digits); Self-Regulation and Emotion Skills (including emotional⁹ and behavioural regulation); Executive Functioning (focusing on attention and inhibition); and Social Skills (with a focus on relationship skills and prosocial behaviour). Figure 8 shows coverage by sub-domain and a breakdown of the number of measures by proficiencies is given in Appendix E.

We found few or no measures for some proficiencies in the shortlist. In Mathematics, measures on place value, estimation, inverse operations, and symbol knowledge were absent or sparse, consistent with their developmental complexity for younger children. In Expressive Arts and Design, we did not identify any measures of early musicality and dance, and measures on creativity were sparse. Within Physical Development, we did not identify measures of strength, while balance and co-ordination dominated. Digital Technology proficiencies (technology awareness and digital literacy) had no measures. Similarly, we did not identify any measures for some Understanding the World proficiencies, such as chronological vocabulary. These gaps may partly reflect how assessments are described, as well as the limitation of mapping to the taxonomy based on publicly available information only, rather than a complete absence of measures. Nonetheless, this highlights areas where there is potential for assessment development.

Figure 8: Coverage, by sub-domain in the shortlist



⁹ Emotional regulation was a proficiency that was attributed to both Self-Regulation and Emotion Skills sub-domain.

Age coverage of measures

Figure 9 shows that measures cluster around ages three to six years, with highest coverage at ages four and five years in both the longlist and shortlist. We observed lower coverage for infants and toddlers (ages birth to two years) but not significantly. We expected this pattern as it reflects known challenges in assessing very early development, where rapid and uneven development means that proficiencies can change quickly, vary widely between children of the same age, and be expressed inconsistently across contexts, making reliable and standardised measurement more difficult. It also reflects challenges around assessing very young children, where limited attention span and emerging communication proficiencies make it difficult for children to engage reliably with standardised tasks, understand instructions, or produce consistent and interpretable responses. This necessitates the use of task-based assessments or reliance on observations or adult-reported measures. The concentration of measures from age three years onwards also reflects the expansion of formal assessment once children enter early education settings, where measurement becomes more pertinent and more routinely embedded in teaching. This is why we created an additional **guide on administering assessments with infants and toddlers (birth to three years) and very young children** (Mazhar *et al.*, 2026).

Even though this was beyond the scope of the EYMD, we also identified substantial numbers of measures targeting age six years and above, particularly in the longlist, as many tools cover a wide range of age groups and may not necessarily target early years only. The presence of a substantial number of measures for six-year-olds and older indicates that the EYMD captures tools relevant to the upper end of the early years age range, despite some academic searches being limited to ages from birth to five years. The overall age distribution remained similar after shortlisting.

Figure 9: Age distribution of measures



Distribution of measures by domain and age

Across domains, Communication and Language and PSED showed the broadest age coverage, with large numbers of measures spanning multiple age bands. Within academic domains,¹⁰ Literacy and Mathematics measures tended to concentrate more strongly in the three to six years age range, reflecting developmental expectations and emphasis that comes with inclusion in school curriculum. Physical Development showed a gradual reduction in older age groups, perhaps indicating the wider range of physical developmental milestones that takes place in earlier ages, whereas we found sparse

¹⁰ In this report, 'academic' domains refer to areas of learning that are embedded within the early years curriculum and form part of children's longer-term educational trajectories, including Literacy, Mathematics, Expressive Arts and Design, and Understanding the World. Academic measures are not intended to capture holistic development, as Communication and Language, PSED, and Physical Development are.

coverage across all ages for Expressive Arts and Design and Understanding the World. We included the Expressive Arts and Design and Understanding the World domains because they align closely with the EYFS and reflect important curricular areas in early years. However, they have historically received less emphasis in evaluation and assessment development than core curriculum domains such as Literacy and Mathematics, which may explain the more limited availability of standardised measures in these domains. Early Learning Environment measures showed relatively even distribution across ages, consistent with their focus on contextual features rather than child-specific developmental milestones.

Figure 10: Age distribution of measures, by domain in longlist

Age/domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Digital technology	Early Learning Environment
0	27	8	7	22	23	1	0	0	16
1	34	9	8	37	26	2	1	0	18
2	47	13	14	54	33	3	3	0	21
3	89	37	37	95	43	6	9	0	26
4	102	60	49	112	39	5	11	0	28
5	104	69	54	126	39	5	11	0	26
6	89	63	49	105	27	4	10	0	18
>6	72	50	36	89	19	1	7	0	12

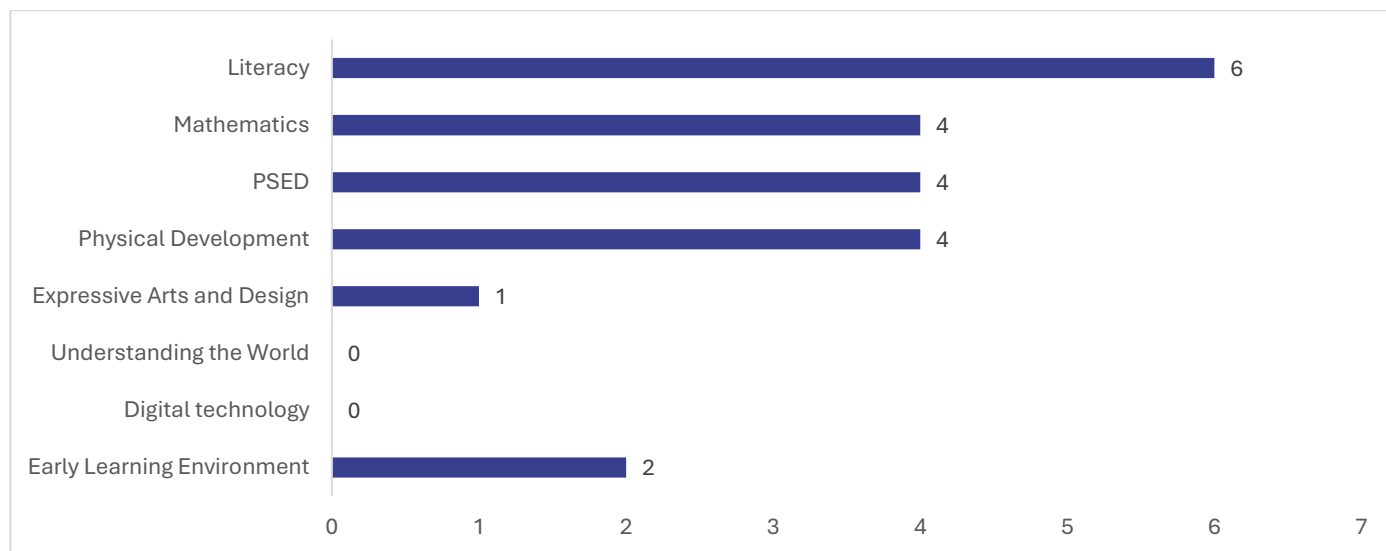
Figure 11: Age distribution of measures, by domain in shortlist

Age/domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Digital technology	Early Learning Environment
0	16	5	4	13	14	1	0	0	8
1	19	6	5	22	16	1	1	0	10
2	29	10	11	30	21	2	3	0	12
3	56	27	28	55	26	5	9	0	18
4	62	45	34	61	23	5	9	0	20
5	60	46	38	67	22	5	8	0	19
6	50	41	33	54	16	4	7	0	12
>6	37	30	21	41	8	1	4	0	8

UK-normed measures

Only a small number of measures in the shortlist had UK norms, and these were mainly in Communication and Language and Literacy. Figure 12 shows that we identified very few UK-normed measures in Mathematics, PSED, Physical Development, or Early Learning Environment domains, while only one measure had UK norms for Expressive Arts and Design and we did not identify any UK-normed measures for Understanding the World. This underscores the need to rely on internationally developed tools in several domains and supports the decision to relax the UK-norms criteria for shortlisting. To support the use of these measures, we also created an additional **guide on adapting international measures** for use in the UK context (Valenzuela *et al.*, 2026).

Figure 12: Number of measures with UK norms, by domain



Quality of measures

We appraised the psychometric properties of shortlisted measures using the framework described in the section ‘Updating the EYMD content’ above. We report the findings from this appraisal below for each key indicator of quality to highlight overall patterns in the strength and consistency of evidence across domains.

We note that the psychometric ratings presented in the EYMD are intended to support informed decision-making rather than to serve as definitive judgements on the quality of individual measures. Users should interpret ratings in relation to the intended use of a measure, the population for which it was designed, and the context in which it is applied. A measure rated lower on certain psychometric criteria may still be appropriate for formative use, exploratory evaluation, or use with specific subgroups, while measures with stronger evidence may not always be appropriate for all evaluation contexts. We explain this in more detail in the user **guide on how to use the EYMD** (Malik *et al.*, 2026).

Validity

We found that evidence of construct validity was generally strong across domains, with around 90% of the measures across all domains except Expressive Arts and Design reporting construct validity (Figure 13). This suggests that most tools included in the shortlist are grounded in explicit theoretical constructs, even where other psychometric properties may be weaker or inconsistently reported.

Figure 13 shows that the quality of the measures identified based on construct validity was high, as the largest proportion of measures across all domains given the highest rating. Construct validity of PSED measures showed slightly more variability, though we still rated most measures.

Figure 13: Construct validity ratings distribution, by domain

Rating/ domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Early Learning Environments
Not reported	12%	11%	10%	11%	9%	17%	10%	0%
✓	3%	7%	3%	0%	5%	0%	0%	0%
✓✓	24%	20%	23%	33%	23%	17%	30%	33%
✓✓✓	62%	61%	63%	56%	64%	67%	60%	67%

We found that criterion validity was more unevenly reported than construct validity. While a majority of measures in Communication and Language, Literacy, Mathematics, PSED, and Physical Development showed strong criterion validity, a substantial proportion of measures, particularly in Expressive Arts and Design and Early Learning Environment, either lacked criterion validity evidence or relied on weaker forms of validation (Figure 14). This reflects both the difficulty of identifying appropriate external criteria in some domains and the evaluative rather than diagnostic intent of many tools.

Figure 141: Criterion validity rating distribution, by domain

Rating/ domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Early Learning Environments
Not reported	26%	16%	27%	17%	27%	50%	10%	17%
✓	6%	7%	7%	8%	5%	17%	10%	22%
✓✓	12%	9%	13%	22%	5%	17%	20%	28%
✓✓✓	56%	68%	53%	53%	64%	17%	60%	33%

Reliability

We found that internal consistency was the strongest and most consistently reported reliability property across domains. Figure 15 shows that the Early Learning Environment and Physical Development domains had higher proportions of measures for which internal consistency was not reported. This is potentially due to the multidimensional and observational nature of these assessments, where internal consistency is less routinely calculated or reported. However, we did not find reported internal consistency for a quarter of the Communication and Language measures.

Figure 15 also shows that, where reported, most measures had high internal consistency ratings across all domains, particularly in Mathematics, Expressive Arts and Design, and Understanding the World, although the latter two domains were represented by fewer measures overall and these small numbers limit conclusions we can draw. Early Learning Environment measures show a lower proportion of strong internal consistency compared to other domains. As stated earlier, this likely reflects the heterogeneous nature of these tools, which capture diverse contextual features rather than a single underlying construct, making high internal consistency less appropriate. Nonetheless, the evidence indicates that overall, items within shortlisted measures across all domains generally function coherently to capture the intended constructs.

Figure 15: Internal consistency ratings distribution, by domain

Rating/ domain	Communication and Language		Literacy	Mathematics		PSED	Physical Development		Expressive Arts and Design	Understanding the World	Early Learning Environment
Not reported	24%	11%	3%	11%	27%	17%	0%	33%			
✓	0%	7%	3%	6%	5%	0%	0%	0%			
✓✓	9%	9%	7%	14%	5%	0%	10%	22%			
✓✓✓	68%	73%	87%	69%	64%	83%	90%	44%			

Test-retest and inter-rater reliability were substantially less well reported.

We did not find test-retest reliability for a large proportion of measures across all domains, especially Expressive Arts and Design and Understanding the World. For other domains, between a third and a half of the measures did not report test-retest reliability. Where reported, we found that the evidence was often strong. However, for PSED and Early Learning Environment, test-retest reliability shows a more mixed pattern compared to some other domains. While a significant proportion of these measures demonstrate strong test-retest reliability, this is lower than in other domains and is accompanied by a notable proportion of measures with only moderate or limited evidence. Given that many of these measures rely on observation, behaviour in context, or adult judgement, this may reflect the influence of situational factors, such as child mood, social dynamics, or assessor interpretation, which can fluctuate over short periods (Figure 16).

Figure 16: Test-retest reliability ratings distribution, by domain

Rating/ domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Early Learning Environment
Not reported	38%	48%	53%	39%	36%	83%	60%	56%
✓	6%	2%	0%	8%	9%	0%	0%	11%
✓✓	9%	16%	13%	14%	5%	0%	10%	22%
✓✓✓	47%	34%	33%	39%	50%	17%	30%	11%

Inter-rater reliability showed a similar pattern. Figure 17 shows that most measures across Literacy, Mathematics, Expressive Arts and Design, Understanding the World and Early Learning Environment did not report inter-rater reliability. This pattern reflects the fact that many tools in academic domains like Literacy, Mathematics, and Understanding the World are designed to be standardised, objectively scored assessments, where rater judgement plays little or no role and inter-rater reliability is either not applicable or not reported, as consistency is ensured through fixed scoring rules rather than agreement between assessors. However, the lack of Early Learning Environment measures reporting inter-rater reliability is a weakness since most measures in this domain are observational and whether different observers note and score the same thing is critical. A further weakness is that where reported, measures of Early Learning Environment do not always have the highest rating, indicating a lack of good quality measures in this domain. For measures in other domains, where inter-rater reliability was reported, we often found strong evidence.

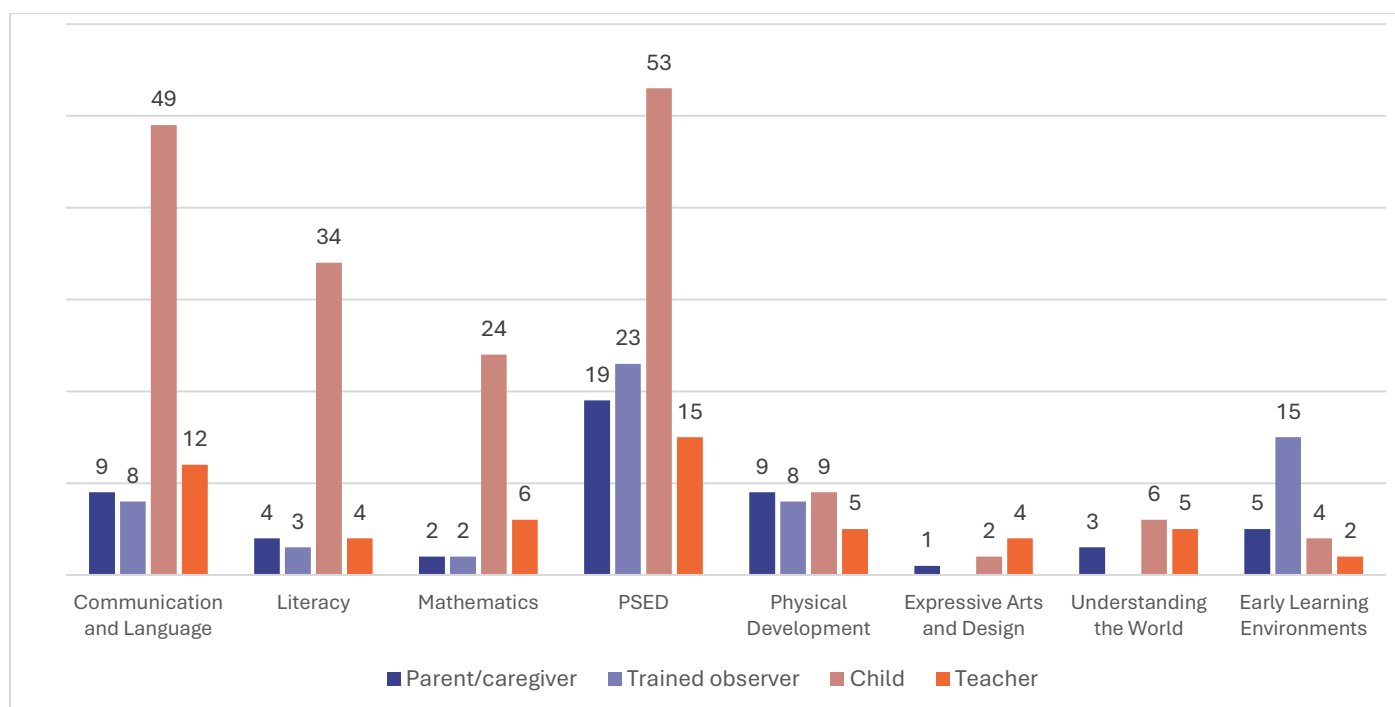
Figure 17: Inter-rater reliability ratings distribution, by domain

Rating/ domain	Communication and Language	Literacy	Mathematics	PSED	Physical Development	Expressive Arts and Design	Understanding the World	Early Learning Environments
Not reported	32%	66%	73%	36%	27%	50%	70%	56%
✓	12%	2%	0%	19%	9%	0%	0%	11%
✓✓	6%	7%	3%	6%	9%	0%	0%	6%
✓✓✓	50%	25%	23%	39%	55%	50%	30%	28%

Distribution of measures by respondent

Figure 18 shows variation in respondent type across domains. Child-completed measures dominate most domains except Early Learning Environment. Child-completed measures are particularly prominent in PSED, Communication and Language, Literacy, and Mathematics, indicating a strong reliance on direct child engagement in these domains. Practitioner involvement is also evident. Teacher-completed measures are most frequent in PSED and Communication and Language, with more limited representation in other domains. Parent/caregiver-completed measures are comparatively fewer overall, though they appear most frequently in PSED and Communication and Language, with more limited representation in other domains. Early Learning Environment domain stands out for its heavier reliance on trained observers rather than direct child participation. The dominance of child-completed measures in academic and socio-emotional domains likely reflects stronger coverage for older children within the early years age span, who are better able to engage in structured tasks. The distribution may also be a reflection of the thinner coverage for younger children, for whom parent-report or observational approaches, which are less represented in the database, are often more appropriate. The concentration of observer-based tools in Early Learning Environment reflects established traditions in environmental quality assessment.

Figure 18: Distribution of measures, by respondent



Findings

This section presents domain-level findings from the review, synthesising evidence on the availability, age coverage, and psychometric quality of early years assessment measures. While the preceding analysis examined cross-cutting patterns across the database as a whole, this section disaggregates the findings by domain to explore differences in coverage, proficiencies assessed, and measurement quality in more detail. To support interpretation for the primary users of the database based in the UK, the findings are presented broadly in line with the structure of the EYFS, beginning with the three prime areas of learning and development (Communication and Language, PSED, and Physical Development), followed by the specific areas that build on them. The prime areas represent universal foundations of early development, reflecting biologically time-sensitive processes that are particularly rapid during the first years of life and underpin children's engagement with the world through relationships, communication, movement, and self-regulation. Our findings are consistent with this developmental logic as the measures identified in the EYMD show particularly strong coverage of early proficiencies within these domains and earlier age ranges, reflecting their role in supporting the development of later specific areas (Literacy, Mathematics, Understanding the World, and Expressive Arts and Design). Together, these domain-level summaries highlight areas of relative strength and constraint within the current measurement landscape and provide context for the implications discussed later in the report.

Communication and Language

Communication and Language is one of the most robustly covered domains, but with stronger coverage for younger children, including those aged from birth to two years, reflecting the stronger emphasis on this domain during infancy and toddlerhood and the fact that early Communication and Language proficiencies provide the foundations for later proficiencies in other domains, particularly Literacy. Measures identified for this domain span a wide range of proficiencies, particularly expressive and receptive vocabulary, phonology, syntax, and pragmatic language (see Appendix E). We found strong psychometric evidence overall, especially for construct validity and internal consistency, although reliability evidence beyond internal consistency is less consistently reported. This reflects the predominance of structured, task-based assessments with fixed scoring rules.

PSED

The PSED measures in the database provide broad coverage across ages and proficiencies, particularly for emotional regulation, behavioural regulation, executive functioning, and social skills (see Appendix E). However psychometric evidence is variable. While construct validity is generally strong, criterion validity and reliability evidence are less consistently reported, potentially reflecting the conceptual breadth of the domain and the diversity of measurement approaches used (e.g. questionnaires and observations). Test developers report test-retest and inter-rater reliability more frequently but still unevenly. There is also a greater balance between direct child assessments and practitioner-reported measures for PSED compared to other domains. While direct child assessments provide concrete evidence of performance, they only capture behaviour within a specific assessment context and time point. Practitioner and observer measures may provide greater ecological validity in these instances but are prone to subjectivity and depend heavily on rater training and reliability. Because of these complexities, PSED measurement in particular, may be an area where a single assessment approach is insufficient on its own and may point to the potential value of triangulating across respondent types to strengthen validity.

Physical Development

The availability of measures for Physical Development shows a distinct pattern, with stronger coverage at younger ages and a gradual decline for older children. Measures focus primarily on co-ordination, locomotion, and fine motor skills, with lack of measures for strength (see Appendix E), though strength may be implicit in some other assessed proficiencies (e.g. strength is needed for throwing, for locomotion). Where reported, psychometric evidence reveals that measures are generally of a good quality, particularly in terms of test-retest and inter-rater reliability, which are more relevant given the observational nature of many tools. Internal consistency is less frequently reported, which may be due to the need to assess discrete motor skills rather than single latent constructs.

Literacy and Mathematics

Literacy and Mathematics are the most extensively represented academic domains in the EYMD. Both domains are well represented in terms of number of measures, particularly for children aged three to six years, reflecting their clearer operationalisation as academic outcomes in specific areas that build on the foundational proficiencies developed in the prime areas and the strong emphasis and tradition of measurement in these areas. Coverage is strongest for foundational proficiencies, such as phonological awareness, early reading, counting, and number recognition, while more complex or abstract proficiencies, such as writing composition, place value, estimation, and inverse operations, are less well covered (see Appendix E). Psychometric quality in both domains is generally high. Construct validity and internal consistency are consistently strong, and a majority of measures report criterion validity. UK-normed measures are more common here than in most other domains, though still limited overall. Test developers rarely report test-retest and inter-rater reliability, largely because measures are typically standardised and objectively scored, making rater agreement less relevant.

Expressive Arts and Design and Understanding the World

We found fewest measures for Expressive Arts and Design and Understanding the World. This is expected because these specific areas tend to receive less emphasis than Literacy and Mathematics as academic domains and are more often assessed through informal, observational approaches without technically robust assessment tools. This is particularly true for the early years age group, where children's development in these areas is typically observed during open-ended activities such as play, creative expression, or exploration of the environment, rather than through short, standardised tasks with clearly defined responses, making it hard to operationalise these domains for assessment purposes. Both domains have relatively few measures across all ages and uneven coverage at the proficiency level. We did not identify or found few measures of key proficiencies, such as early musicality, dance, creativity, and chronological vocabulary (see Appendix E). Where measures exist, construct validity and internal consistency are often rated highly, but small numbers limited the conclusions we can draw. Criterion validity and reliability evidence are sparse, reflecting both the difficulty of defining external benchmarks and the limited number of assessments in these domains.

Early Learning Environment

Early Learning Environment measures form a distinct group, focusing on contextual rather than child-level constructs. Coverage is relatively even across all age groups for this reason. Overall coverage is moderate, but psychometric evidence is more variable than in child-focused domains. Internal consistency is less frequently reported and less often strong, reflecting the multidimensional nature of these tools. Criterion validity and inter-rater reliability are also inconsistently reported, which is a notable weakness given the observational nature of many measures and their use in evaluation contexts. In addition, a relatively high proportion of these measures are normed or primarily validated outside the UK, which may limit their contextual relevance and comparability for UK-based settings without further adaptation or local validation evidence. We outline how this can be done in our user **guide on adapting international measures** for use in the UK context (Valenzuela *et al.*, 2026).

Digital Technology

We did not identify any standalone measures for Digital Technology. This represents a clear gap between conceptual recognition of digital proficiencies in early development and the availability of tools to measure these proficiencies directly.

Implications

Use of the EYMD

The updated EYMD provides a robust and transparent resource for identifying and selecting outcome measures for early years research, evaluation, and practice. It makes the current measurement landscape visible by highlighting where there is a strong supply of well-evidenced tools and where choices are more limited.

In domains where coverage is strongest, such as Communication and Language, Literacy, Mathematics, and PSED, users can draw on a relatively wide range of well-established measures with strong construct validity and internal consistency, especially for children aged three to six years. In these domains, the EYMD can support evaluators and commissioners to make confident decisions around selecting measures that align with intervention aims and target age groups, while being aware of the practical constraints. However, psychometric evidence should be interpreted with care in relation to the underlying assessment design. For example, indicators such as inter-rater reliability may be less relevant or not reported for standardised, objectively scored Literacy and Mathematics assessments, and should not be taken as a marker of weaker quality in such cases.

In domains where coverage and quality are more uneven, including Physical Development, Expressive Arts and Design, Understanding the World, and Early Learning Environment, the database helps users navigate trade-offs between conceptual relevance, age appropriateness, and psychometric strength. The findings also underline the need for caution in interpreting psychometric ratings in domains characterised by observational, multidimensional, or context-level constructs, where some indicators (such as internal consistency) may be less applicable or less consistently reported. Many proficiencies relevant to these domains may also be captured within measures from other areas. For example, pattern recognition skills under Mathematics support children's understanding of sequence, comparison, and natural processes within Understanding the World, and emotional and social skills within PSED underpin proficiencies such as sensitivity to difference, social observation, and engagement with people, culture, and communities. Similarly, many proficiencies relevant to Expressive Arts and Design are also captured within measures from other domains, such as Communication and Language proficiencies underpinning narrative and role play, physical proficiencies supporting dance, and mathematical proficiencies such as spatial reasoning supporting visual composition. We address this in more detail in the **user guide on how to use the EYMD** (Malik *et al.*, 2026).

Future tool development

The results point to several clear priorities for future tool development.

- **We identified substantial gaps in Expressive Arts and Design, Understanding the World, and Digital Technology domains.** The first two domains were deliberately included because of their prominence within the EYFS. Despite their curricular importance, these areas have historically received less attention in assessment development. So, the limited numbers of measures, uneven proficiency coverage, and sparse psychometric evidence constrain evaluation. In addition, the absence of any identified measures of Digital Technology proficiencies for this age group highlights a clear gap between conceptual recognition of digital proficiencies and their operationalisation in early years assessment.
- Across domains, **we identified that several developmentally important proficiencies remain poorly captured**, including creativity, early musicality, chronological understanding, and more advanced early mathematics concepts such as estimation and place value. These gaps reflect three potential challenges: i) some proficiencies (such as creativity and early musicality) have not historically been the focus of formal assessment development in the early years; ii) some proficiencies (such as estimation and place value) are more commonly assessed at later ages and are therefore less well served by age-appropriate early years measures; and iii) some proficiencies (such as chronological understanding) may be reflected indirectly within broader skills (e.g. narrative sequencing within Communication and Language), but they are rarely assessed explicitly as distinct constructs. Addressing these gaps will require both increased demand and investment in measures targeting historically under-measured proficiencies, and the adaptation of existing

tools to capture more advanced concepts at younger ages, particularly to mitigate the risk of ceiling effects when such proficiencies are omitted.

- **Future tool development could prioritise multi-respondent approaches where feasible to strengthen validity across contexts.** The distribution of respondent types across domains reflects established measurement traditions. PSED measures are more likely to be adult-reported, Early Learning Environment measures rely heavily on structured observation, and Communication and Language, Literacy, and Mathematics measures are predominantly task-based and child-completed. While these patterns are often appropriate, greater triangulation across respondent perspectives would strengthen confidence in inferences drawn. In evaluation contexts, this may involve using multi-respondent approaches by selecting a more objective or independent measure as a primary outcome while using additional respondent perspectives as secondary outcomes to provide complementary evidence. However, decisions about whether to use multi-respondent approaches will also involve trade-offs between measurement robustness and the cost and feasibility of data collection in evaluation contexts.
- **We identified a need for more age-appropriate measures for children aged birth to two years.** While some tools exist for this age group, these are mainly for Communication and Language and Physical Development. This concentration of tools in a small number of domains may shape which outcomes are evaluated in infancy and toddlerhood, limiting the ability to capture change across a broader range of developmental areas at this age. For example, there are very few measures capturing pre-reading skills like print awareness for birth to two-year-olds. Similarly, there are few measures for early numeracy skills such as pattern awareness and shape recognition. Expanding the range of child-friendly and robust assessments for very young children would support earlier intervention and evaluation.
- **The limited availability of UK-normed measures beyond core academic and language domains can constrain evaluation.** While international measures remain available, further UK norming and validation would strengthen interpretability and comparability. This is particularly true for PSED measures, especially social skills, as these skills are highly context dependent. Greater UK norming and validation would improve cultural relevance. We outline how this can be done in **the user guide on adapting international measures** for use in the UK context (Valenzuela *et al.*, 2026).
- Finally, **although Early Learning Environment measures are available, their psychometric quality is more variable than for child-focused domains.** Future tool development should prioritise improving reliability, particularly inter-rater reliability, and clarifying construct definitions, given their importance.

Overall, the EYMD supports both better use of existing measures and more strategic decision-making about where future investment in assessment development and validation is most needed, helping to align measurement practice more closely with the full breadth of early learning priorities.

Implications for the EEF trials

For the EEF, these findings have implications for both evaluation design and commissioning.

- **Evaluators should justify outcome choices by clearly linking them to what an intervention is designed to change, rather than defaulting to outcomes with the strongest available measures, particularly outside Literacy and Mathematics.** The findings show that strong, standardised psychometric evidence is concentrated in Literacy and Mathematics, while other domains rely more heavily on observational, multidimensional, or context-level measures with unevenly reported psychometric properties. This implies that outcome selection should be led by the Theory of Change rather than tool-led, with psychometric quality standards applied proportionately across domains, to avoid systematically disadvantaging outcomes that are central to intervention intent but less robust in their psychometric properties.
- **In domains where standardised tools are limited or absent, evaluations may need to develop bespoke, evaluation-specific outcome measures** tailored to the intervention and context, where appropriate. This

is particularly relevant for domains such as Expressive Arts and Design, Understanding the World, Physical Development, and Early Learning Environment, where no single standardised measure adequately captures intended outcomes and where carefully designed, fit-for-purpose tools may provide more meaningful evidence of change. In such cases, measures may first be piloted on a small scale (e.g. at the pilot stage of the EEF's pipeline) to test feasibility and establish initial evidence of validity and reliability, with further evidence gathered during trials using a larger and more representative sample. Decision-making about the development process will involve trade-offs between measurement robustness and the cost and practicality of data collection, which the EEF and evaluators will need to consider carefully together.

- **The EEF may wish to play a market-shaping role**, signalling demand for new or adapted measures through Terms of References, funded pilots, or partnerships with measure developers. Where bespoke, evaluation-specific measures would require specialist design or validation expertise beyond that typically held by evaluation teams, the EEF may need to commission or support measure development separately from the evaluation itself.
- **The EEF may wish to require robust assessor training for use of observational measures within trials and potentially pilot measures before trial.** In the Early Learning Environment domain, inter-rater reliability is not reported for approximately half of the identified measures and, where reported, is often variable. Given the observational nature of these tools, the EEF might want to emphasise the need for proper training and quality checks when using these measures. It may be useful to pilot the measures at the pilot stage prior to full trial implementation, so evaluators can assess feasibility and reliability in advance, reducing risk at the main trial stage.
- **The EEF may wish to consider mechanisms for making technical information more visible through its reports.** Validation evidence and technical documentation are not always easily accessible, particularly for paywalled tools. Trial reports could make this information visible by including information on reliability (including internal consistency, test-retest or inter-rater reliability), validity (including construct or criterion validity), and any available information on norms or standardisation, if applicable, of measures when justifying the outcome of choice and potentially compare this with other measures considered. Alternatively, trials may include summary psychometric properties of measures when used during trial in the appendices to the report. In addition to the outcome distribution and ceiling and floor effects that are already reported in the EEF trial reports, this could include:
 - internal consistency (e.g. Cronbach's alpha);
 - inter-rater reliability (if applicable);
 - test-retest reliability (if repeated measures are collected);
 - if feasible, responsiveness or sensitivity to change over the course of the trial using standardised effect sizes (e.g. Cohen's d or standardised response mean); and
 - where relevant, analyses could also explore subgroup consistency (e.g. by key demographic characteristics) to assess whether measures perform similarly across groups.

This may have additional cost implications, but systematic reporting of technical characteristics in trial documentation would support cumulative learning and reduce duplication of appraisal effort for future evaluators.

- Where feasible, **trials should consider triangulating across respondent types to strengthen confidence in findings.** The distribution of respondent types across domains suggests that outcome interpretation may vary. This pattern is also reflected in the reliability evidence presented above, where measures relying on observation or adult report, particularly in domains such as PSED and Early Learning Environment, show greater variability in test-retest reliability compared to other domains. Adult-reported PSED measures can capture behaviour over time and across contexts but may be biased. Similarly, observational Early Learning Environment tools can provide rich contextual insight but depend heavily on rater training and inter-rater reliability. Task-based Literacy and Mathematics assessments offer structured, standardised evidence of performance but typically reflect behaviour within a specific assessment context and time point. This is

particularly important in early years research, where children's skills are often still emerging, rapidly developing and highly context dependent. Between ages from birth to six years, small changes in language exposure, adult interaction, or classroom routines can meaningfully influence observed behaviour, and performance may fluctuate substantially across settings and time points. Unlike older pupils, young children cannot reliably self-report many socio-emotional constructs, and even task-based assessments may capture only a snapshot of performance influenced by attention, familiarity with the assessor, or regulation in the moment. As a result, reliance on a single respondent type or assessment mode may provide an incomplete picture of development and impact. Combining perspectives in early years may therefore provide a more balanced and robust picture of impact.

Project team

The project was delivered by a team from Oxford MeasurEd (see Table 5).

Table 5: Project team

Name	Role	Institutional affiliation
Dr Lydia Marshall	Principal investigator and project director	Oxford MeasurEd (director of research)
Dr Rachel Outhred	Expert advisor	Oxford MeasurEd (managing director)
Dr Abbie Raikes	Expert advisor	ECDMeasure (founder and chief scientific officer)
Robert Wishart	Expert advisor	Oxford MeasurEd (associate)
Sara Bashir Malik	Project manager and UX lead	Oxford MeasurEd (senior consultant)
Paulina Valenzuela	Analyst – assessment specialist	Oxford MeasurEd (senior consultant)
Dr Anushay Mazhar	Analyst – early years and social-emotional learning specialist	Oxford MeasurEd (consultant)
Dr Astrid Pickenpack	Analyst	Oxford MeasurEd (consultant)
Tasvid Khan	Analyst	Oxford MeasurEd (associate)

References

- Britto, P.R., Lye, S.J., Proulx, K., Yousafzai, A.K., Matthews, S.G., Vaivada, T., Perez-Escamilla, R., Rao, N., Ip, P., Fernald, L.C.H., MacMillan, H., Hanson, M., Wachs, T.D., Yao, H., Yoshikawa, H., Cerezo, A., Leckman, J.F., Bhutta, Z.A.; Early Childhood Development Interventions Review Group, for the Lancet Early Childhood Development Series Steering Committee. (2017) 'Nurturing Care: Promoting Early Childhood Development'. *The Lancet*, 389: 10064, 91–102. [https://doi.org/10.1016/S0140-6736\(16\)31390-3](https://doi.org/10.1016/S0140-6736(16)31390-3)
- Department for Education (DfE). (2021) '*Changes to the Early Years Foundation Stage (EYFS) Framework*'. GOV.UK. Available at: www.gov.uk/government/publications/changes-to-the-early-years-foundation-stage-eyfs-framework/changes-to-the-early-years-foundation-stage-eyfs-framework (accessed 20 04 2026).
- Department for Education (DfE). (2025a) '*Changes to the EYFS Framework*'. Department for Education. Available at: <https://help-for-early-years-providers.education.gov.uk/support-for-practitioners/changes-to-the-eyfs-framework> (accessed 20 04 2026).
- Department for Education (DfE). (2025b) '*Areas of Research Interest*'. London: Department for Education. Available at: https://assets.publishing.service.gov.uk/media/68136b1bb0ef2c985052541c/DfE_areas_of_research_interest.pdf (accessed 20 04 2026).
- Dockrell, J., Hurry, J., Cowan, R., Flouri, E. and Dawson, A. (2017) '*Review of Assessment Measures in the Early Years: Language and Literacy, Numeracy and Social Emotional Development and Mental Health*'. London: Education Endowment Foundation. Available at: https://discovery.ucl.ac.uk/id/eprint/10039319/8/Dockrell_Report.pdf (accessed 20 04 2026).
- Early Years Alliance. (2026) '*Government to Publish New Guidance on Screen Time for Parents and Carers of Under-Fives*'. Early Years Alliance. Available at: www.eyalliance.org.uk/news-events/news/government-to-publish-new-guidance-on-screen-time-for-parents-and-carers-of-under-fives/ (accessed 20 04 2026).
- European Federation of Psychologists' Associations (EFPA). (2013) '*EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests. Test Review Form and Notes for Reviewers. Version 4.2.6.b*'. Available at https://mlp.fi/wp-content/uploads/2023/07/DiSC-EFPA_TestReviewModel_Report.pdf (accessed 18 08 2026).
- Malik, S.B., Marshall, L., Mazhar, A., Pickenpack, A., Outhred, R. and Wishart, R. (2025) '*Update of the EEF's Early Years Measures Database: Protocol*'. London: Education Endowment Foundation. Available at: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/early-years-measures-database-update> (accessed 20 04 2026).
- Malik, S.B., Marshall, L., Mazhar, A., Pickenpack, A. and Wishart, R. (2026) '*How to Use the EEF's Early Years Measures Database: User Guide*'. London: Education Endowment Foundation. Available at: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/early-years-measures-database> (accessed 08 10 2026).
- Mazhar, A., Outhred, R., Malik, S.B. and Marshall, L. (2026) '*Conducting Assessments with Infants and Toddlers (Birth to Three Years): User Guide*'. London: Education Endowment Foundation. Available at: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/early-years-measures-database-update> (accessed 08 10 2026).
- Tierney, A.L. and Nelson, C.A. 3rd. (2009) 'Brain Development and the Role of Experience in the Early Years'. *Zero to Three*, 30: 2, 9–13.
- Valenzuela, P., Outhred, R., Malik, S.B. and Marshall, L. (2026) '*Adapting International Measures for Use in the UK: User Guide*'. London: Education Endowment Foundation. Available at: <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/early-years-measures-database-update> (accessed 08 10 2026).

World Bank. (2017) 'Early Childhood Development Measurement Inventory'. In: Fernald, L., Prado, E., Kariger, P. and Raikes, A. *'A Toolkit for Measuring Early Childhood Development in Low- and Middle-Income Countries'*. Washington, DC: World Bank. Available at: <https://www.worldbank.org/en/programs/sief-trust-fund/publication/a-toolkit-for-measuring-early-child-development-in-low-and-middle-income-countries> (accessed 20 04 2026).

Appendix A: Taxonomy

Appendix A Table 1: Domain, sub-domain, and proficiency definitions

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
Communication and Language	Children's proficiency in understanding and conveying information, ideas, and feelings, including through words and sentences	Communication	Children's proficiency in conveying and understanding information, ideas, or feelings	Joint attention	Proficiency in sharing focus with one or more individuals on the same object or event, co-ordinated through eye gaze, pointing, or other verbal/non-verbal cues
				Gesticulation	Proficiency in using gestures, such as hand movements or facial expressions, to convey meaning alongside or instead of speech
				Turn taking	Proficiency in engaging in reciprocal communication with another person
				Speech	Proficiency in producing spoken language sounds clearly and fluently to convey meaning
				Pragmatic skills	Proficiency in adapting verbal and non-verbal communication to different social contexts
				Code switching	Proficiency in alternating between two or more languages or language varieties depending on context and audience
		Language	Children's proficiency in understanding words and sentence structures and using them to express meaning	Expressive vocabulary	The range of words a child can produce and use in speech, writing, or sign language
				Receptive vocabulary	The range of words a child can understand when heard or read
				Syntax	Proficiency in arranging words to make meaning
				Morphology	Proficiency in using word parts, such as prefixes and suffixes, to create meaning
				Grammar	Proficiency in using the system of rules governing the structure of words, phrases, and sentences, to create meaning
				Narrative skills	Proficiency in structuring and telling a story or recounting events coherently
				Inferencing	Proficiency in understanding information that is implied but not directly stated, using context clues and prior knowledge
				Language reasoning	Proficiency in using language to explain their thinking and reasoning

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
Literacy	Children's proficiency in engaging with, understanding and conveying meaning through written text	Reading	Children's proficiency in decoding and making meaning from written text	Print awareness	Understanding that print carries meaning and knowing how to handle books and written materials
				Letter naming	Proficiency in recognising and naming the letters of the alphabet
				Phonological awareness	Proficiency in recognising and manipulating the sound structures of language (e.g. syllables, phonemes, rhyme, and alliteration)
				Letter-sound correspondence	Proficiency in matching written letters/graphemes and the sounds they represent
				Word reading	Proficiency in reading familiar and unfamiliar words
				Rapid automatised naming	Proficiency in quickly naming familiar items, such as letters, words, numbers, or colours, when presented visually
				Code switching	Proficiency in alternating between two or more languages or language varieties depending on context and audience
				Reading comprehension	Proficiency in understanding and interpreting the meaning of written text
				Reading fluency	Proficiency in reading text accurately, quickly, and with appropriate expression
		Writing	Children's proficiency in generating and recording ideas using symbols to communicate meaning	Graphomotor skills	Proficiency in co-ordinating fine motor movements for writing and drawing
				Mark making	Proficiency in creating marks on a surface as a precursor to writing
				Spelling	Proficiency in writing words with correct letter sequences to represent spoken language
				Handwriting	Proficiency in forming letters and words legibly and consistently
				Written composition	Proficiency in organising ideas to create meaningful written text
Mathematics	Children's proficiency in engaging with and understanding ideas about number, relationships, quantity, and space	Number	Children's proficiency in understanding and expressing ideas about quantity, counting, and numerical patterns	Pattern awareness / matching	Proficiency in recognising and reproducing patterns using shapes, colours, numbers, objects, or events
				Recognising and naming digits	Proficiency in identifying and naming written numbers
				Linking number words, digits, and quantities	Proficiency in identifying the relationship between spoken number words, written digits, and their quantities
				Counting	Proficiency in skills including correctly saying number words in a sequence, assigning each number word to a single object (one to one correspondence), and recognising that the final number word used represents the total number of objects in the set (cardinality)
				Subitising	Proficiency in visually seeing a number of objects instantly without needing to count them out one at a time
				Number composition	Understanding that a number can be made up of two or more smaller numbers
				Place value	Understanding the value of a digit based on its position within a number

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
				Comparing numerical magnitude	Proficiency in identifying which numbers are worth more or less than each other
				Numerical patterns	Proficiency in recognising and reproducing patterns using numbers, and what this means about their relationships
				Mathematical talk	Proficiency in having spoken or written conversation with teachers, children, and self in order to problem solve or aid understanding of mathematical concepts
		Operations	Children's proficiency in grouping, sharing, and combining numbers	Additive reasoning	Proficiency in solving problems involving combining, separating, or comparing quantities using addition and subtraction
				Calculation	Proficiency in carrying out mathematical operations such as addition, subtraction, multiplication, or division
				Inverse operations	Understanding that some mathematical operations can reverse the effect of others
				Symbol knowledge	Proficiency in recognising and understanding mathematical symbols and their meanings
				Strategy selection	Proficiency in choosing appropriate methods to solve mathematical problems
				Comparing numerical magnitude	Proficiency in identifying which numbers are worth more or less than each other
				Numerical patterns	Proficiency in recognising and reproducing patterns using numbers, and what this means about their relationships
				Mathematical talk	Proficiency in having spoken or written conversation with teachers, children and self in order to problem solve or aid understanding of mathematical concepts
		Measurement	Children's proficiency in understanding, describing, and comparing physical attributes including size, weight, and capacity	Non-standard units	Proficiency in measuring objects or quantities with everyday items rather than formal units, such as using blocks, footsteps, or hand spans
				Estimation	Proficiency in making a reasonable guess about quantity, size, or outcome before measuring or counting
				Comparing physical magnitude	Proficiency in comparing and ordering objects and shapes by size, weight and capacity
				Comparing quantities	Proficiency in identifying which quantities are worth more or less than each other
				Pattern awareness / matching	Proficiency in recognising and reproducing patterns using shapes, colours, numbers, objects or events
				Mathematical talk	Proficiency in having spoken or written conversation with teachers, children and self in order to problem solve or aid understanding of mathematical concepts
				Shape recognition	Proficiency in identifying and naming common shapes

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
		Geometry and Spatial Thinking	Children's proficiency in understanding, describing and comparing shapes, space and spatial relationships	Shape decomposition	Understanding that complex shapes can be broken down into smaller, simpler shapes, and ability to do so
				Two dimensional (2D)–three dimensional (3D) relationships	Understanding how 2D representations relate to 3D objects
				Spatial language	Proficiency in using words to describe the position, direction, and movement of objects in space
				Perspective taking	Proficiency in understanding and representing how objects or scenes appear from different viewpoints
				Comparing physical magnitude	Proficiency in comparing and ordering objects and shapes by size, weight, and capacity
				Pattern awareness / matching	Proficiency in recognising and reproducing patterns using shapes, colours, numbers, objects, or events
				Mathematical talk	Proficiency in having spoken or written conversation with teachers, children, and self in order to problem solve or aid understanding of mathematical concepts
PSED	Children's proficiency in forming relationships, understanding themselves, and making choices about their emotions, behaviour, and learning	Personal Development	Children's understanding of themselves and capacity to take actions to look after themselves	Self-esteem	Confidence in their own worth or abilities
				Self-concept/ sense of self	Beliefs and knowledge about their own attributes and qualities
				Self-care and healthy behaviours	Capacity to engage in activities and habits that maintain physical and mental health, such as rest, hygiene, and healthy eating
				Safety behaviours	Capacity to engage in activities and habits to avoid harm and ensure personal safety
		Social Skills	Children's proficiency in adapting to and interacting effectively with others in group and one-on-one contexts	Theory of mind / mentalisation	Understanding that others have thoughts, feelings, and perspectives different to their own
				Empathy	Proficiency in understanding and sharing others' feelings
				Relationship skills	Proficiency in interacting effectively with others, building positive relationships, and managing social challenges
				Prosocial behaviour	Capacity to engage in actions intended to benefit others, such as helping, supporting, or sharing
				Conflict resolution skills	Proficiency in resolving disagreements in respectful and constructive ways
		Emotion Skills	Children's proficiency in recognising, understanding, and expressing emotions in themselves and others	Emotion recognition	Proficiency in identifying their own emotions through feelings and through verbal and non-verbal communication
				Emotion expression	Proficiency in communicating feelings verbally and non-verbally
				Emotion vocabulary	The range of words a child can use to describe and express feelings and emotions
				Emotion understanding	Proficiency in understanding a range of distinct emotions that they can experience

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
		Self-Regulation	Children's proficiency in adapting their thoughts, behaviours, and emotions to achieve their goals	Emotional regulation	Proficiency in planning, monitoring, and adjusting their emotional responses
				Emotional regulation	Proficiency in planning, monitoring, and adjusting their emotional responses
				Behavioural regulation	Proficiency in planning, monitoring, and adjusting their actions and behaviours
				Cognitive regulation	Proficiency in planning, monitoring, and adjusting their thinking processes to achieve goals
				Persistence	Proficiency in continuing working toward a goal despite challenges
				Inhibition	Proficiency in controlling impulses and refrain from action
		Executive Functioning	Children's proficiency in using cognitive skills to carry out goal-directed behaviour	Meta-cognition	Proficiency in thinking about and reflecting on their own thinking and learning processes
				Cognitive flexibility	Proficiency in adjusting to situational demands, shifting perspectives, and balancing competing needs
				Working memory	Proficiency in holding and manipulating information in mind for short periods
				Attention control	Proficiency in focusing attention on relevant information while ignoring distractions
				Persistence	Proficiency in continuing working toward a goal despite challenges
				Inhibition	Proficiency in controlling impulses and refrain from action
				Task initiation	Proficiency in independently starting a task or activity
Physical Development	Children's proficiency in the physical skills needed for everyday tasks and active play	Gross Motor Skills	Children's proficiency in co-ordinating their large muscles to balance and move around	Stability / balance	Proficiency in maintaining control of their body position whether moving or still
				Co-ordination	Proficiency in using different parts of their body together smoothly and efficiently
				Body awareness	Understanding where their body is in space and how it moves
				Strength	Proficiency in exerting force to move objects or their body
				Locomotion	Proficiency in moving from one place to another using actions such as crawling, walking, running, or jumping
		Fine Motor Skills	Children's proficiency in co-ordinating the small muscles in their hands and fingers to perform precise tasks	Grasping and releasing	Proficiency in grasping and letting go of objects intentionally
				In-hand manipulation	Proficiency in moving and positioning objects within one hand
				Hand-eye co-ordination	Proficiency in using visual information to guide their hand movements
				Palmar grasp	Proficiency in holding objects securely using the whole hand, with the fingers wrapped around the object
				Pincer grasp	Proficiency in holding small objects between their thumb and forefinger

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
				Graphomotor skills	Proficiency in co-ordinating fine motor movements for writing and drawing
Expressive Arts and Design	Children's proficiency in exploring and expressing their ideas, feelings, and creativity	Performing Arts	Children's proficiency in exploring and expressing ideas, feelings, and creativity through movement, music, and drama	Early musicality	Proficiency in perceiving, producing, and responding to musical elements such as rhythm and pitch
				Role play	Proficiency in taking on and acting out roles in imaginative scenarios
				Narrative	Proficiency in structuring and telling a story or recounting events coherently
				Dance	Proficiency in using movement intentionally to express ideas, emotions, or stories
				Creativity	Proficiency in generating and expressing original ideas
		Visual Arts	Children's proficiency in exploring and expressing ideas, feelings, and creativity through drawing, painting, and other visual media	Colour knowledge	Proficiency in identifying different colours and their properties
				Manipulating materials	Proficiency in handling and shaping materials to create or modify objects
				Creativity	Proficiency in generating and expressing original ideas
		Design and Technology	Children's proficiency in creating and adapting materials, structures, and tools to express ideas and solve problems	Understanding function	Understanding how tools, machines, or objects help people in daily life
				Selecting materials	Proficiency in choosing appropriate materials or tools for an identified problem or purpose
				Manipulation materials	Proficiency in handling and shaping materials to create or modify objects
				Creativity	Proficiency in generating and expressing original ideas
Understanding the World	Children's proficiency in making sense of the social and material world around them	Past and Present	Children's proficiency in engaging with and understanding ideas about the order of events and the passage of time	Understanding of sequence	Understanding of the order of events or steps in a process
				Chronological understanding	Understanding the passage of time and how events are ordered in relation to each other
				Chronological vocabulary	The range of words a child can use to describe time and sequence
				Comparison skills	Proficiency in identifying similarities and differences between objects, concepts, or events
				Observation skills	Proficiency in closely watching and noticing details in the environment
		The Natural World	Children's proficiency in engaging with and understanding ideas about living things and the material environment	Comparison skills	Proficiency in identifying similarities and differences between objects, concepts, or events
				Observation skills	Proficiency in closely watching and noticing details in the environment
				Natural process knowledge	Knowledge of phenomena related to nature and the environment
			Children's proficiency in engaging with, understanding	Comparison skills	Proficiency in identifying similarities and differences between objects, concepts, or events

Domain	Domain definition	Sub-domain	Sub-domain definition	Proficiency	Proficiency definition
		People, Culture, and Communities	and appreciating the social world	Observation skills	Proficiency in closely watching and noticing details in the environment
				Religious and cultural knowledge	Knowledge of different religious and cultural beliefs, practices, and traditions
				Sensitivity to difference	Awareness and respect for individual and cultural differences
Digital Technology^a	Children's understanding of proficiency in using digital tools and devices	Technology Awareness	Children's understanding of the purposes and functions of Digital Technology		
		Digital Literacy	Children's proficiency in using digital tools to find, create, and communicate information and solve problems		
Early Learning Environment	The physical, social, and emotional conditions that shape children's early development and learning	Home Learning Environment	The physical, social, and emotional conditions that shape children's early development and learning at home	Physical environment	The physical surroundings where learning and development take place
				Social and emotional climate	The relationships, emotional support, and interactions in a learning environment
				Adult-child interactions	Adults' interactions with children
				Peer interactions	Children's interactions with each other
				Health and hygiene routines	Children's understanding and practise of everyday habits that keep the body clean, healthy, and safe
				Learning resources	The materials and tools available to support children's learning and development
				Cognitive stimulation	The experiences and activities introduced to support children's learning and problem-solving
		Early Years Setting Learning Environment	The physical, social, and emotional conditions that shape children's early development and learning in a non-home environment (e.g. formal early childhood education settings, outdoor learning provision)	Physical environment	The physical surroundings where learning and development take place
				Social and emotional climate	The relationships, emotional support, and interactions in a learning environment
				Adult-child interactions	Adults' interactions with children
				Peer interactions	Children's interactions with each other
				Health and hygiene practices	Health, safety, and hygiene routines and protocols
				Learning resources	The materials and tools available to support children's learning and development
				Cognitive stimulation	The experiences and activities introduced to support children's learning and problem-solving

^a Included as a set of skills underpinning learning and development in other domains.

Appendix B: Search terms

Changes made to the search terms following publication of the protocol (Malik *et al.*, 2025) are highlighted in red, with deleted terms shown using strikethrough (see Appendix C for a rationale for these changes).

Population keywords

(child* OR toddler* OR infant* OR preschool* OR pre-school* OR "early childhood" OR "young children" OR "ECE" OR "ECEC" OR "early years" OR "nursery" OR "kindergarten" OR "reception" OR "baby room" OR "pre primary" or pre-primary)

Measurement/assessment keywords

(measure* OR tool* OR instrument* OR assessment* OR test* OR scale* OR checklist* OR observation*)

Domain-specific keywords

Language and Communication

(language OR communication OR vocabulary OR grammar OR syntax OR morphology OR speech OR "turn taking" OR "pragmatic skills" OR "pragmatic language" OR "joint attention" OR "shared attention" OR gestur* OR "gesticulation" OR "narrative skills" OR "storytelling" OR "narrative competence" OR inferenc* OR "language reasoning" OR "verbal reasoning" OR "code switching" OR "language switching" OR "bilingual language use" OR "EAL")

Literacy

("literacy" OR "pre-literacy" OR "reading" OR "writing" OR "print awareness" OR "print concepts" OR "concepts of print" OR "print knowledge" OR "emergent literacy" OR "book handling skills" OR "understanding of print" OR "text directionality" OR "recognition of print" OR "print recognition" OR "letter naming" OR "letter name knowledge" OR "alphabet knowledge" OR "letter identification" OR "letter recognition" OR "naming letters" OR "phonological awareness" OR phoneme OR phonics OR "phonological processing" OR "sound awareness" OR "sound discrimination" OR "syllable segmentation" OR "rhyme" OR "letter-sound " OR "grapheme-phoneme correspondence" OR "sound-symbol knowledge" OR alphabe* OR decoding OR "early word recognition" OR "rapid automatized naming" OR "rapid automatised naming" OR "RAN" OR "rapid naming" OR "naming speed" OR "rapid serial naming" OR "naming fluency" OR "code switching" OR "language switching" OR "bilingual language use" OR "language alternation" OR "code mixing" OR translanguaging OR "comprehension" OR "prosody")

Mathematics

(math OR maths OR mathematics OR numeracy OR pre-numeracy OR "pattern awareness" OR "pattern matching" OR "digit recognition" OR "naming digits" OR "number word*" OR counting OR "one-to-one correspondence" OR cardinality OR subitising OR "number composition" OR "place value" OR "comparing *magnitude" OR "numerical patterns" OR "additive reasoning" OR calculation OR "inverse operations" OR "symbol knowledge" OR "strategy selection" OR "non-standard units" OR "estimate*" OR "sequenc*" OR "shape recognition" OR "shape decomposition" OR "2D-3D relationships" OR "spatial language" OR "perspective taking")

PSED

("personal development" OR socioemotion* OR socio-emotion* OR self-esteem OR self-concept OR "sense of self" OR self-care OR "healthy behaviours" OR "safety behaviours" OR "social skills" OR "theory of mind" OR mentalis* OR empathy OR "relationship skills" OR "prosocial behaviour" OR "conflict resolution skills" OR "emotion skills" OR "emotion recognition" OR "emotion expression" OR "emotion vocabulary" OR "emotion naming" OR "emotion understanding" OR "emotion regulation" OR "emotional regulation" OR "self-regulation" OR "behavioural regulation" OR "behaviour regulation" OR "behavioral regulation" OR "behavior regulation" OR "cognitive regulation" OR persistence OR inhibit* OR "executive function*" OR metacognition OR "cognitive flexibility" OR "working memory" OR "attention control" OR "sustained attention" OR "task initiation")

Physical Development¹¹

("gross motor skills" OR stability OR balance OR coordination OR "body awareness" OR strength OR locomotion OR "fine motor skills" OR grasping OR releasing OR "in-hand manipulation" OR "hand-eye coordination" OR "hand eye coordination" OR "palmar grasp" OR "pincer grasp" OR "graphomotor skills" OR "tool manipulation")

Expressive Arts and Design

(art* OR arts* OR "design and technology" OR music* OR "role play" OR drama OR dance OR creativ* OR imaginat* OR sing OR "dramatic play" OR "colour knowledge" OR "manipulation of materials" OR sculpting OR paint* OR draw*)

Understanding the World

("understanding the world" OR "past and present" OR "order of events" OR "passage of time" OR "understanding of sequence" OR "chronological understanding" OR "chronological vocabulary" OR chronology OR "comparison skills" OR "observation skills" OR "the natural world" OR nature OR biology OR science OR "living things" OR "material environment" OR natur* OR geography OR "people, culture and communities" OR "religious and cultural knowledge" OR "religious education" OR "cultural education" OR "sensitivity to difference" OR "respect for diversity" OR "appreciation of differences" OR "cultural sensitivity" OR "cultural awareness" OR inclusivity OR open-mindedness OR "acceptance of others" OR "intercultural understanding" OR "social awareness" OR "valuing diversity" OR "recognition of individual differences" OR "awareness of social and cultural norms" OR "understanding of diverse perspectives")

Digital Technology

("technology awareness" OR "digital literacy" OR "technology skills" OR "ICT" OR "computing skills" OR "digital skills" OR "media literacy" OR "online literacy" OR "technological fluency" OR "digital competence" OR "digital awareness" OR "tech literacy" OR "information literacy" OR "computer literacy" OR "digital understanding" OR "digital navigation skills")

Early Learning Environment

("learning environment*" OR "learning resource*" OR "learning process quality" OR "adult-child interaction*" OR "teacher-child interaction*" OR "caregiver-child interaction*" OR "peer interaction*" OR "health routine*" OR "hygiene routine*" OR "care routine*" OR "cognitive stimulation" OR "teaching process quality" OR "instructional quality" OR "teacher quality" OR "teaching effectiveness" OR "classroom quality" OR "educational environment" OR "teacher-student interaction" OR "learning support" OR "early childhood education quality" OR "curricul*" OR "pedagog*")

¹¹ We did not include Physical Development as a search term as we believed the scope will be too broad. We noted that this would be revisited if the more specific search terms yield insufficient results. However, this was ultimately not needed.

Appendix C: Changes and decisions following protocol publication

Appendix C documents refinements to the taxonomy and review process made after publication of the protocol (Malik *et al.*, 2025). These changes reflect learning from the review itself and were made to improve feasibility, clarity, and usability.

Changes to taxonomy

We were explicit from the outset that the taxonomy would be iterated and finalised throughout the duration of the project, and that it might be updated as the review progressed and as we examined which proficiencies are most commonly measured within each sub-domain. However, relatively few changes were required to the originally proposed taxonomy. The proficiencies included in the taxonomy effectively captured the breadth of outcomes targeted by identified measures, suggesting that the combination of EYFS alignment, desk review, and expert consultation had resulted in a robust and well-calibrated structure.

We stated that we would not remove proficiencies simply because no measures were available. However, we also did not add new proficiencies during the review. Adding further proficiencies would have made the taxonomy too granular, reducing usability and increasing the risk of fragmentation without clear practical benefit. Moreover, proficiencies identified during the review could be reasonably mapped onto existing proficiency categories, which meant that adding more proficiencies was not necessary. The only substantive change to the taxonomy was the removal of 'bilateral co-ordination' as a proficiency from the Physical Development domain. This was because we found that it overlapped heavily and was often hard to distinguish from with co-ordination and hand-eye co-ordination when reviewing measures. Retaining it risked double-counting the same underlying proficiency. Moreover, the proficiency was rarely measured as a standalone construct or assigned to measures in early years assessment tools.

Some proficiencies and sub-domains were retained despite having few or no associated measures. In the Digital Technology domain, for example, no tools explicitly measured technology awareness or digital literacy.

Other proficiencies with few or no measures, such as inverse operations, place value, estimation, musicality, dance, strength, and chronological vocabulary, were largely those expected to be developmentally advanced for much of the early years age range or inherently difficult to capture through standardised assessment. We have deliberately retained these proficiencies in the taxonomy as it highlights gaps in the current measurement landscape and identifies areas that may warrant further tool development, instead of implying that these outcomes are unimportant. We discuss the distribution of measures across domains and proficiencies in more detail in the report section 'Description of database'.

Changes to the process of identifying a longlist of measures

The protocol (Malik *et al.*, 2025) set out PsycINFO, SCOPUS, and ERIC as the academic research databases to be used to identify measures. In practice, we conducted academic searches using only PsycINFO for all the domains other than Self-Regulation and Executive Function, which had initially been explored through pilot Google Scholar searches. This was because PsycINFO allowed search results to be exported, making it integrate effectively with EPPI-Reviewer. We do not anticipate this having major implications for the evidence base for several reasons. First, PsycINFO has extensive coverage of psychological and developmental research, and many early years measures are developed and validated within child development and psychology journals rather than education-focused outlets. There is also substantial overlap in indexing across databases—ERIC in particular tends to have a stronger focus on Unites States-based education research, whereas PsycINFO has broader international coverage. Second, measures with a strong evidence base that would have met the shortlisting criteria are likely to be indexed across multiple databases. Third, the database search was supplemented by targeted approaches, including expert recommendations, the existing EYMD, and searches of major assessment publishers and organisations (Pearson, GL Assessment, Riverside Insights, Brookes Publishing, Hogrefe, PAR, PRO-ED). Together, these strategies provided broad coverage of widely used and well-established measures, as was indicated by the number of measures that were longlisted. As such, we do not expect that searching additional databases would have substantially expanded the pool of measures identified. We managed results using EPPI-Reviewer, which allowed records to be organised, screened, and de-duplicated in a more structured and transparent manner. The use of PsycINFO, with

exportable search results, enabled systematic screening of all retrieved records and therefore more comprehensive coverage than manual Google Scholar searches.

The PsycINFO search was limited to studies relating to children aged from birth to five years, as this was the age range available as a filter within the database. We do not believe this resulted in the omission of relevant measures for older children as measures targeting six-year-olds were well represented through other longlisting sources. In total, 219 measures in the longlist targeted children aged six years, second only to those targeting four- and five-year-olds, indicating that coverage was not impacted through the use of this filter.

During screening, **papers that did not mention a specific assessment tool in the title or abstract were excluded in the first place**. These records were tracked separately from clearly irrelevant studies, allowing targeted follow-up if there was low coverage of a specific domain. This ensured that the process was feasible without running the risk of missing relevant measures embedded within broader studies. Only sources describing or validating assessment tools were included. Meta-analyses, intervention studies, and studies reporting outcomes without detailing the measurement instrument were excluded.

While longlisting, **coverage in the Early Learning Environment domain was initially sparse**. Based on internal expert judgement that additional relevant measures were likely to exist, targeted supplementary searches were undertaken to strengthen coverage in this area. These included expert recommendations from the project team and external advisors, targeted searches using Google Scholar to identify commonly cited instruments, and manual searches of key reviews and reference lists within relevant papers. In these additional searches, we used similar search strings to those applied in the main search strategy. We also reviewed publications by researchers known to work in this area and examined relevant papers to identify measures referenced within them. These additional searches resulted in the identification of further measures, which were then screened using the same criteria as the rest of the longlist to ensure consistency.

As set out in the **protocol (Malik et al., 2025)**, **we committed to documenting any changes to the search strategy following publication**. Only minor refinements were made to improve precision and narrow down the search results. For mathematics, the term “measur*” was removed as it frequently captured out-of-scope references, and we restricted searching for the keywords to abstracts only instead of the full text. Similarly, for Expressive Arts, “design*” was removed to avoid unrelated uses, “arts*” was refined to “art OR arts” to prevent retrieval of terms such as “article”, and the asterisk was removed from “draw” to exclude references to sampling or study design. We also restricted searching for the keywords to abstracts only for the Expressive Arts domain. These refinements improved the relevance of search results and reduced the number of search results to be scanned. The final search terms used are documented in Appendix B.

For each longlisted measure, basic descriptive information was recorded, including assessment name and version, a brief summary provided by the author or publisher where available, the domain(s) targeted, age range, and publisher or author details with links to the assessment. Although we initially proposed recording sub-domains at the longlisting stage, we only included information on the domains assessed. Sub-domain information was rarely explicit in publicly available documentation and would have required substantial interpretation, with limited added value at this early stage. Resources were therefore prioritised for more detailed appraisal of shortlisted measures, where sub-domain and proficiency-level coding was both more feasible and more useful to users of the database.

Changes to the process of appraising shortlisted measures

During quality assurance of the detailed appraisal, **we identified a subset of measures that were not norm-referenced and had been validated on non-representative samples**. As these measures had already undergone full appraisal, we reviewed our approach and decided that exclusion solely on the basis of non-normed status would remove tools that remain useful in research and practice contexts. Therefore, we retained shortlisted measures that were not standardised as long as their intended use aligned with how the measure is likely to be applied, rather than whether it enables comparison against a nationally normed sample. For example, where a measure is designed to monitor progress, inform teaching, or guide interventions, rather than to benchmark individuals against population norms, national standardisation is not essential, and non-normed measures can still provide meaningful and actionable information, and the measure may still be considered fit for purpose. To ensure consistency, we applied this revised criterion to all measures. Norming status and

sample characteristics are explicitly captured in separate fields within the database and are displayed distinctly in the user interface.

While the **protocol** (Malik *et al.*, 2025) specified reliance on publicly available information, **a small number of potentially relevant measures had key technical information available only behind a paywall**. In agreement with the EEF, resources originally allocated for evaluator interviews were instead used to access this information, primarily through the purchase of technical manuals. Of 11 measures initially identified as candidates for this approach, a further assessment of relevance and cost-effectiveness narrowed the list down to eight. Publishers were first contacted to request access to technical manuals without purchase. Where this was not possible, materials were purchased following agreement with the EEF and included in the appraisal.

We included optional evaluator interviews in the protocol (Malik *et al.*, 2025) to gather additional information on the experiences of evaluators using measures in the EEF trials if trial reports did not provide sufficient detail. However, information available in the EEF published reports was sufficient for appraisal, and evaluator interviews were not conducted, as they were unlikely to add substantive insight beyond the documented evidence.

Appendix D: Extraction template

[illegible]

[illegible]

Proficiencies Covered	Proficiencies																												
Proficiencies Covered (separated by a comma)	Communication						Language								Reading										Writing				
	Joint attention	Gesticulation	Turn taking	Speech	Pragmatic skills	Code switching	Expressive vocabulary	Receptive vocabulary	Syntax	Morphology	Grammar	Narrative skills	Inferencing	Language reasoning	Print awareness	Letter naming	Phonological awareness	Letter-sound correspondenc e	Word reading	Rapid automatised naming	Code switching	Reading comprehension	Reading fluency	Graphomotor skills	Mark making	Spelling	Handwriting	Written composition	

Number										Operations								Measurement					
Pattern awareness/ matching	Recognising and naming digits	Linking number words, digits, and quantities	Counting	Subitising	Number composition	Place value	Comparing numerical magnitude	Numerical patterns	Mathematical talk	Additive reasoning	Calculation	Inverse operations	Symbol knowledge	Strategy selection	Comparing numerical magnitude	Numerical patterns	Mathematical talk	Non-standard units	Estimation	Comparing physical magnitude	Comparing quantities	Pattern awareness/ matching	Mathematical talk

[illegible]

Self-regulation					Executive functioning							Gross motor skills						Fine motor skills						
Emotional regulation	Behavioural regulation	Cognitive regulation	Persistence	Inhibition	Meta-cognition	Cognitive flexibility	Working memory	Attention control	Task initiation	Persistence	Inhibition	Stability/ balance	Coordination	Body awareness	Strength	Locomotion	Grasping and releasing	In-hand manipulation	Hand-eye coordination	Palmar grasp	Pincer grasp	Graphomotor skills		

[illegible]

Home learning environment															Psychometric Properties					
Home learning environment							Early Years Setting Learning Environment							Standardised? (Yes/No)	UK Standardisation sample size		UK Standardisation sample characteristics		Non-UK Standardisation sample	Validation sample size and characteristics
Physical environment	Social and emotional climate	Adult-child interactions	Peer interactions	Health and hygiene routines	Learning resources	Cognitive stimulation	Physical environment	Social and emotional climate	Adult-child interactions	Peer interactions	Health and hygiene practices	Learning resources	Cognitive stimulation	If yes, fill columns GA-GE (as applicable) If no, fill column GF	Rating	Explanation	Rating	Explanation		

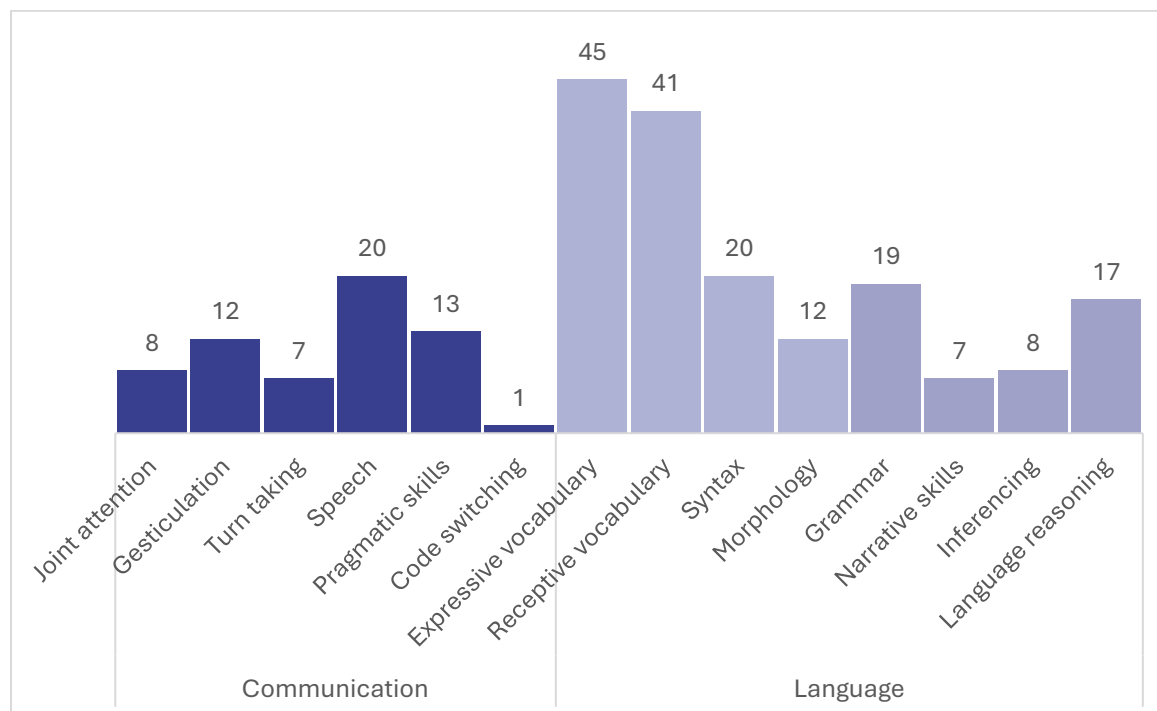
[illegible]

Implementation Properties													
Administration format	Response Mode		Item Response Types		Respondent	Prior Qualifications Required	Ease of Scoring	Availability of guidance	Time Required			Cost	Availability
	Response Mode	If 'Other', give details	Item Response Type	If 'Other', give details					Prep Time	Administration Time	Scoring Time		

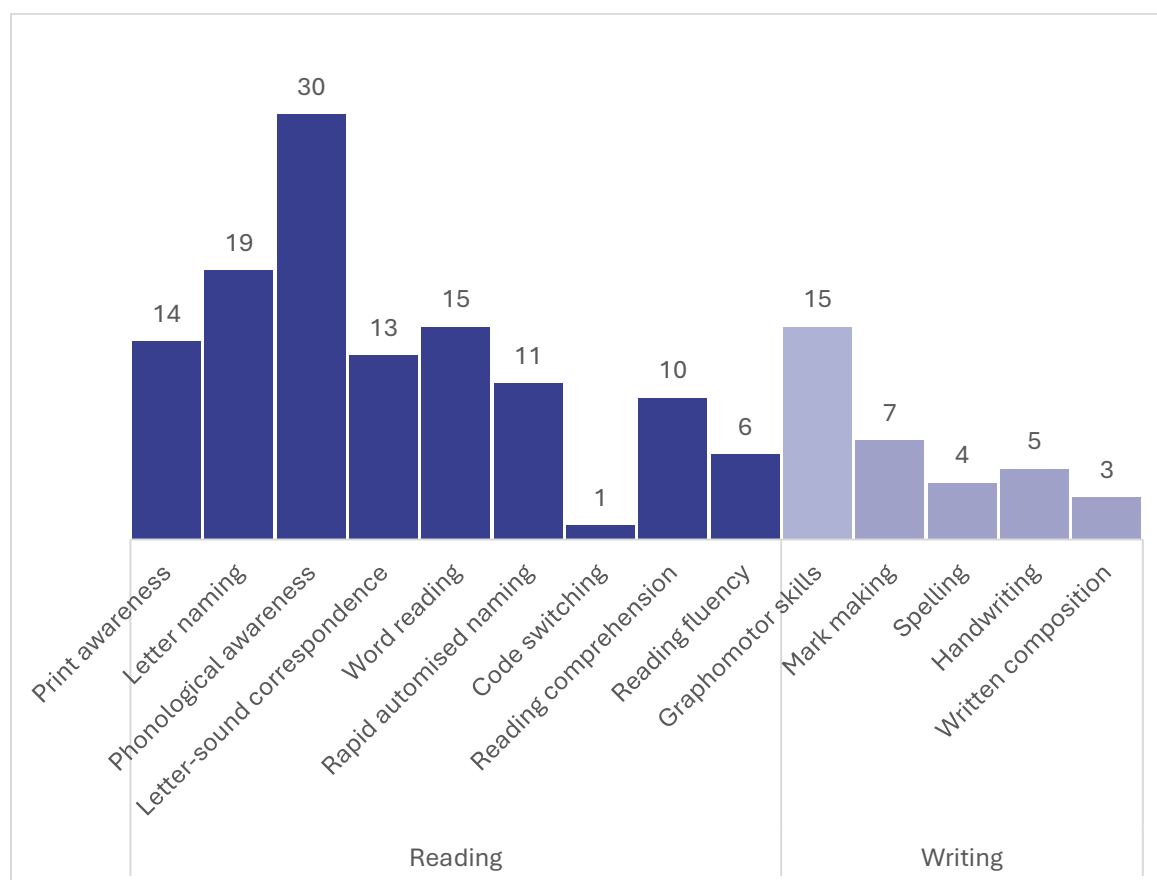
Use in EEF Trials (as of December 2025)							
Description of use(s)	Key Learnings	Adaptations made	Sources of information				
			Reference 1	Reference 2	Reference 3	Reference 4	Reference 5

Appendix E: Proficiencies coverage by domain

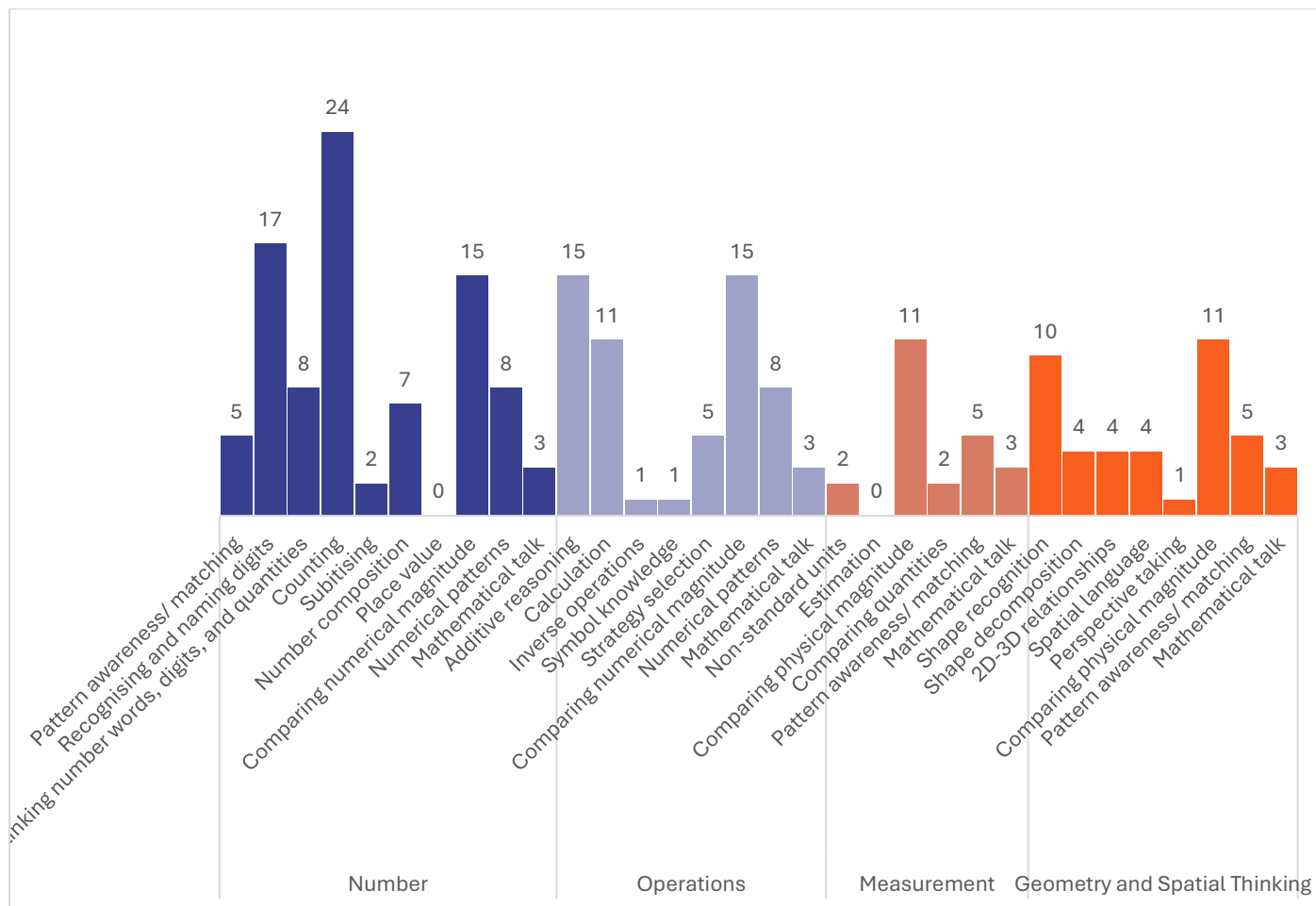
Appendix E Figure 1: Proficiencies coverage for Communication and Language



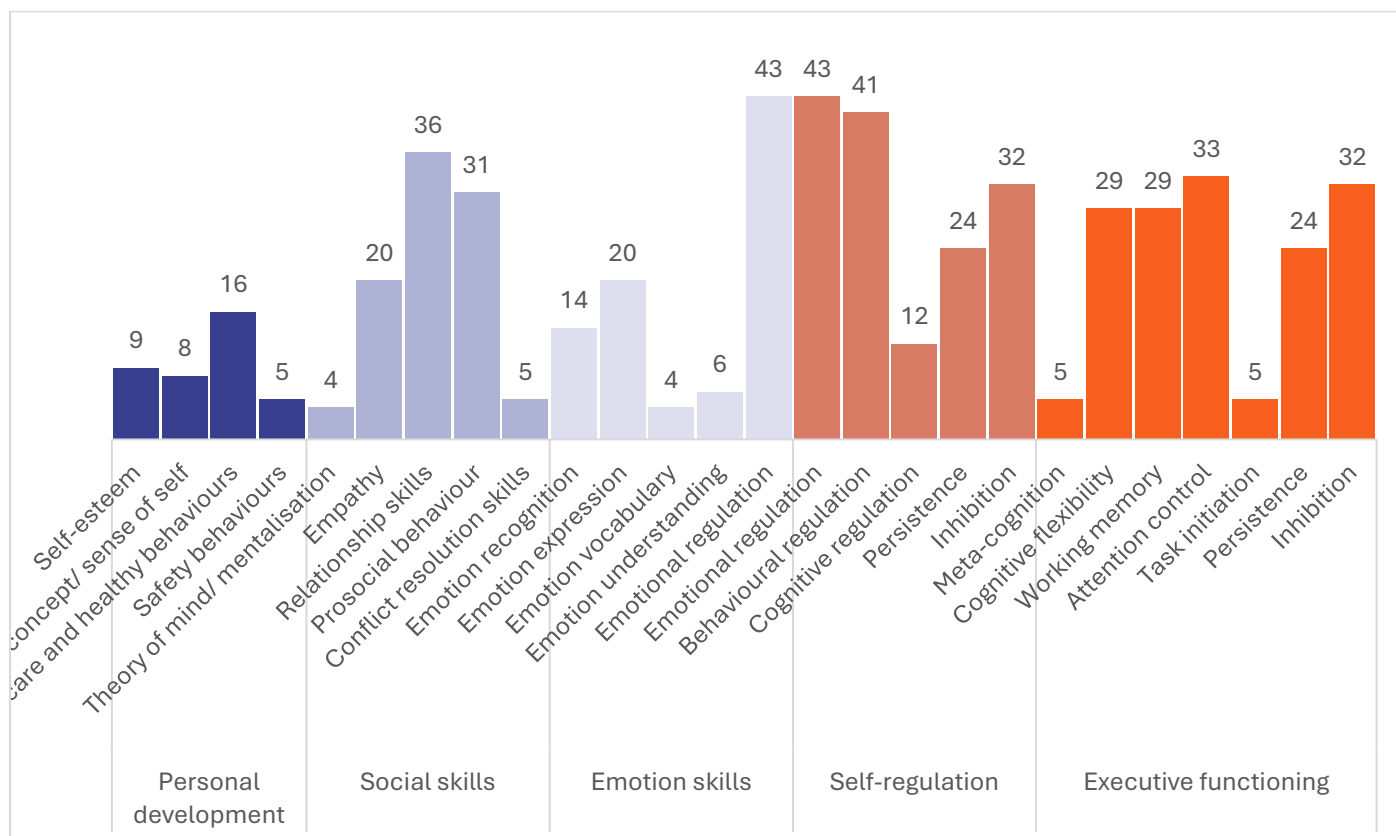
Appendix E Figure 2: Proficiencies coverage for Literacy



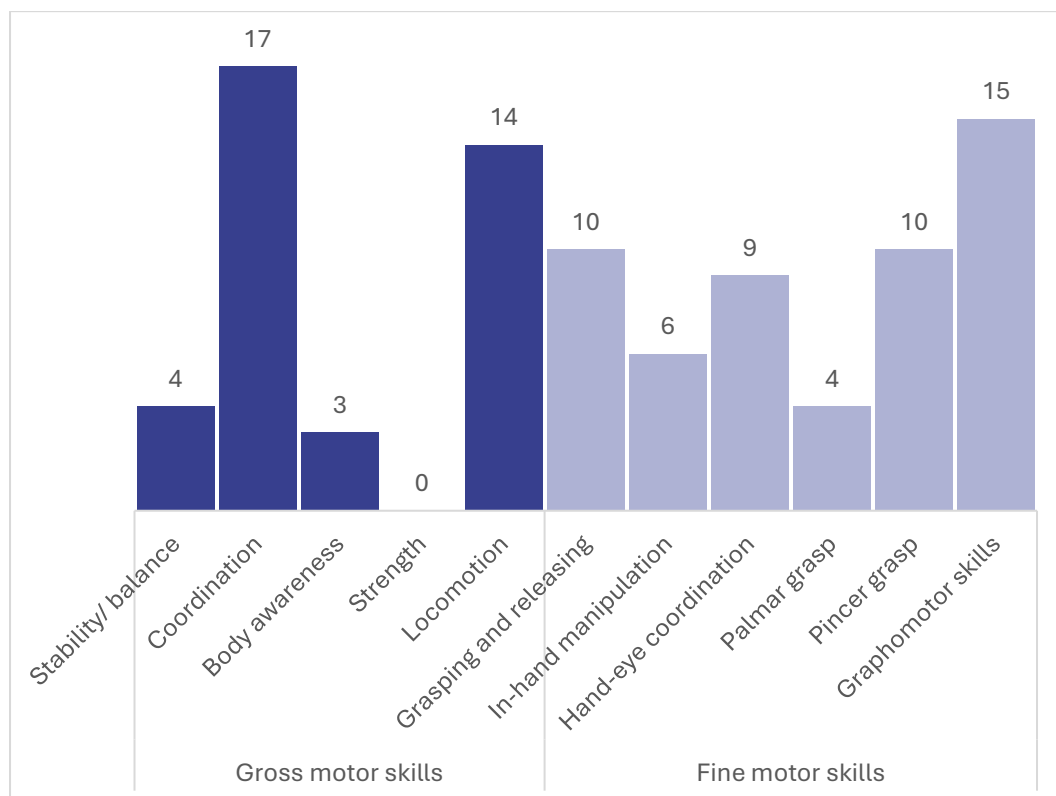
Appendix E Figure 3: Proficiencies coverage for Mathematics



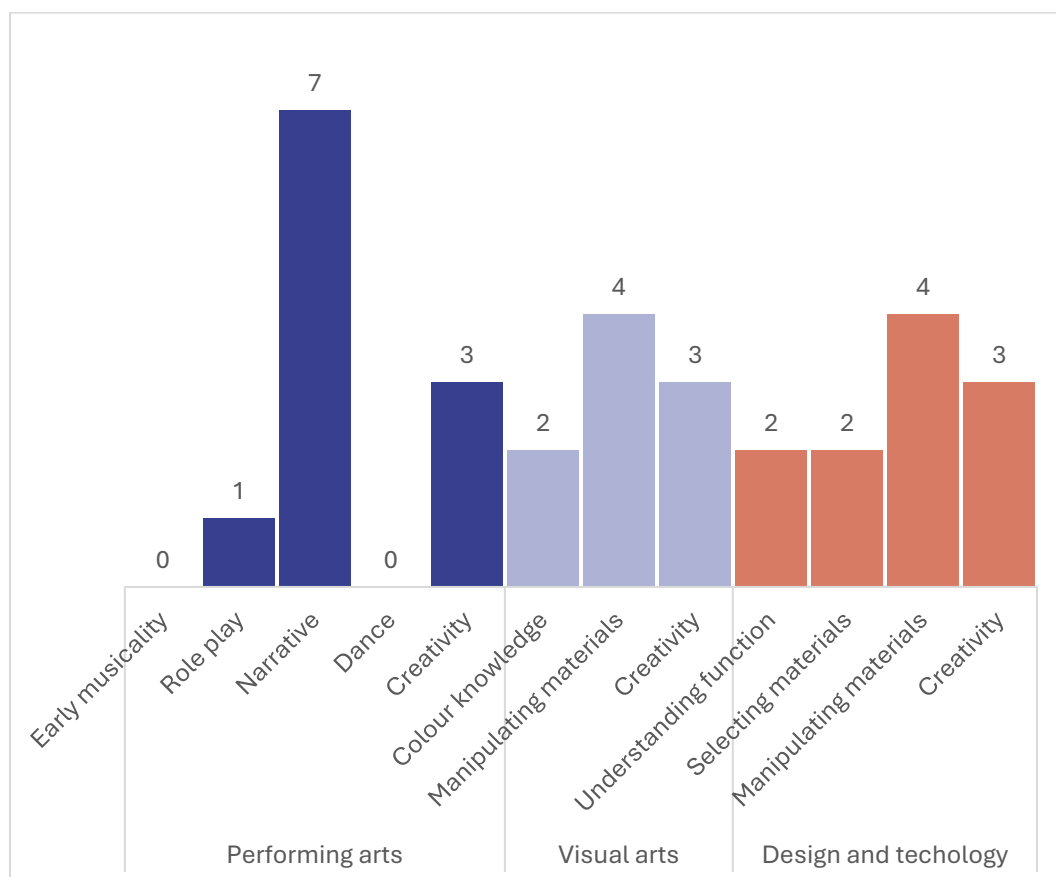
Appendix E Figure 4: Proficiencies coverage for PSED



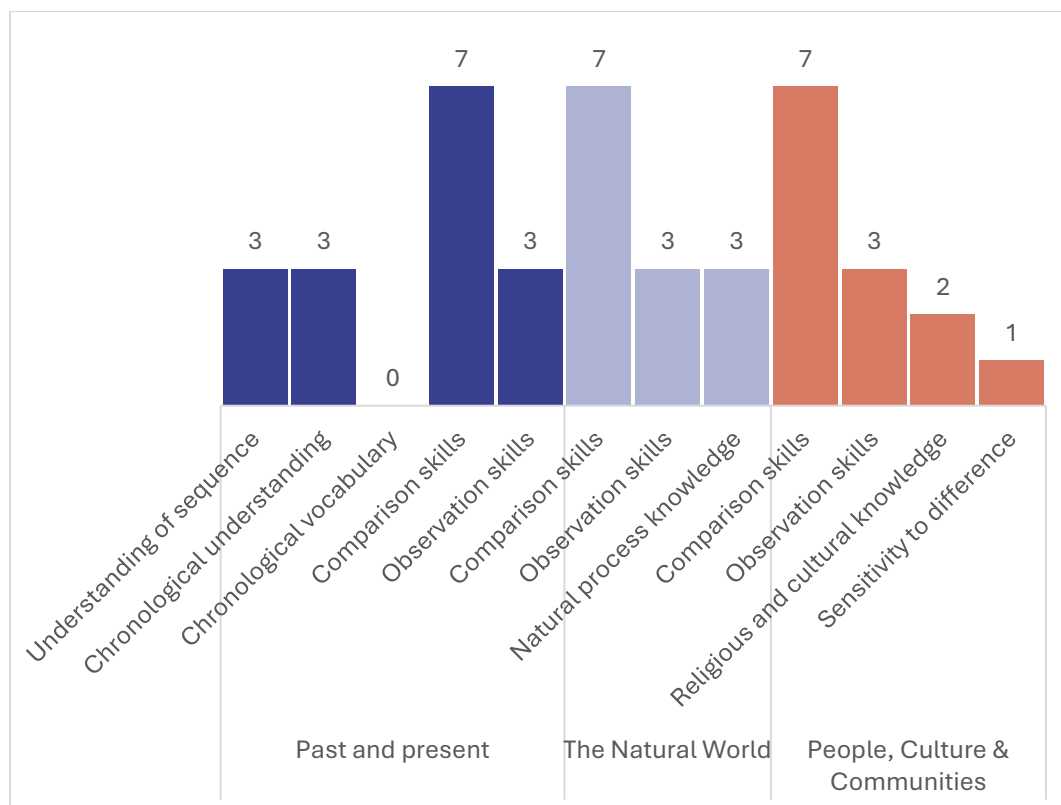
Appendix E Figure 5: Proficiencies coverage for Physical Development



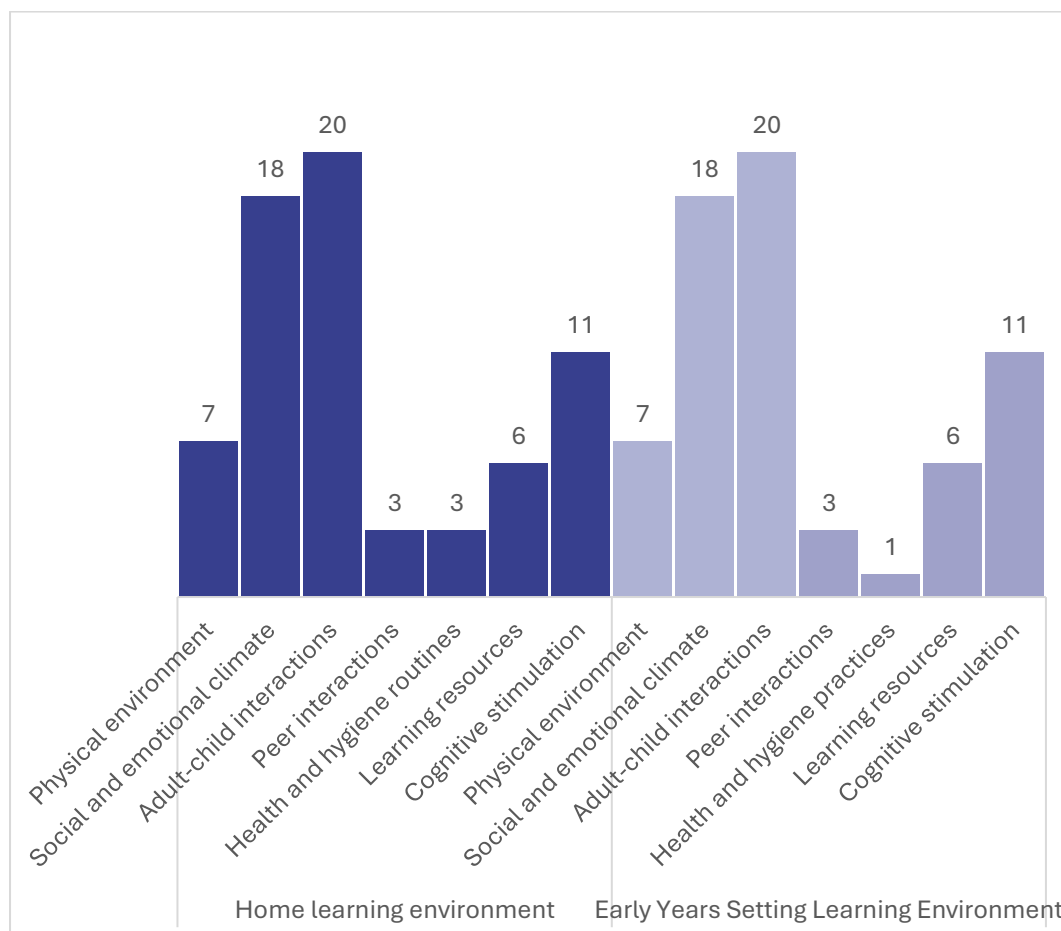
Appendix E Figure 6: Proficiencies coverage for Expressive Arts and Design



Appendix E Figure 7: Proficiencies coverage for Understanding the World



Appendix E Figure 8:2 Proficiencies coverage for Early Learning Environment

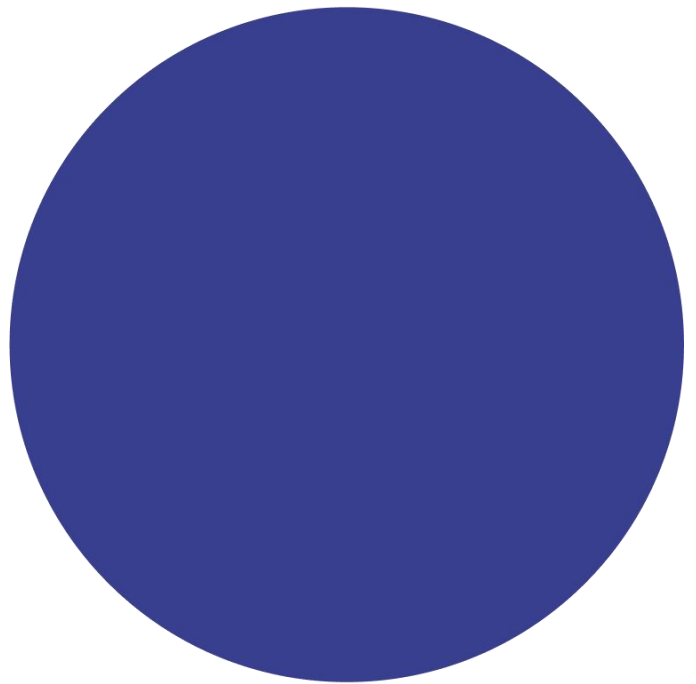
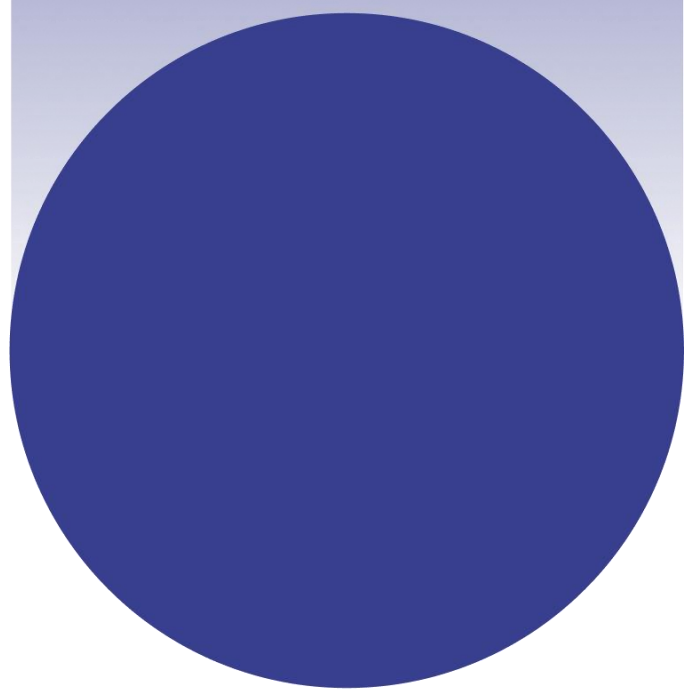


You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower,
21–24 Millbank,
London,
SW1P 4QP

<https://educationendowmentfoundation.org.uk>

 @EducEndowFoundn

 [Facebook.com/EducEndowFoundn](https://www.facebook.com/EducEndowFoundn)