Education
Endowment
Foundation

A systematic review of standardised measures of attainment in literacy, mathematics, and science

Evidence Review

June 2021

Helen L. Breadmore and Julia M. Carroll

Research Centre
Global Learning

Coventry University

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education. Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

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

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

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator

Principal investigators: Dr Helen L. Breadmore and Prof Julia M. Carroll

Centre for Global Learning,
Coventry University,
Priory Street,
Coventry,
CV1 5FB

Dr Breadmore, **ab8179@coventry.ac.uk.**

Date of search: July 2020

Disclaimer: The review was conducted by an independent team based on inclusion and exclusion criteria determined a priori for each phase and documented in a protocol. The list of measures included in the report and corresponding database are a result of this individual review, but we do not claim that this is an exhaustive list of all tests available to measure attainment in literacy, mathematics, and science.

Executive summary

This review of measures of attainment in literacy, maths, and science provides much-needed guidance to support the selection of standardised tests. This report, and the associated database, will help anyone seeking a measure to benchmark students' performance against national attainment. The review identifies measures that fulfil minimal reporting criteria, summarises information on the reliability and validity of test data, as well as providing practical information about test administration and implementation to help you to determine which test best suits your needs, whether that is to understand strengths and weaknesses, measure progress over time, or evaluate an intervention.

In all cases we aimed to focus on measures of attainment rather than cognitive abilities or specific skills. Literacy, mathematics, and science are complex, multi-dimensional subjects and the key constructs of knowledge that determine attainment change over the course of development as different skills develop and subject knowledge is taught. Here, we seek to evaluate measures of overall attainment in each subject. Nonetheless, consideration will be given to whether tests are specific or general measures of attainment as this has relevance for structural validity. Specific tests measure only one key concept or area of content knowledge. For example, a spelling test would be a specific measure of literacy attainment while an arithmetic test would be a specific measure of mathematics attainment. General measures of attainment are multi-dimensional, assessing more than one key concept or area of content knowledge.

Objectives

Our approach focuses on tests of particular relevance to educators and evaluators in the U.K. who wish to measure the attainment of children and adolescents aged 6 to 18 years benchmarked against a nationally representative sample. The evidence is summarised here in a written synthesis and also presented in a searchable database.

The research questions are:

1. How can teachers and evaluators assess attainment and progress in literacy, mathematics, and science in the U.K.?
2. What is the psychometric quality and implementation utility of the standardised tests identified through this review for use with pupils aged 6 to 18 years old?

Inclusion and exclusion criteria and rationale

There were two phases to the systematic review search process: (1) test identification and (2) publication identification. Inclusion and exclusion criteria were determined a priori for each phase and documented in the systematic review protocol (Breadmore and Carroll, 2020).

The test identification phase formed the long list database of 231 tests. Initially during this phase, tests were identified with the support of our advisory panel of experts, by reading EEF studies and communicating with the EEF, hand searching 18 publisher and distributor websites, and searching the ERIC database. To be included at this stage, tests had to be:

- used to assess literacy, mathematics, or science attainment—in all cases we aimed to focus on overall measures of attainment rather than cognitive abilities or specific skills;
- published in or since 2000—to ensure relevance of test content; and
- suitable for English-speaking 6- to 18-year-olds.

Tests identified in this way were then screened, additional information (such as test manuals) was gathered from publishers and through systematic searches for peer-reviewed publications, and eligibility checks were performed to ensure that the information needed to evaluate the measures was available. During these screening and eligibility checks, a number of tests were excluded for not meeting certain criteria:

- 11 tests were not available for review—because not available in the U.K. or out of print;

- 94 tests were criterion-referenced or not norm-referenced—these tests can be very useful for assessing attainment but cannot be evaluated using the same benchmarks as norm-referenced tests;
- 3 tests were not applicable to sample—for example, tests intended for use with clinical populations were excluded to ensure relevance of test content and norms to target sample;
- 50 had not been normed on a U.K. sample—this is essential to ensure applicability of norms to target sample;
- 32 did not have recent norms available—recent norms are essential to ensure the test results can be generalised to the target sample, hence the test must have been published since 2010, or had updated norms published since then; and
- 4 tests were removed due to insufficient information available for evaluation—validity and reliability could not be evaluated from the information we gathered.

Methodology

Thirty-seven tests were subjected to full evaluation using selected questions about implementation utility, reliability, validity, and quality of norms from the European Federation of Psychologists' Associations test review model (Evers, Hagemeister, et al., 2013).

Outcome of search and evaluation

How can teachers and evaluators assess attainment and progress in literacy, mathematics, and science in the U.K.?

We identified 231 tests, which are included in the long list database. However, only 37 were eligible for full evaluation. We considered the availability of tests to measure attainment in each subject for primary- and secondary-aged pupils. Note that some tests are suitable for assessing both primary and secondary pupils, or measure both literacy and mathematics. Those tests were counted multiple times in these analyses.

- For primary-aged pupils, there were 18 tests of attainment in literacy, 16 in mathematics and 1 in science.
- For secondary-aged pupils, there were 15 tests of attainment in literacy, 9 in mathematics, and 1 in science.

A large proportion of tests were removed because they were criterion-referenced or not norm-referenced and therefore could not be evaluated using the criteria chosen for this review. Many of those tests are well established measures and their exclusion from this evaluation should not be seen as implying inadequacy. Some subject areas, including science, might lend themselves more readily to criterion-referenced testing to assess attainment. Further research should evaluate the reliability and validity of these criterion-referenced tests.

Of the tests on the long list that were normed, a large proportion of the norms could not be generalised to the target population because the norms were old or generated from non-U.K. samples. Again, this included some well-established tests. Publishers and test developers should be encouraged to conduct re-standardisation trials to update the norms.

It was notable that only one science test fulfilled eligibility criteria for evaluation. This is a significant gap, which test developers should be encouraged to resolve.

What is the psychometric quality and implementation utility of the tests identified through this review for use with pupils aged 6 to 18 years?

'Implementation utility' refers to how easily a test can be used. It is subjective, and dependent on a multitude of factors including the purpose for the assessment, availability of resources (including facilities, money and time), the child(ren) being assessed, and the tester. As such, implementation factors were summarised in the evaluation phase but were not rated. Implementation factors considered include:

- the need for the person administrating or scoring the test to have appropriate prior experience, training, or accreditations;
- the costs associated with the test (in terms of time, resources and equipment); and
- the format of administration and scoring.

Many of the psychometric properties of tests can be evaluated objectively. This evaluation crucially depends on considering the validity and reliability of the test results as well as the quality of norms. We examined 'construct validity', 'criterion validity', and 'reliability'.

Construct validity examines the extent to which the test actually measures what it sets out to measure or, instead, partially or mainly measures something else. We rated construct validity on a 0–4 point scale (0 indicating insufficient information was available to evaluate construct validity, 1 indicating weak descriptive and statistical evidence of validity, 4 indicating strong descriptive and statistical evidence of validity). While construct validity was typically moderate to good, only five tests achieved the highest score. Users of the measures database should be reminded that construct validity also depends upon alignment of their intended target construct to the construct measured by the test.

Criterion validity considers the extent to which test scores are related to scores on a real world measure of the construct, such as national key stage tests or GCSEs. This review established that evidence criterion validity was only available for ten tests. Those tests were rated on a 0–4 point scale (0 indicating that no evidence was available to review, 1 indicating a single source of inadequate statistical evidence of validity, 4 indicating multiple sources of strong statistical evidence of validity).

Reliability refers to the extent to which the test scores are likely to be reproducible. Reliability was rated on a 0–4 point scale (0 indicating that no evidence was available to review, 1 indicating weak descriptive and statistical evidence of validity, 4 indicating strong descriptive and statistical evidence of validity). In most cases, reliability was moderate or good. Most tests present a single measure of reliability (usually internal consistency) but do not assess temporal or equivalence reliability.

Conclusion

The majority of tests on the market do not have recent U.K. norms; this is particularly true in science.

Our review highlights that information about the validity and reliability of measures of attainment is often difficult to access, lacking, or of low quality. Where such information was available, it was often found in technical manuals, which are not usually available for users to review until after a test has been selected and purchased. We recommend that publishers provide accessible summaries of this information on their websites.

In addition, we noted that that very few measures reported criterion validity—that is, the relationship between the measure of attainment and school outcome tests such as key stage tests or GCSEs. This is disappointing and problematic because in many cases these attainment measures are marketed as a way for schools to predict performance on these tests. In some cases, predicted national test grades are one of the measures that the test will provide. We would urge significant caution in teachers using these predicted grades and suggest that in many cases their own professional judgement of students they have known over a period of time would be a better predictor of outcome. Test users are reminded that all measures of attainment are based on observations on the day of the assessment and should correctly be considered an estimate of the examinee's true level of attainment combined with some degree of measurement error.

We also note that test manuals often recommend adaptations to administration without presenting any evidence of equivalence reliability: for example changing between paper based and digital administration, or between group and individual administration. Sometimes small adaptations are necessary for fairness in testing, or for practical reasons, however, we urge caution in assuming that changes to test administration have no effect on the reliability and validity of outcome scores.

Background and review rationale

This review evaluates measures of attainment in literacy, mathematics, and science in order to support educators and evaluators when selecting tests. Test selection must be informed by consideration of the purpose of the assessment, which leads to consideration of what construct the user should measure and how.

A distinction should be made between tests and assessments. Here, we adopt definitions from the U.S. Standards for Educational and Psychological Testing (Joint Committee of the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014, henceforth 'the Standards'): a *test* applies standardised procedures to sample, evaluate, and score an examinee's behaviour in a specified domain; *assessment* is a broader term, where test information is combined with other sources such as several tests, educational history, and so on.

Educators and evaluators need to measure attainment in order to:

- track pupil attainment over time;
- understand individual pupil's patterns of strengths and weaknesses;
- identify individual pupils who may benefit from targeted support;
- consider the effectiveness of changes in teaching methods and resources at pupil, class, or school level; and
- evaluate the effectiveness of interventions.

This review focuses on *how* to measure attainment but the reader is reminded to consider first what to measure and why. The validity of test data depends upon a good match between the aims of the user and of the assessment. For example, some tests will provide data that is more useful for making decisions about individual children while others are better for describing classes, schools, or making recommendations for public policy. Some measures are ideal for measuring change over time through repeated testing while others are designed to be given only on a single occasion.

There are many measures of attainment available but it is not always easy to decide which to use. To select the most appropriate test it is essential to consider both the psychometric properties of the test as well as practical implementation factors (Evers, Muñiz, et al., 2013). The psychometric properties of the test indicate whether the assessment is a valid and reliable measure of the constructs of interest and for the population of interest. Evaluation of implementation factors reflects how easy it is to use the test in a particular situation.

While the psychometric properties of a test can be evaluated objectively, preference over implementation factors is more subjective. Preference depends on the user, the context of the assessment, the resources available, and the purpose for the assessment. Implementation factors to consider include:

- the need for the person administrating or scoring the test to have appropriate prior experience, training, or accreditations;
- the costs associated with the test (in terms of time, resources, and equipment); and
- the format of administration and scoring (such as whether responses are multiple choice or open ended, recorded on paper or electronically, and whether the test is delivered to a group of students or an individual).

The core skills of literacy, mathematics, and science are essential to learning across all educational domains. Attainment in these subjects are key indicators of individual, school, national, and international scholastic achievement more broadly. For example, these subjects are the focus of assessment and comparison in the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA https://www.oecd.org/pisa) and the International Association for the Evaluation of Educational Achievement (IEA) through the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS https://timssandpirls.bc.edu).

The National Curriculum in England (DfE, 2014) defines English, mathematics, and science as 'core subjects', compulsory throughout every key stage of education. Moreover, it explicitly states that teachers should develop

language, literacy, numeracy, and mathematics across every relevant subject because these skills underpin success in all other areas of the curriculum.

Currently, there are few sources of impartial guidance and information to find and compare tests of literacy, mathematics, and science attainment in school-age children in the U.K. The Education Endowment Foundation SPECTRUM database and Early Years Measures Database evaluate tests for other constructs and populations but, to our knowledge, there is not a comparable resource for literacy, mathematics, and science attainment in school-age children in the U.K. Indeed, much of the information that users need to make an informed choice is within test manuals and therefore behind a paywall.

The aim of this review is to provide publicly available guidance on selection of appropriate measures of attainment in each subject (literacy, mathematics, and science), paired with accessible summaries about the range and nature of the tests available. Selected questions from the European Federation of Psychologists' Associations (EFPA) test review model for the description and evaluation of psychological and educational tests (Evers, Hagemeister, et al., 2013) are used to evaluate tests. This information is also summarised within a searchable database. This written synthesis outlines the systematic review methodology used to form the database.

The database is somewhat comparable to the aforementioned Early Years measures database but includes additional information and filters. A rating system based on the psychometric properties of the test will transparently indicate the quality of each test. In contrast to the system applied to the Early Years database, implementation factors will not be rated. Instead, information about implementation will be provided as filters to sort the database and shortlist measures that match the users' needs. Given that the audience for the database is diverse (including teachers, evaluators, and researchers) this is important, ensuring that implementation factors are considered a preference and are not misinterpreted as relating to the quality of a test.

Objectives

This review provides much-needed guidance to support the selection of measures of attainment in literacy, mathematics, and science. Our approach focuses on tests of particular relevance to educators and evaluators in the U.K. who wish to measure the attainment of children and adolescents aged 6 to 18 years. The evidence is summarised here in a written synthesis and also presented in a searchable database.

The research questions are:

1. How can teachers and evaluators assess attainment and progress in literacy, mathematics, and science in the U.K.?
2. What is the psychometric quality and implementation utility of the tests identified through this review for use with pupils aged 6 to 18 years old?

This written synthesis begins by describing the search results of the systematic search protocol. The systematic review methodology is presented in Appendix 1: Methodology (see also Breadmore and Carroll, 2020). Search results are summarised and presented in Appendix 2: PRISMA flow diagram. The results are organised by the research questions. In each case, we begin with definitions of key terminology used in the review such as the definition of attainment, how to evaluate the psychometric properties (reliability, validity, standardisation process, and the nature of norms), and how to interpret this information when selecting tests. The results of the review are then presented in terms of summaries of all of the tests subjected to evaluation, as well as descriptive summaries such as the proportion of tests rated as having 4*, 3*, 2*, 1*, or 0* psychometric properties, as well as the identification of gaps in the availability of tests. Finally, we discuss the implications and limitations of this review.

Search results

Search results are presented in Appendix 2: PRISMA flow diagram. There were two elements to the search process: (1) test identification and (2) publication identification.

Test identification

The test identification phase established the long-list of tests of attainment for consideration. The methods used in test identification are described in detail in Appendix 1. Search criteria included an initial screen to ensure that measure were:

- used to assess literacy, mathematics, or science attainment;
- published in or since 2000 (see also Denman et al., 2017); and
- suitable for English-speaking 6- to 18-year-olds.

The long-list did not include tests of underlying abilities; however, if there was any uncertainty about the construct measured within a test, it was initially included in the long-list and later excluded during screening and eligibility checks (306 records for tests that were initially identified were later excluded). Two hundred and seventy-three relevant tests were identified by hand-searching websites (see Table **11** in Appendix 1 for a complete list of websites), 35 were identified by the advisory panel, 17 from a list of tests used as outcome measures in EEF trials, 31 from personal communication with the EEF, 4 were later added to the long-list through an iterative process (identified during later literature searches or because of recommendations from publishers when gathering materials), and 8 tests were identified from a search of the ERIC database (search terms are presented in Appendix 1: Methodology and Search Terms). Minimal information was recorded about all tests that fulfilled initial search criteria (see Table **1**).

Table 1: Basic information recorded for all tests on the long-list

Criterion	Minimal information to include in the database	Exclusion criteria
Basic test information.	Name of test. Current version/edition number. Name and acronym of previous/original version(s) of the test (if applicable). Subject (literacy, mathematics, science, or generic).[1]	Does not meet search criteria.

[1] Generic tests are included only if they have subtest(s) that measure literacy, mathematics, or science.

After removing duplicates, 231 tests were identified for inclusion in the long-list database of tests (see Supplementary Materials 1). In line with the recommendations from the EFPA test review model (Evers, Hagemeister, et al., 2013), information about the tests was sourced from publisher's websites, marketing materials, or personal communication with publishers or authors. In some cases, test manuals were also consulted for clarification at this stage (this was sometimes necessary to clarify whether an test measured language or literacy, or had criterion- or norm-referenced scores). Initially, minimal information was gathered to screen tests using the exclusion criteria reported in Table 1 and then Table 2 (see also Appendix 1: Methodology and Search Terms). After screening, 41 tests were subject to evaluation.

Table 2: Screening criteria for tests, minimal additional test information, and summary exclusion criteria included in the database

Criterion	Minimal information to include in the database	Exclusion criteria
Basic test information (additional information added during screening).	List of subscales (if applicable). Additional references/hyperlinks for other sources of information about the test (e.g., supplementary norms, academic peer-reviewed publications, as applicable). Brief description of test using content from publisher website (if available).	
Availability of administration guidelines and scoring criteria.	Authors. Publisher. Hyperlink for source of test.* Administration guidelines not available.	Test is not available for review.
Norm-referenced scores.		Criterion-referenced.
Suitable for target sample (6 to 18 years).	Specific population and age range that publisher states the test is intended/suitable for. Key Stage(s) applicable to.	Test is not applicable to sample.
U.K. standardisation sample.	Yes/No.	No U.K. standardisation available.
Published or re-normed since 2010.	Publication date. Date of re-norming (if applicable).	No recent norms available.

Criteria highlighted in **bold** are new exclusionary criteria. Note that 'test is not applicable to sample' is not redundant with the inclusion criteria 'suitable for English-speaking 6- to 18-year-olds' because inclusion criteria were applied leniently to maximise the number of tests included in the initial screening phase. Tests excluded on this basis included those designed for use with clinical populations.
* The hyperlink enables users to obtain additional information that may change over time, such as the cost of materials required for administration.

Publication identification

The next phase was to identify sources of information to enable evaluation of the 41 shortlisted tests—the publication identification phase. Administration and technical manuals were obtained from publishers, along with any grey literature—other meaningful sources of information that were provided by publishers that might assist our evaluation. Grey literature obtained from publishers was only included in the review if the sources were published with intention to be available to the public (for example, available on the publisher's website or distributed with test materials) or was from peer-reviewed publications. Grey literature intended for private distribution, such as internal reports, were not included in the review. In total, 130 records that were received from publishers contributed to the review process. This information was further supplemented by the results of systematic searches of the ERIC and PsycInfo databases, which enabled us to identify peer-reviewed evidence of the psychometric properties of tests (see Table 12 in Appendix 1: Methodology and Search Terms). Five hundred and fifteen peer-reviewed publications were initially identified from systematic searches. After removing duplicates and screening publications (see Table 13 in Appendix 1: Methodology and Search Terms for screening criteria), only four peer-reviewed publications obtained through these systematic searches contributed to the review.

In the final phase of identifying tests for evaluation, all sources were reviewed for information that could contribute to the evaluation. This information was combined across multiple sources for each test and then eligibility checks were conducted to consider whether there was enough information available from these sources to justify further evaluation (see Table **3** and Appendix 1: Methodology and Search Terms for further information about eligibility criteria). In the shortlisted database, 37 out of 41 tests fulfilled these eligibility criteria and were subjected to full evaluation; four tests were not eligible because the information provided by publishers and sources from peer-reviewed publications was insufficient to evaluate validity and reliability.

Table 3: Test eligibility criteria

Criterion	Minimal information to include in the database	Response options	Exclusion criteria
At least one measure of construct or criterion validity.	Validity measures available?	Yes/No.	No measure(s) of validity available.
	Measure(s) indicated and the value provided.	Construct validity, structural validity, internal structure, item construct validity, concurrent validity, convergent validity, predictive validity, discriminant validity, contrasted groups validity, identification accuracy, diagnostic accuracy, cross-cultural validity, criterion validity.	
	Source(s) of validity measures	Free text	
At least one measure of reliability.	Reliability measures available?	Yes/No.	No measure(s) of reliability available.
	Measure(s) indicated and the value provided (e.g., Pearson's r =, Cronbach's α =, Cohen's κ =).	internal consistency/reliability, content sampling, convention item analysis, inter-rater/scorer reliability, intra-rater/scorer reliability, test-retest reliability, temporal stability, time sampling, parallel forms reliability, measurement error, standard error of measurement, smallest detectable change, limits of agreement.	
	Source(s) of reliability measures	Free text	

Results of review

How can teachers and evaluators assess attainment and progress in literacy, mathematics, and science in the U.K.?

Defining key constructs

Before considering which measures of attainment are available, we needed to define what we mean by 'attainment' and 'progress' in each subject area. To do so, we consulted recent evidence reviews commissioned by the EEF on literacy, mathematics, and science (Breadmore, Vardy, Cunningham, Kwok, and Carroll, 2019; Hodgen, Foster, Marks, and Brown, 2018; Nunes et al., 2017), national and international policy documents (DfE, 2014; OECD, 2019), and worked with key experts in each field and in assessment more broadly through our advisory panel (see Table **10** on page 35). It is clear from all of these sources that attainment in all three subject areas is dependent upon multi-faceted sources of knowledge and skills.

Definition of attainment

Attainment and progress can be defined in terms of age-related expectations based on the curriculum. In England, statutory programmes of study and attainment targets are set out in the National Curriculum in England (DfE, 2014). The national curriculum specifies that numeracy and mathematics and language and literacy must be developed through all relevant subjects. Programmes of study in English, mathematics, and science are particularly detailed throughout the four key stages of compulsory education. Furthermore, statutory assessments (SATs or national curriculum assessments) measure attainment in these core subjects. Performance on statutory assessments reflects the attainment and progress of individual pupils and results are used by the DfE for school accountability purposes (DfE, 2019b, 2020). As such, statutory assessments provide criterion measures of attainment for both pupils and schools. If the purpose of measurement is to predict attainment, the important consideration is alignment to the national curriculum and statutory assessments.

Our aim with this review was to evaluate measures of attainment in literacy, mathematics, and science. At this point, we need to draw a distinction between *academic attainments***,** *specific skills***,** and *cognitive abilities*. Academic attainment could be defined as performance in skills taught and measured by the curriculum. As such, academic attainment can be considered outward measurable behaviour.

Academic attainments depend upon a wide variety of factors including the teaching a child receives, their wider environment, their underlying cognitive abilities, and their specific skills. The skills and abilities that are considered key in different subject areas have been determined and described in academic research and theory (summarised in Breadmore et al., 2019; Hodgen et al., 2018; Nunes et al., 2017). Some abilities are domain general (such as working memory and reasoning) and impact on attainment across core subjects. Other abilities are subject specific (for example, number sense). Specific skills may be learnt or innate, but form some component of a broader attainment measure. This could include, for example, decoding nonsense words, handwriting, remembering times tables or facts, or knowing the periodic table.

As a concrete example, in 2020 attainment in Key Stage 2 was planned to be assessed at the end of Year 6 through a series of national curriculum tests in maths and English (note that these tests were cancelled as a result of the COVID-19 pandemic). The English grammar, punctuation, and spelling test was comprised of two papers focused on the relevant elements of the English Programme of Study. The English reading test focuses on the comprehension elements of the national curriculum and includes a mixture of text types (DfE, 2019a). Performance on the Key Stage 2 reading test is a valuable criterion measure of attainment in literacy, which depends upon the combination of a range of underlying skills including word reading, vocabulary and language skills, reading fluency, drawing upon prior knowledge, building coherent mental models, literal and inferential comprehension, and comprehension monitoring as well as more distal abilities such as executive functions (see Breadmore et al., 2019). However, the Key Stage 2 reading test does not directly measure these individual skills and abilities. If the purpose for measuring attainment is to predict performance in Key Stage 2 reading, then a good measure would be a reading comprehension task with a similar design. However, such a measure is unlikely to provide diagnostic information about why a child is struggling—children with word reading

or comprehension difficulties may perform similarly on the measure but would need quite different interventions to improve attainment.

Understanding the profile of underlying skills and abilities is very important for understanding the causes of low attainment, identifying individuals at risk of low attainment, understanding where to target teaching resources to raise attainment, or considering how or why an intervention is effective. However, searching for measures of each underlying ability is beyond the scope of this review.

In all cases we aimed to focus on measures of attainment rather than cognitive abilities or specific skills. However, this was not always straightforward and some readers may disagree with our decisions. For example, we decided to exclude a measure of handwriting speed because we felt it was a measure of a specific skill rather than attainment, and we decided to exclude the Cognitive Abilities Test 4 (GL Assessment, 2020) as it is sold as a measure of underlying ability rather than attainment. However, we have included measures of word reading, which some might regard as a specific skill. In general, we have aimed to be inclusive rather than exclusive while following our protocol as closely as possible.

Literacy, mathematics, and science are complex multi-dimensional subjects and the key constructs of knowledge that determine attainment change over the course of development as different skills develop and subject knowledge is taught. Here, we seek to evaluate measures of overall attainment in each subject. Nonetheless, consideration will be given to whether tests are specific or general measures of attainment as this has relevance for structural validity. Specific tests measure only one key concept or area of content knowledge. For example, a spelling test would be a specific measure of literacy attainment while an arithmetic test would be a specific measure of mathematics attainment (with potential alignment to Key Stage 2 criterion assessments). General measures of attainment are multi-dimensional, assessing more than one key concept or area of content knowledge.

Norm-referenced tests versus criterion-referenced tests

Tests of attainment come in two main formats. Norm-referenced tests allow you to compare the performance of a given child to the range of attainments for children of that age or that school year. Criterion-referenced tests, on the other hand, give information as to whether a child understands particular concepts or can use particular methods or approaches. The national tests for Key Stages 1 and 2 are criterion-referenced in that there are certain criteria which must be met in order to judge that a child has met the 'expected' level for their age. This allows the Department for Education to set objective criteria of what is expected at each age, and then to state what proportion of children meet these criteria in different schools, or from different ethnic groups, and so on. This in turn allows comparison over time to assess whether standards are improving. It is important to note that 'scaled scores' used to describe performance on these tests are not based on normative data (see page 20 for further discussion).

Although curricula always aim to develop skills and abilities, curriculum and statutory assessment content is also subject to pedagogic and political decision-making. As a result, the content of criterion measures of attainment will vary between different political jurisdictions and change over time. For example, Roberts (2020) describes 16 key developments in government reforms to Key Stage 1 and Key Stage 2 SATs in England from 2010 to 2020. In many ways, the criteria for 'expected levels' have increased in recent years.

GCSE grades, instead, are marked using a type of norm referencing: overall scores are ranked and grade boundaries are set so that similar proportions of students get each grade each year, regardless of the difficulty of the particular exam set. This norming process means that scores are consistent even if students get a hard exam or an easy exam.

When examining attainment in terms of progress and identifying potential difficulties, norm-referenced tests are vital because they allow a statistical comparison with other children or other time-points, and they give an indication of how much progress is to be expected over time. They avoid the variability due to changes in curricula that can make criterion-referenced tests difficult to compare. This is why we have chosen to focus only on evaluating norm-referenced tests in this review. The exclusion of criterion-referenced tests from the evaluation phase of this review should not be misinterpreted as suggesting that norm-referenced tests are necessarily better. Criterion-referenced tests are also essential guides to understanding what has been learned. However, these two types of tests cannot be evaluated along the same benchmarks and for this reason criterion-referenced tests were included in the long-list but were not evaluated here.

Changes in norms over time

One might assume that standard scores remain relatively constant over time, but in fact they vary in complex ways. Scores on tests of intellectual ability, such as nonverbal reasoning, actually show rises over time with cohorts scoring higher than similar cohorts ten or twenty years ago. This is known as the Flynn effect (Flynn, 2012). Scores on verbal tests tend to show less variation with time, but still increase by a few points each generation.

Changes with scores in attainment tests over time have not been as closely studied, but one would expect that as curriculum demands change and underlying abilities change, attainments will also change. This is also shown in practice: for example, the technical manual of the British Ability Scales 3 demonstrates how a ten-year-old obtaining a raw score of 72 on the same reading test would achieve a standard score of 113 on the 1996 BAS II standardisation and a standard score of 100 on the BAS 3 standardisation in 2011 (Elliott and Smith, 2011).

THE EFPA test review model recommends that when evaluating normative studies, norms over 20 years old should be considered inadequate, norms between 15 and 19 years old are adequate, norms between ten and 14 years are good, and less than ten years old are excellent (Evers, Hagemeister, et al., 2013). In this review, to allow for publication lag both in terms of the sources of information about normative studies and of the measures database itself, the long-list database includes tests published since 2000, but we only evaluated tests that have been published or re-normed since 2010. This is to avoid recommending tests in which both the norms and test-content are likely to be significantly out of date.

Description of the evidence base

The long-list database includes 231 tests that measure attainment in literacy, mathematics, or science. Those tests are included in the long-list database and listed in Supplementary Materials 1. However, after screening and checking for eligibility, only 37 were found to be relevant for evaluation:

- 11 were not available to review;[1]
- 94 were criterion-referenced (or did not have norms at all);
- 3 were not applicable to the sample (for example, were intended for use with clinical populations);
- 50 had not been standardised on a U.K. sample;
- 32 did not have recent U.K. norms (published since 2010) available; and
- 4 were initially shortlisted, but were excluded from the evaluation because there was insufficient information available to the review team to evaluate validity and reliability.

We were able to gather enough information on validity and reliability to evaluate the remaining 37 tests.

Findings and gaps in the evidence base

The reasons for screening or determining a test as ineligible for evaluation reveals some interesting findings about trends in test design. For example, most of the tests that were found were not norm-referenced. Another notable point is that the requirement for recent, local (U.K.) norms excluded 82 norm-referenced tests from the evaluation. This included some measures that appear of high quality. Therefore, publishers and test developers should be encouraged to resolve this by conducting re-standardisation trials.

Further, there are signs of variation in the nature of tests in different subjects. Examining the tests that were included in the evaluation phase reveals some important differences between subjects (see Table 4: Summary of the number of tests eligible for evaluation by subject area and suitability to different age ranges). Educators and evaluators have a good deal of choice when selecting measures of attainment in literacy, particularly in primary aged pupils, but very little choice when selecting measures of attainment in science, where only one measure was found that fulfilled criteria to be considered in the evaluation. This is an important and noticeable gap, which test developers should look towards resolving.

[1] Because the test was not available to purchase in the U.K. or was out of print.

Table 4: Summary of the number of tests eligible for evaluation by subject area and suitability to different age ranges

Measures of attainment in…	Suitable for primary aged pupils	Suitable for secondary aged pupils
Literacy	18	15
Mathematics	16	9
Science	1	1

Note: The sum of cells is greater than the total number of tests eligible for review because some tests fall into more than one age group (for example, tests that are suitable for assessing both secondary and primary aged pupils, or that include measures of both literacy and mathematics).

What is the psychometric quality and implementation utility of the tests identified through this review for use with pupils aged 6 to 18 years old?

Before discussing the psychometric quality and implementation utility of the tests that entered the evaluation phase, we must first define the terminology and concepts used in this element of the systematic review. In the evaluation phase, we relied heavily on recommendations from the EFPA test review model (Evers, Hagemeister, et al., 2013) and combined information from multiple sources using methodology aligned to the COSMIN risk of bias checklist (Mokkink, de Vet, et al., 2018). The procedures used to summarise and evaluate these factors are described in Appendix 1: Methodology and Search Terms. Table **5** describes the implementation factors and Table **6** describes the psychometric properties that are summarised in the shortlisted database.

The EFPA review model was developed by the European Federation of Psychologists' Associations Board of Assessment (http://www.efpa.eu/professional-development/assessment) to support the description and evaluation of psychological and educational tests. This review model informed inclusion and exclusions criteria in this review, and selected questions from 'Part 2: Evaluation of the Instrument' were used to evaluate tests in the final stage of the review (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013). Further detail about how this review model was applied is provided in the Appendix 1.

Definition of implementation utility

'Implementation utility' refers to how easily a test can be used. It is subjective and dependent on a multitude of factors including the purpose for the assessment, availability of resources (including facilities, money, and time), the child(ren) being assessed, and the tester. As such, implementation factors were summarised in the evaluation phase but were not rated. This will enable individual users of the database to consider information that is relevant to their decision-making when selecting a test, and to ignore irrelevant information. The implementation factors evaluated were selected in consultation with our advisory panel and largely align to Part 1 of the EFPA test review model section 'Description of the instrument' (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013). The additional information included in the database is summarised in Table **5** (see Appendix 1: Methodology and Search Terms for further detail). Most of this terminology should be self-explanatory. Below we provide some definitions for terminology that might be less familiar to users of the database. For example, one of the response format options is 'manual (physical) operations', which is used to categorise responses such as pointing, manipulating blocks, and so forth. The sources used for the evaluation (administration and technical manuals and other meaningful sources of information provided by publishers) are described in 'publication identification' (see page 42).

Table 5: Evaluation stage—additional implementation information included in the database (not rated)

Criterion	Minimal information to include in the database	Response options
Basic test information (added during evaluation).	Note whether additional versions are available (e.g., short/long versions, and which is subject to review).	Free text
	Note whether subtests can be administered in isolation (if applicable).	Free text
Administration format.	Administration group size.	Individual/small group/whole class
	Administration duration.	Total time in minutes
	Description of materials needed to administer test	Free text (e.g., user manual, licence, computer, internet access, headphones, digital recorder, etc.)
	Any special testing conditions?	Free text
Response format.	Response mode.	Oral/paper and pencil/manual (physical) operations/electronic*
	* If electronic, what device is required	Free text (e.g., computer, tablet)
	Question format.	Multiple choice/open ended/mixed
	Progress through questions.	Adaptive/flat
Assessor requirements.	Is prior knowledge/training/profession accreditation required for administration?	Yes*/No/Not stated
	* If yes, what is required? Where possible, distinguish between requirements for administration and scoring.	Free text
	Is administration scripted?	Yes/No
Scoring.	Description of materials needed to score test	E.g., user manual, supplementary norms
	Types and range of available scores	Raw/centiles/deciles/z-scores/standard scores/stens/Stanines/T-scores/other (specify)
	Score transformation for standard score.	Not applicable (no standard scores available)/not-normalised/age standardised/grade standardised/other (specify)
	Age bands used for norming.	E.g., 3 months, 1 year

Criterion	Minimal information to include in the database	Response options
	Scoring procedures	Computer scoring with machine readable paper forms/computer scoring with direct entry by test taker/computer scoring with manual entry of responses from paper form/simple manual scoring key – clerical skills required/complex manual scoring – training required/bureau service (scored by publisher/distributor)/other (describe)
	Automatized norming	None/machine readable/computerised/online/bureau service

Adaptive or flat

Progress through questions in a test may be flat or adaptive. 'Flat' progress means that every examinee taking the same test answers the same questions. 'Adaptive' tests present different items to different examinees, the goal being to reduce testing time and frustration by minimising exposure to questions that are too easy or too hard for the examinee. Often, adaptive tests will be constructed of a series of blocks with accuracy across those blocks determining whether the examinee will receive easier or harder items. These tests have a rule that determines the basal (easiest) and ceiling (hardest) blocks. Alternatively, some computer-based tests can select easier or harder questions in an algorithmic process based on performance on the earlier items to allow targeted testing of a child's attainment.

Assessor requirements

The International Test Commission Guidelines for Test Use (International Test Commission, 2001) highlight the knowledge, understanding, and skills that are essential competencies for all test users. However, some test administration manuals further specify that users should have a certain level of prior knowledge, qualifications, or professional accreditation in order to administer, score, or interpret a test. If this is the case, it should be clearly indicated by the publisher prior to purchase of a test and described in the administration manual for the test. This requirement might be related to the specific skills required to administer certain subtests, to score the test, or to interpret the results in an assessment. It can be the case that different levels of skill are required for different elements of administration and interpretation. Some manuals are explicit about this and others are not.

When considering assessor requirements, it is important not only to review the information presented in the administration manual but also to also think about the purpose of testing. Additional user qualifications are often imposed for high stakes testing. For example, evidence requirements for applications for Disabled Student Allowances are determined by the Department for Education. These currently include the requirement that diagnostic assessments are conducted by a specialist teacher assessor with an Advanced Practicing Certificate or a practitioner psychologist registered with the Health and Care Professions Council. On the other hand, more lenience may be appropriate in low stakes testing where aggregate or group level data will be considered or where conduct will be overseen by an advanced assessor such as a chartered psychologist. This is commonly the case when assessments or subtests are used for research purposes (International Test Commission, 2014).

Types of scores

Test manuals should describe how the measures were calculated and offer information on how to interpret results. During the evaluation process for this review, we noticed wide variation in the level of information provided to test users to guide their interpretation of these measures. Below we provide a general guide to interpretation of the scores that are most commonly used.

Raw scores and weighted raw scores

A child's **raw score** is simply the sum of their correct responses or points awarded on a test or subscale. It can be informative to look at which particular items a child made errors on as well as the nature of those errors. However, raw scores have limited interpretability. For example, if subtests have different numbers of items or vary in difficulty then raw scores cannot be meaningfully compared, or if the test is adaptive different participants will have completed different items and hence raw scores are not comparable. **Weighted raw scores** account for differences between items sets, and so could be used to compare children of the same age who complete different items (Wechsler, 2018). However, derived scores (such as standard scores, scaled scores, stanines, or T-scores) should be used to make comparisons or aggregations of performance of children of different ages or on different item sets. These scores are described below. Criterion-referenced tests do not include these derived scores and hence were screened from the review.

Standard and scaled scores

Raw scores on tests of attainment are influenced both by age and experience—for example, older children are likely to gain higher scores. **Standard scores** offer the most precise description of performance allowing the examiner to compare children of different ages and performance on different tests by comparing the child's performance compared to a normative sample. The average (mean or median depending on standardisation procedures) performance of the

normative sample is converted to a standardised score of 100 with a standard deviation of 15. Hence, the reliability and validity of standard scores depends upon the representativeness of the norm-derived sample. Assuming that the norm-derived sample is free from bias, by converting a raw score to a standard score an examiner can therefore consider performance against scores obtained by children of the same age or cohort: a standard score of 100 indicates average performance, 85 indicates performance 1SD below the mean, 70 indicates performance 2SD below the mean; 68% of children of the same age would obtain standard scores between 85 and 115, and so this is typically considered the normal range of ability. Meanwhile, 96% of children of the same age would obtain a standard score between 70 and 130 (Wechsler, 2018). Descriptive classifications will typically assign the following qualitative descriptors to standard scores:

- 69 and below: low;
- 70–84: below average;
- 85–115: average;
- 116–130: above average; and
- 131 and above: high.

Age standardised scores enable the examiner to consider how performance of one chid compares to their same-age peers (Wechsler, 2018). **Year or cohort standardised scores** enable comparison against other children in their cohort but do not account for age (Ruttle et al., 2020). When looking at a single class at a single point in time, these two types of standardisation will be very similar, but there are situations in which one is more useful than the other. Year or grade standardised scores are particularly helpful when considering aggregate scores within a year group, for example, to look at overall class progress over time. For measures that span multiple years or underlying abilities, age is likely to be the strongest predictor of performance and so age-standardised scores are invaluable. However, the strongest predictor of performance for tests that measure progress in learning the curriculum will be experience (Cunningham and Carroll, 2011). Hence, in some circumstances year or cohort standardised scores can be more informative (Ruttle et al., 2020).

The reliability and precision of standardised scores depend on the quality of the normative sample, hence we evaluated whether there were any risks of bias in the normative sampling. Possible risks of bias include normative samples that are unlikely to be nationally representative, for example, because of recruitment techniques, because data was collected from a small number of schools or a limited geographic area, or by using sampling procedures that otherwise fail to adequately represent the socio-economic and cultural diversity of the general population. Normative data collected at a single time-point either towards the beginning or end of the academic year can impact on generalisability if the test is used at other time-points. Other considerations that reflect sensitivity include the range of standardised scores and the size of the age band used to convert raw scores to standard scores. Note, however, that because development and attainment are usually nonlinear and begin to plateau in older ages, it is common and nonproblematic for intervals to increase with age.

Standardised tests sometimes provide alternate derived scores such as **scaled scores**, **stanines**, or **T-scores**. These scores are similar to standard scores but use alternate ranges: stanines, for example, are on a nine-point scale with a mean of five and standard deviation of two. The use of stanines is recommended in circumstances where the user wishes to prevent overinterpretation of small differences in scores (Wechsler, 2018).

It is important to note that derived scores on a norm-referenced test are quite different from the **scaled scores** as defined by the Standards and Testing Agency on behalf of the Department for Education. A scaled score of 100 or greater on Key Stage 2 national curriculum tests indicates that a child has performed at or above the ‘expected standard’ on the test (Standards and Testing Agency, 2020). This is not based on normative data or mean performance of a sample, but rather on an a priori judgement of what standard should be expected. Scaled scores range from 80 to 120 and a scaled score of 99 or below indicates that a child has not met the expected standard (Standards and Testing Agency, 2020). In contrast, a standard score of 100 on a norm-referenced test indicates that a child has average performance, and a standard score of 99 would be considered well within the normal range of performance.

Percentile ranks

Percentile ranks (also sometimes called percentiles or centiles) are closely linked to standard scores but reflect the percentage of the population that would be expected to perform at or below that of the examinee. Average performance

(that is, a standard score of 100) would achieve a percentile rank of 50 indicating that this examinee performed as well or better than 50% of the normative sample of children of the same age (Wechsler, 2018). While percentile ranks appear easily understood, care should be taken to ensure that they are not misinterpreted as percentage accuracy, and it should be noted that the nature of rank ordering means that the differences between scores are not equivalent. The difference in ability between pupils on the 50th and 60th percentiles is not the same as the difference in ability between pupils on the 10th and 20th percentiles. As a result, percentiles cannot be meaningfully aggregated, added, or subtracted (Connolly, 2013; Wechsler, 2018).

Age equivalents

Age equivalents reflect the age at which an average student performs as well as the examinee (Wechsler, 2018). Age equivalents have limitations as indicators of ability and should be used with caution as they have limited interpretability (Connolly, 2013; Elliott and Smith, 2011; Wechsler, 2018). They may give an indication of a child's instructional level but do not imply the level of the curriculum that the examinee should receive next. For example, if an eight-year-old achieves a reading age of 13 years old in a reading comprehension test, this means that they got the same number of correct responses as the average 13-year-old on this particular test (Wechsler, 2018). They won't necessarily have the requisite world knowledge to understand or enjoy texts intended for 13-year-olds. Like percentiles, age equivalents are not on an equal-interval scale. Furthermore, age equivalent typically has a non-linear relationship with raw score. This is because development is rarely linear and typically plateaus in older age ranges. Age equivalents should not be aggregated to look at group differences or added or subtracted to consider longitudinal trajectories (Connolly, 2013).

Confidence intervals and standard error of measurement

All measures of attainment are based on observations on the day of the assessment and should correctly be considered an estimate of the examinee's true level of attainment combined with some degree of measurement error. **Confidence intervals**, **standard error of measurement**, and **standard deviations** provide an indication of the reliability of the measure and should be reported where possible as a reminder of the measurement error inherent to all assessments (Wechsler, 2018).

Behavioural observations during testing are also essential to interpretation of test scores. In order to interpret and communicate the results of a test appropriately, due consideration should be given to whether the examinee became distracted, demotivated, or rushed (Connolly, 2013; Wechsler, 2018). This can also be very informative when considering next steps. Tests often allow examiners to note behavioural observations on response sheets, which may reveal something about the validity of the test results in an assessment. However, behavioural observations do not usually alter test scores.

Definitions for evaluating the psychometric properties of tests

Whereas implementation utility is difficult to evaluate objectively due to variation in user preferences, many of the psychometric properties of tests can be evaluated objectively. This evaluation crucially depends on consideration of validity and reliability of the test results as well as the quality of norms. The International Test Commission Guidelines on Test Use highlights the importance that this

> 'evidence should be accessible to the test user and available for independent scrutiny and evaluation. Where important evidence is contained in technical reports that are difficult to access, fully referenced synopses should be provided by the test distributor.' *(International Test Commission, 2001, p. 96.)*

For this reason, we only reviewed evidence relating to validity and reliability if it was available to the test user and did not include information from grey literature unless it was distributed with test materials or available on the publisher's website. The evaluation protocol enabled us to supplement this with information from peer-reviewed publications, which arguably are not readily available to the typical test user. However, in reality very little useful information was gathered from these sources (only four peer-reviewed publications contributed to the evaluation). We strongly encourage publishers to summarise information about the reliability and validity of tests on their websites. We note that although this information is not typically summarised on U.K. publishers' websites, it is commonly available on U.S. publishers' websites (for example, https://www.parinc.com/, https://www.proedinc.com/), presumably as a result of adherence to the U.S. Standards in Educational and Psychological Testing (The Standards, 2014). Table 6 indicates the summary

information used to evaluate the psychometric properties of the tests that were included in the shortlist. In the following sections we define different types of validity and reliability and discuss what was considered during the evaluation of the quality of norms. This will support interpretation of the information summarised in the shortlisted measures database, as well as providing key considerations to keep in mind during test selection.

Note that the more high stakes a test is, the more important the psychometric properties of that test as they give an indication of how likely the test results are to be a true reflection of a student's underlying abilities.

Table 6: Evaluation stage—evaluation of psychometric properties on a four-star scale

Criterion	Minimal information to include in the database	Response options	Exclusion criteria/Rating
Construct validity	Does it adequately measure literacy, mathematics, or science?	0–4	Overall construct validity score ≥ 3/4 = Star
	Does it reflect the multidimensionality of the subject? Is it a generic (e.g., literacy) or specific (e.g., word reading) assessment of attainment?	Generic/specific	
	Summary of information available on construct validity (and reference for source).	Free text	
Criterion validity	Predictive/concurrent/postdictive validity: Does test performance adequately correlate with later, current, or past performance?	0–4	Overall criterion validity score ≥ 3/4 = Star
	Summary of available comparisons (e.g., the measures compared to assess concurrent validity) and correlation (value(s) reported, with citation).	Free text	
Reliability	Is test performance reliable?	0–4	Overall reliability score ≥ 3/4 = Star
	Summary of available comparisons (e.g., specify the types of reliability measured and how) and value(s) reported, with citation.	Free text	
Is the norm-derived population appropriate and free from bias?	Is population appropriate and free from bias?	Yes/No*	Yes = Star
	* If any biases are noted in sampling, these will be indicated here.	Free text	

Validity

Validity describes how well a test serves its intended purpose—how well it measures what it intends to and results in accurate interpretation of results. As a result, the validity of the results of a test depend not only on test properties, but also on the decisions of the user and how they intend to apply and interpret the results. Results will only be valid if users have selected the correct measure for their needs. It is the interpretation of scores that can be validated, not the test itself (The Standards, 2014). Therefore, although we provide an evaluation of validity for the purposes of this review, users of the database must be alerted to their own responsibility in determining validity, which depends upon their interpretation of test scores for their intended purpose.

There are numerous different types of validity that have been described in the literature. Here, in line with the recommendations from the EFPA test review model (Evers, Hagemeister, et al., 2013), we focus on evaluating construct and criterion validity, which are defined below. Table 6 indicates the summary information that we included in the measures database and how scores were assigned (see also Appendix 1: Methodology and Search Terms). The sources used to inform the evaluation of validity included administration and technical manuals, grey literature provided by publishers, and peer-reviewed publications (described in further detail in Appendix 1, 'publication identification', see page 42). We acknowledge that validity is now typically viewed as a unitary concept using all available evidence to consider the extent to which test score interpretation is appropriate for the intended use (The Standards, 2014). Nonetheless, we felt it was important to distinguish between evidence for construct and criterion validity because construct validity is more subjective and likely to be influenced by the aims of the user.

Construct validity

Construct validity examines the extent to which the test actually measures what it sets out to measure or instead partially or mainly measures something else. The first step in determining construct validity is to consider what the target of measurement is. Care must be taken to distinguish between our own target construct (attainment) and the construct identified by the authors of a test. Another challenge for measuring attainment is that the target construct can change when the national curriculum changes. This can present a problem for measuring attainment over time, which can also limit the criterion validity of a test. We return to this discussion later.

Tests vary in terms of the extent to which the target of measurement has been identified, explained, and justified. It is rare that only one possible meaning could be interpreted from test responses so it is essential that test developers carefully specify the intended scope of the construct of measurement (The Standards, 2014). The EFPA test review model encourages consideration of the quality of the explanation of the rationale in relation to theoretical foundations and test development procedures in addition to statistical evidence of validity (Evers, Hagemeister, et al., 2013). Hence, in the shortlisted measures database we evaluated construct validity using information reported in all available sources combined with our subjective assessment of how well that target construct adequately measured general attainment in literacy, mathematics, or science. In addition to our subjective rating of construct validity in response to the question, 'Does it adequately measure literacy, mathematics or science?', we also provide information on structural validity in response to the question, 'Does it reflect the multidimensionality of the subject?' Further, we provide a summary of evidence relating to the construct validity of the tests as described by the test developers, with citations to sources. Before describing how ratings were assigned, we will first define different aspects of construct validity and what can be measured statistically.

Face or content validity is assessed subjectively by considering how well the test can reasonably be expected to measure the target construct described by the test developers. To some extent, it is easier to assess construct validity for attainment measures than it is for ability measures in that attainment refers to the outward behaviour shown rather than an underlying skill. Hence, many of our measures highlighted that they tested skills very similar to those required on the national curriculum in order to demonstrate face or content validity. Face or content validity is necessary but not sufficient for high construct validity: it should also be accompanied by adequate statistical evidence to support construct validity (described below).

Structural validity is an aspect of construct validity that refers to the extent to which the test adequately measures the underlying structure of the target construct. Since we assume literacy, mathematics, and science to be multidimensional constructs, measures of attainment with high structural validity would reflect this multidimensionality. For example, a test of literacy attainment that included subscales for word reading, reading comprehension, spelling, and writing

composition would be considered to have higher structural validity than a literacy test that purely measured spelling. However, a different research question would alter the structural validity of the test. For example, to evaluate the effectiveness of a spelling intervention a spelling measure would have high structural validity because it measures the specific literacy skill that was treated in the intervention (spelling). In contrast, a general multidimensional measure of literacy attainment would be less valid if applied to the question of whether the spelling intervention had a positive impact on spelling. On the other hand, the multidimensional measure of literacy attainment would have high validity if the research question is whether the spelling intervention impacts on literacy attainment more broadly. Hence, we provide descriptive information about structural validity in response to the question, 'Does it reflect the multidimensionality of the subject?', with the response options 'generic/specific', rather than by providing a rating. Structural validity can be measured statistically, for example, by using exploratory and confirmatory factor analyses (Evers, Hagemeister, et al., 2013).

Construct validity more broadly is not only determined by test design. It is essential that users think about how well the target construct of the test aligns to their own construct of interest. For example, if users want to measure attainment in terms of progress on the curriculum, they need to select a test which similarly has this as the target of measurement. If not, their interpretation of the data will not have construct validity regardless of what is reported in the test manual.

Statistical measurement of construct validity

Having determined the target of measurement, we can then consider the extent to which items in the test are effective in measuring that target construct. This can be measured statistically. For example, we can consider whether test scores and profiles change in line with expected patterns of developmental progression. Often, this is achieved by correlating test performance with age and reporting the strength of that correlation. However, these correlations can be difficult to interpret because, as previously discussed, progress in attainment is rarely linear throughout development. Similarly, if a test is intended to be used at regular intervals and has been standardised accordingly (for example, termly assessments), normative data should be collected at each intended time of assessment and therefore age will not be a strong predictor of attainment (Ruttle et al., 2020). In the shortlisted database we present a summary of the nature of the statistical evidence available to support the construct validity of each test. This includes a qualitative descriptor of the strength of this evidence, which also influenced our evaluation of construct validity. In addition to correlations with age, other measures of construct validity include contrasted group, convergent, and discriminant validity. These measures of construct validity are described further below.

Comparisons of performance between groups that would be expected to differ on the construct (**contrasted group validity**) can be more informative, particularly if those analyses indicate whether effective decisions can be made on the basis of the assessment (Connolly, 2013). For example, one might consider whether a measure of reading distinguishes effectively between children with and without a diagnosis of dyslexia.

Correlations with other established assessed assessments can also be used to explore whether the same or different constructs are measured (**convergent** or **discriminant validity** respectively). To aid interpretation of the strength of the evidence for convergent validity, in the summaries provided in the shortlisted database we provided a qualitative descriptor based on the recommendations from the EFPA test review model: inadequate, $r < 0.55$; adequate, $0.55 \leq r < 0.65$; good, $0.65 \leq r < 0.75$; excellent, $r \geq 0.75$ (Evers, Hagemeister, et al., 2013).

Ratings used to evaluate construct validity in the shortlisted database

In line with the recommendations from the EFPA test review model (Evers, Hagemeister, et al., 2013), our overall ratings of construct validity were formulated by subjectively considering the evidence from all available sources with due consideration to the quality of each source of evidence. The evidence used to formulate these ratings is also summarised in the database (with citations to sources). We rated construct validity in response to the question, 'Does it adequately measure literacy, mathematics, or science?'

- A rating of 0 indicates that insufficient information was available to evaluate construct validity. This rating was effectively redundant because a test would not have proceeded to the evaluation phase without being able to make any judgement on the target construct.
- A rating of 1 indicates that the available information does not well support construct validity—descriptive and statistical evidence was problematic.

- A rating of 2 indicates limited support for construct validity—descriptive evidence provides some suggestion of face validity, but this has not been adequately tested (statistically).
- A rating of 3 indicates adequate evidence for construct validity—there is some evidence suggesting the test has construct validity but this is not extensive. This could be because there is inadequate description of the construct of measurement, or that the description of the target of measurement is extensive but the statistical evidence that the test successfully measures this target construct is weak.
- A rating of 4 indicates that the evidence very well supports the construct validity of the test. A detailed description of the construct of measurement was accompanied by a good deal of statistical evidence indicating that the test successfully measures this target construct.

Criterion validity

Criterion validity considers the extent to which test scores are related to scores on a real world measure of the construct (Evers, Muñiz, et al., 2013). Hence, in this review, criterion validity would be measured using evidence from comparisons against national key stage tests, GCSEs, or A-levels, which are gold standard measures of attainment in England. As noted previously, because the national curriculum is subject to change over time, this can present a problem for measuring criterion validity.

In the shortlisted measures database, we subjectively evaluated criterion validity using information reported in all available sources combined. Before describing how ratings were assigned, we first describe how criterion validity can be assessed.

When using test data, most educators will be interested in criterion validity because they want to know about **predictive validity**—how well does performance on the test predict future outcomes in key stage assessments. However, concurrent and postdictive validity might also be considered in order to determine whether the test adequately measures the construct that the user wants to measure. **Concurrent validity** considers how well performance on the test predicts current outcomes on a criterion measure. **Postdictive validity considers** how well performance on the test predicts prior performance on a criterion measure.

Whether predictive, concurrent, or postdictive, correlations between tests scores on the test and the criterion measure provide a statistical measure of validity. In the shortlisted measures database, where the criterion validity of a test was reported, we provide a summary of the evidence indicating the type of criterion validity measured and the strength of the statistical evidence (with citations to sources). To aid interpretation, we provided a qualitative descriptor based on the recommendations from the EFPA test review model: inadequate, $r < 0.20$; adequate, $0.20 \leq r < 0.35$; good, $0.35 \leq r < 0.50$; excellent, $r \geq 0.50$ (Evers, Hagemeister, et al., 2013).

Ratings used to evaluate criterion validity in the shortlisted database

A combination of consideration of the breadth and quality of the evidence was used to formulate the ratings used in the evaluation of criterion validity.

- A zero rating indicated that insufficient information was available to evaluate criterion validity.
- A rating of 1 indicated that the available information did not support criterion validity—measures of criterion validity are reported but the qualitative descriptors suggest the strength of the correlations are inadequate ($r < 0.2$).
- A rating of 2 indicated that there was limited evidence of criterion validity—one or two measures of criterion validity with correlations of inadequate to adequate strength ($r < 0.35$).
- A rating of 3 indicated that there was adequate evidence of criterion validity—at least one measure of criterion validity indicating a correlation of good to excellent strength ($r > 0.35$).
- A rating of 4 indicated that there was very good evidence of criterion validity—multiple measures indicating good to excellent criterion validity ($r > 0.35$).

While evaluating the validity of tests, it became apparent that few measures of attainment report any form of evidence for criterion validity. That is, only ten of the 37 evaluated assessments provided information as to how closely their measure was associated with actual performance on national tests. This was surprising to us because several measures will provide users with predictions for performance in key stage tests or GCSEs based on performance on the measures without having measured the predictive validity of these measures. As highlighted in the conclusions, we believe that—in the absence of adequate evidence of criterion validity—these predictions should be treated with caution.

Reliability

Reliability refers to the extent to which the test scores are likely to be reproducible. Any measurement is subject to random errors caused by inconsistencies in examinee's and examiner's behaviour, as well as test content (Connolly, 2013). Measures of reliability explain the degree to which the test is free from such **measurement error**. Tests often include measures such as confidence intervals or the standard error of measurement (the spread of observed scores around true score) which can give some information on the reliability of test scores in terms of precision; these are also helpful reminders of the measurement error inherent to the test when reporting test scores. Although the reliability and precision of measurement is always important, the impact during interpretation of test results increases with the stakes of decisions made on the basis of those test results (The Standards, 2014). There are a number of different ways to conceptualise and measure reliability, not all of which will be relevant to all tests. Hence, in the shortlisted database of tests we rated the reliability of a measure based on overall evidence for reliability, supported by a summary describing of the nature and strength of available evidence. In addition to considering the strength of the statistical evidence on reliability, it is also important to consider the quality and range of sources of evidence of reliability (Evers, Muñiz, et al., 2013). These different types of reliability are summarised below, before then describing how we formulated our ratings of reliability in the shortlisted database.

Internal consistency refers to the interrelatedness of items in the test. Items in a test invariably measure a variety of subskills and might also differ in how well they measure the target skill. As a result, an examinee's performance may differ if they were given different items (Connolly, 2013). Internal consistency includes internal reliability (the consistency of results across items within a test) and content or item sampling (the consistency of results on subsets of items) (Mokkink et al., 2010). This is often evidenced statistically by reporting Cronbach's alpha coefficient, Lambda-2, greatest lower bound, or factor analysis (for further information about how to calculation measures of validity and reliability also see Newton and Shaw, 2014). When considering the evidence for internal consistency, you should not only consider the size of the coefficients but also the size and quality of sampling and the number of different sources of reliability (Evers, Hagemeister, et al., 2013).

Note that high levels of internal consistency in a test is not, in itself, an indication of a good quality test. Consistency must be considered alongside other factors. Measures on which all children get all items correct or all items incorrect will be highly reliable, but not practically useful. Conversely, measures that test complex multi-dimensional skills may not be highly reliable because of the many skills involved, but may be useful. Finally, it can be difficult to measure internal consistency in adaptive tests in which not all participants receive the same items.

Item response theory (IRT) is a different approach to test creation that focuses on finding a set of individual items that are accurate measures of attainment rather than an overall test with high consistency. Item response theory assumes that item difficulty and child level of attainment can be measured on a single scale and therefore the percentage of children who get an item right gives an indication of that item's difficulty. This information can then be used to select a set of items that have varying levels of difficulty (Renaissance, 2019a, 2019b). Normally, tests created using IRT will start with a large number of items given to a wide range of children and select only the most accurate items for the final measure. This allows the test as a whole to be a reliable measure of attainment. Because tests which use IRT to select items may not be appropriate for measures of internal consistency, we note whether this approach was used to select items.

Temporal stability or test-retest reliability considers whether assessment results are consistent over time by comparing scores after either a short or a long duration between testing (Evers, Hagemeister, et al., 2013; Mokkink et al., 2010). Over short durations, this measures the magnitude of effects of incidental and temporary changes in mental state of the examinee or examiner (for example, motivation, fatigue; Connolly, 2013) and can also provide evidence of practice effects. Even where practice effects have not been reported, users are encouraged to consider whether they can reasonably assume that the nature of the test makes practice effects likely to impact on results (Evers, Hagemeister, et

al., 2013). Tests can minimise the impact of practice effects by indicating a suitable test-retest period where these effects can be avoided, or by producing alternate versions of a test.

Equivalence reliability examines whether test results are consistent across different versions of a test by examining performance by the same person on different parallel or alternate forms on the same occasion (Evers, Hagemeister, et al., 2013; Mokkink et al., 2010). Not all tests have parallel or alternate forms and so this measure of reliability is not always applicable. However, tests do commonly indicate that variations in administration are permissible and here too equivalence reliability should be considered and documented. For example, assessments are commonly available in both digital and paper format or for use with small groups or individuals. Evidence should be presented to indicate that test results remain reliable despite these variations in administration.

Inter-rater/inter-scorer reliability compares scores by different examiners on the same occasion. Intra-rater/scorer reliability compares scores by the same examiner on different occasions (Evers, Muñiz, et al., 2013; Mokkink et al., 2010). Evidence for these measures of reliability are particularly important when a qualitative assessment is made, and far less so when administration or scoring are highly manualised or even automatized digitally.

Ratings used to evaluate reliability in the shortlisted database

We used the recommendations from the EFPA test review model to formulate our reliability ratings, which ensured that we minimally considered evidence for internal consistency, temporal stability, equivalence reliability, and inter-rater reliability (Evers, Hagemeister, et al., 2013). Summary information is also reported describing the nature of comparisons that were available (with citations to sources) and highlighting whether additional measures of reliability that were not available would be beneficial to evaluation. We accompanied the description of the nature and size of reliability coefficients with a qualitative descriptor using the values recommended in the EFPA test review model: inadequate, $r < 0.70$; adequate, $0.70 \leq r < 0.80$; good, $0.80 \leq r < 0.90$; excellent, $r \geq 0.90$ (Evers, Hagemeister, et al., 2013). The ratings of reliability that were assigned in the shortlisted database were as follows.

- A zero rating indicated that insufficient information was available to evaluate reliability—this rating was redundant due to inclusion/exclusion criteria.
- A rating of 1 indicated that the available information did not support reliability—measure(s) of reliability are reported but the qualitative descriptors suggest the strength of the evidence is inadequate ($r < 0.7$).
- A rating of 2 indicated that there was limited evidence of reliability—one or two measures of reliability with inadequate to adequate strength ($r < 0.8$).
- A rating of 3 indicated that there was adequate evidence of reliability—at least one measure of reliability indicating good to excellent strength ($r > 0.8$).
- A rating of 4 indicated that there was very good evidence of reliability—multiple measures indicating good to excellent reliability ($r > 0.8$).

Quality of norms

Finally, evaluation of the quality of the available norms is important to determine whether the norms are meaningful. This includes consideration of the nature of sampling and representativeness of the norm-derived population (including recruitment methods and sample size) to examine whether the norms are appropriate to your target sample and free from bias. To some extent, these considerations are independent from the research question and therefore we were able to reflect on whether the population was appropriate and free from bias. However, when thinking about the representativeness of the norm-derived population, users must also consider whether the norms are representative of the specific characteristics of the sample they intend to apply the norms to. Here, we were specifically seeking assessments that are suitable for typically-developing 6- to 18-year-olds in the U.K. Hence the summary information in the database reflects any sampling biases that were noted in comparison to this population. However, database users seeking measures for subgroups of this population (for example, children in the U.K. with SEN or EAL) should consider what norm-derived population is appropriate to their research question.

Note that the tests that were included in the shortlisted database had already been selected after considering the norms. The screening criteria excluded tests that were criterion-referenced or without norms, norms that were not suitable for

the target population (for example, clinical sample, age range, or no U.K. standardisation sample), or that were considered out of date (see Table **2** and Appendix 1: Methodology and Search Terms).

Description of the evidence base

The systematic review culminated in the evaluation of 37 short-listed tests; 121 publications (assessment manuals, grey literature from publishers, and peer-reviewed publications) were reviewed in order to summarise implementation and evaluate the psychometric properties of these tests. The procedures used to summarise and evaluate these factors are described in detail in Appendix 1: Methodology and Search Terms. Table **5** describes the implementation factors and Table **6** describes the psychometric properties that are summarised in the shortlisted database. Summary information for each test is presented in Supplementary Materials 2.

Findings and gaps in the evidence base

Table **7** summarises the number of tests that received a rating of zero to four for construct validity.

- A zero rating indicates that insufficient information was available to review construct validity.
- A rating of one indicates that the available information does not support the construct validity of the measure.
- A rating of four indicates excellent construct validity was illustrated—this score was generally reserved for tests that demonstrated high face validity that was supported by at least one source of good statistical evidence for construct validity (see page 25 for further detail on how these ratings were formulated).

The notable pattern here is that although very few tests achieved a rating of four, few were rated as one or two. Thus, most tests on the shortlist did provide some indications of construct validity. This means that the tests should indeed measure attainment in the target construct that they set out to measure. However, it is worth noting that construct validity was often determined on the basis of a subjective judgement using the descriptive evidence of face or content validity rather than any objective statistical analysis.

Table 7: Summary of the construct validity rating of tests eligible for evaluation by subject area

	0 (insufficient information for evaluation)	1 (Weak)	2	3	4 (Strong)
Literacy	0	3	4	14	3
Mathematics	0	0	4	10	3
Science	0	0	0	1	0
OVERALL	0	3	6	23	5

Note: The sum of cells is greater than the total number of tests eligible for review because some tests include measures of both literacy and mathematics. The overall row counts each test once.

Table **8** summarises the number of tests that received ratings of zero to four for criterion validity.

- A zero rating indicates that insufficient information was available to review criterion validity.
- A rating of one indicates that the available information does not support the criterion validity of the test.
- A rating of four indicates excellent criterion validity was illustrated (see page 26 for further information about how these ratings were formulated).

The most obvious finding from this analysis is that only a small proportion of the tests present any information about criterion validity and very few present strong evidence.

Table 8: Summary of the criterion validity rating of tests eligible for evaluation by subject area

	0 (insufficient information for evaluation)	1 (Weak)	2	3	4 (Strong)
Literacy	19	0	0	3	2
Mathematics	11	0	1	2	3
Science	1	0	0	0	0
OVERALL	27	0	1	5	4

Note: The sum of cells is greater than the total number of tests eligible for review because some tests include measures of both literacy and mathematics. The overall row counts each test once.

Table 9: Summary of the reliability rating of tests eligible for evaluation by subject area

	0 (insufficient information for evaluation)	1 (Weak)	2	3	4 (Strong)
Literacy	0	1	6	9	8
Mathematics	0	0	1	11	5
Science	0	0	0	1	0
OVERALL	0	1	7	18	11

Note: The sum of cells is greater than the total number of tests eligible for review because some tests include measures of both literacy and mathematics. The overall row counts each test once.

Table **9** summarises the ratings for the reliability of tests in different subject areas. A similar scale is used here too.

- A zero rating indicates that insufficient information was available to review reliability.
- A rating of one indicates that the available information does not support the reliability of the measure.
- A rating of four indicates excellent reliability was illustrated. To achieve this, we would expect to see multiple measures indicating high levels of reliability.

Here, we see most tests rated as two or three (see page 28 for further information about how these ratings were formulated). Mostly, tests present a single measure of reliability (usually internal consistency) but do not assess temporal or equivalence reliability. This is particularly problematic for measures that are intended for use longitudinally.

Implications

Implications for policy and practice

Our review highlights that information about the validity and reliability of measures of attainment are often difficult to access, lacking, or of low quality. In particular, we noted that that very few measures reported criterion validity—that is, the relationship between the measure of attainment and school outcome tests such as key stage tests or GCSEs. This is disappointing and problematic because in many cases these attainment measures are marketed as a way for schools to predict performance on these tests. In some cases, predicted national test grades are one of the measures that the test will provide. We found that there was a worrying lack of detail in how these predictions were made. In some cases, scores were simply ranked and grade boundaries created. This is potentially appropriate for GCSEs, which are norm referenced, but only if the measure has been shown to be highly correlated with GCSE score (for example, having good criterion validity). It is not appropriate for criterion referenced measures such as Key Stage 1 and Key Stage 2. We would urge significant caution in teachers using these predicted grades and suggest that in many cases their own professional judgement of students they have known over a period of time would be a better predictor of outcome. Indeed, evidence from an EEF research paper analysing trial outcome data suggests that commercial tests are not strongly correlated with later key stage test performance (Allen, Jerrim, Parameshwaran, and Thompson, 2018).

An additional area of concern is that although tests commonly do report some measure of reliability, temporal stability and equivalence reliability are rarely assessed. This is quite a concern for tests that are used to measure progress over time, as is commonly the case for educators and evaluators. We have highlighted the importance of considering the purpose of assessment when selecting tests. If the purpose is to predict future attainment in national tests and qualifications, then users should particularly avoid tests that have low criterion validity and temporal stability. Since this, probably, is a key reason why educators use standardised tests, test developers should endeavour to provide this information in a readily accessible format.

We should acknowledge that there is a conflict between publishing up-to-date norms—as detailed above, tests should be re-normed every 10 to 15 years—and publishing criterion validity. The most useful form of criterion validity would be predictive validity, where individuals complete the standardised test at some time-point well in advance of the outcome or criterion measure. However, this would entail a longitudinal study which would both be costly and result in publication of older norming data, itself a threat to validity. An alternative, for measures that cover a range of year groups, could be to present a combination of good temporal stability between year groups and concurrent criterion validity in the age groups that take part in national tests. However, as already noted, temporal stability is also under-reported in general.

One publisher that has dealt with this conflict and does report criterion validity is Renaissance, publishers of Star Maths and Star Reading (Renaissance, 2019a, 2019b). Star Maths and Star Reading are online, adaptive measures which participants complete on the computer. Data is collected from all participants and used to update the norms. This approach is becoming increasingly popular with tests that are completed or scored online and it has the potential to substantially improve norming as long as some background information—such as date of birth, gender, date of test, and school—is recorded.

While evaluating implementation factors we noticed another threat to the validity and reliability of test data that was rarely acknowledged within the user manuals. Many tests offer alternate formats for implementation without publishing any evidence of equating studies. For example, tests very often suggest that administration group size can be whole class, small group with the support of a teaching assistant, or individual. The impact of behavioural factors in these different conditions should not be underestimated. While these adaptations might be necessary for practical reasons, user manuals should give clear guidance on how to determine group size and what impact this might have on the reliability and validity of results.

Similarly, we noted that several tests were available in both paper and digital formats. Anecdotally, our impression is that interest in digital assessment is increasing as it can be more cost effective and efficient. We would like to urge caution in the assumptions made when transitioning from paper to digital assessment. There is evidence that children perform less well on digital tests than paper tests (Støle, Mangen and Schwippert, 2020). The testing situations for these two types of test are often quite different (for example, some digital versions progress adaptively through the questions while their paper equivalents have flat administration). Yet, there is a paucity of studies examining whether the two were equivalent or not. Publishers often seem to assume that the standardisation for a pencil and paper version could be

used for the digital version, or vice versa. This may or may not be a reasonable assumption, but in the absence of evidence from standardisation trials such data should be interpreted with caution. Digital assessments might have advantages for some groups of individuals (for example, if they include accessibility functions) but might disadvantage others disproportionately (for example, if they depend upon digital literacy).

Connected to these issues about variation in test administration is the issue of fairness in testing, another fundamental threat to validity and potential source of bias in results. Testing situations must ensure that all examinees are equally able to show their abilities without characteristics such as age, disability, race or ethnicity, gender, or language impacting on their performance (The Standards, 2014). However, achieving fairness is quite a challenge. For example, the ability to use background knowledge to make inferences is an important component of reading comprehension. In an assessment situation, care must be taken to ensure that such background knowledge is equally available to examinees with different socioeconomic and cultural backgrounds. Standardised procedures have been considered fundamental to ensuring that all examinees have equal opportunity to demonstrate their abilities. Yet, inflexible standardised procedures can in fact limit accessibility and create construct-irrelevant barriers to performance (The Standards, 2014). For example, examinees with limited English language proficiency can be unfairly disadvantaged by the language of assessments. If language proficiency is irrelevant to the target construct, care must be taken to enable such individuals to demonstrate their abilities on the construct of interest.

Fairness can be improved through flexibility, to create accessible testing situations. However, whether a test has flexibility by design or a user deviates from standardised procedures to increase accessibility, evidence of the validity of score interpretations is still necessary. Some adaptations for accessibility may fundamentally alter the construct of measurement and therefore present a significant threat to validity (The Standards, 2014). For example, the provision of a reader to a pupil with dyslexia can support accessibility and limit the impact of construct-irrelevant bias due to their limited literacy skills; however, this adaptation would alter the construct of measurement in reading comprehension test (the adapted test assesses language comprehension rather than reading).

In many cases, we were not able to judge various aspects of the measures we evaluated. Often, details such as validity, reliability and, in particular, the methods used for standardisation such as participant recruitment, sample size and demographics, and evidence of floor or ceiling effects were simply not reported leaving us unable to evaluate the measure fully. This was especially true for computer-based tests with automated scoring where user manuals typically focused on administration and did not describe how norms were generated at all. Another notable issue is that even though we were able to review this information, it was only provided in technical manuals and not readily available to potential test users prior to purchase of a test. We found that most publishers were very forthcoming with inspection copies of materials when requested for the purposes of this review, however, gathering and sifting this information was extremely time consuming and would be beyond the scope of typical users. We suggest that there should be general guidelines on the information expected in the technical information for standardised tests, and that test publishers should ensure that basic information needed to evaluate the psychometric quality and implementation utility of assessments is readily available to potential users before test purchase (for example, summarised on the publisher website). The Standards for Educational and Psychological Testing (2014) provides a comprehensive source of guidance and we strongly encourage test developers to apply these standards to measures of attainment.

Implications for research

We have identified several aspects of available tests that need further research to provide evidence of the validity and reliability of existing measures. In addition, we have identified particular subject areas that have a particularly limited choice of high quality standardised tests.

Early in the process of agreeing the protocol, a decision was made to exclude tests which did not have a U.K. standardisation. This excluded a number of well-established, valid, and reliable measures that have been standardised within other English-speaking populations, particularly in the U.S. Test users should carefully consider whether this is an appropriate decision in the context of their own research question and methods. For example, the EFPA test review model (Evers, Hagemeister, et al., 2013) suggests that a normative sample size of 1,000 is excellent. Hence, large scale randomised controlled trials may collect sufficient test data that re-standardisation using data from pupils in the control condition would be as high quality (or better) than published normative data. If this is the case, test users might not want to rule out measures that have been standardised within other populations, or even tests that do not have norms at all.

Another important area for future research is to evaluate the validity and reliability of ability measures as these measures are often more informative if users wish to understand why children are not attaining as well as hoped, or to understand why an intervention is effective.

Limitations

There was only one measure of science attainment that passed the screening measures in contrast to the range of measures available for literacy and mathematics. Eighteen science assessments were excluded because they were not norm-referenced. This is clearly a significant gap and we would welcome more science attainment measures with U.K. standardisations. One possible reason for this gap is that because science is not formally assessed in national tests until GCSE, test developers might not consider there to be sufficient interest in formal assessment in this subject. An alternative possibility, though, is that the nature of scientific attainment and how science is taught does not lend itself well to standardised tests. There is a great deal of variation between schools in terms of which science subjects are taught and when, particularly through Key Stages 2 and 3. This makes it very difficult to design a science assessment that aligns with the curriculum and is predictive of future attainment. The reasons for this gap in the availability of tests should also be explored further in order to resolve this issue. The decision to exclude criterion-referenced assessments in the present study might have been more appropriate in literacy and mathematics, where knowledge builds systematically on previous knowledge, but may have been less appropriate in science where content knowledge is likely to have a greater impact on attainment and may build in different areas relatively independently. Further research is needed to make recommendations on how to evaluate criterion-referenced assessments, and to apply these recommendations to evaluate the assessments that are available to measure attainment in science.

Team

Dr Helen Breadmore (PI, helen.breadmore@coventry.ac.uk), Associate Professor (Research), Centre for Global Learning, Coventry University, oversaw the project, leading development of the database, systematic searches, evaluated measures, and was first author of the narrative synthesis.

Professor Julia Carroll (Co-I, julia.carroll@coventry.ac.uk), Professor, Centre for Global Learning, Coventry University, co-authored the narrative synthesis, evaluated measures ,and reviewed information coded within the searchable database.

Three research assistants—Blaire Sweeney, Katie Baker, and Steve Raven—supported the PI and Co-I by following the procedures outlined in this protocol to gather the information needed for evaluation of measures (hand searching websites, conducting systematic database searches, and requesting inspection copies of materials from publishers).

The advisory panel (see Table 10) was formed of experts in the fields of literacy, mathematics, science, assessment design, and evaluation. The advisory panel supported the development of the systematic review protocol.

Table 10: Advisory panel members

Name	Job title	Affiliation
Dr Katie Baker	Specialist mathematics teacher, PhD evaluated a mathematics intervention programme.	Coventry University
Kate Blundell	Specialist dyslexia teacher, member of the SpLD Assessment Standards Committee, studying for a PhD in dyslexia diagnosis.	Coventry University
Dr Michelle Ellefson	Reader in Cognitive Science.	University of Cambridge
Dr Judith Hillier	Associate Professor of Science Education (physics), Vice President and Fellow of Kellogg College.	University of Oxford
Professor Jeremy Hodgen	Professor of Mathematics Education.	UCL Institute of Education
Wayne Jarvis	Senior Network Education Lead.	STEM Learning
Professor Duncan Lawson MBE	Director of Sigma and Professor of Mathematics Education.	Coventry University
Lynne McClure	Director.	Cambridge Mathematics
Dr Sue Stothard	Independent Consultant.	Stothard Education
Helen Wilson	Affiliate Lecturer (science).	Oxford Brookes

Conflicts of interest

The review team do not have any conflicts of interest. Some members of the advisory board have been involved in development of assessments or are affiliated with organisations that develop or publish assessments. Their expertise was invaluable to development of the review protocol. The advisory board did not, however, have any involvement in the systematic review or influence the evaluation of measures.

The review was commissioned by the Education Endowment Foundation, which also reviewed the protocol.

References

Allen, R., Jerrim, J., Parameshwaran, M. and Thompson, D. (2018) 'Properties of Commercial Tests in the EEF Database', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf

Breadmore, H. L. and Carroll, J. M. (2020) 'Protocol for a Systematic Review of Measures of Attainment in Literacy, Mathematics and Science', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Attainment_Review_Protocol.pdf

Breadmore, H. L., Vardy, E. J., Cunningham, A. J., Kwok, R. K. W. and Carroll, J. M. (2019) 'Literacy Development: Evidence Review', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Literacy_Development_Evidence_Review.pdf

Connolly, A. J. (2013) 'KeyMaths3UK: Diagnostic Assessment Manual', London: Pearson Assessment.

Cunningham, A. and Carroll, J. (2011) 'Age and Schooling Effects on Early Literacy and Phoneme Awareness', *Journal of Experimental Child Psychology*, 109, pp. 248–255.

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y. W. and Cordier, R. (2017) 'Psychometric Properties of Language Assessments for Children Aged 4–12 Years: A Systematic Review', *Frontiers in Psychology*, 8, p. 1515. DOI: 10.3389/fpsyg.2017.01515

DfE (2014) 'The National Curriculum in England: Framework Document', London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381344/Master_final_national_curriculum_28_Nov.pdf

DfE (2019a) 'Key Stage 2 Assessment and Reporting Arrangement (ARA)', London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/838961/2020_KS2_assessment_and_reporting_arrangements.pdf

DfE (2019b) 'Primary School Accountability in 2019: Technical Guide', London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/854515/Primary_school_accountability_in_2019_technical_guide_2_Dec_2019.pdf

DfE (2020) 'Secondary Accountability Measures: Guide for Maintained Secondary Schools, Academies and Free Schools' (DfE-00026-2020), London: Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872997/Secondary_accountability_measures_guidance_February_2020_3.pdf

Dockrell, J., Llaurado, A., Hurry, J., Cowan, R., Flouri, E. and Dawson, A. (2017) 'Review of Assessment Measures in the Early Years: Language and Literacy, Numeracy and Mental Health', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Review_of_assessment_measures_in_the_early_years.pdf

Elliott, C. D. and Smith, P. (2011) 'British Ability Scales 3: Technical Manual', London: GL Assessment.

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J. and Sjöberg, A. (2013) 'EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests: Test Review Form and Notes for Reviewers', EFPA Board of Assessment Document 110c. http://www.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A. and Bartram, D. (2013) 'Assessing the Quality of Tests: Revision of the EFPA Review Model', *Psicothema*, 25 (3), pp. 283–291. DOI: 10.7334/psicothema2013.97

Flynn, J. R. (2012) *Are We Getting Smarter?: Rising IQ in the Twenty-First Century*, Cambridge University Press.

GL Assessment (2020) 'CAT4 and Strategies for Learning: Cognitive Abilities Test', London: GL Assessment.

Hodgen, J., Foster, C., Marks, R. and Brown, M. (2018) 'Improving Mathematics in Key Stages 2 and 3: Evidence Review', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/improving-mathematics-in-key-stages-two-and-three/

International Test Commission (2001) 'International Guidelines for Test Use', *International Journal of Testing*, 1 (2),pp. 93–114. DOI: 10.1207/s15327574ijt0102_1

International Test Commission (2014) 'ITC Statement on the Use of Tests and Other Assessment Instruments for Research Purposes'. https://www.intestcom.org/files/statement_using_tests_for_research.pdf

Joint Committee of the American Educational Research Association, American Psychological Association and National Council on Measurement in Education (2014) 'Standards for Educational and Psychological Testing', Washington, DC: American Educational Research Association.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. and Group, P. (2009) 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement', *BMJ*, 339, b2535. DOI: 10.1136/bmj.b2535

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M. and Terwee, C. B. (2018) 'COSMIN Risk of Bias Checklist for Systematic Reviews of Patient-Reported Outcome Measures', *Quality of Life Research*, 27 (5), pp. 1171–1179. DOI: 10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C. and Terwee, C. B. (2018) 'COSMIN Methodology for Systematic Reviews of Patient—Reported Outcome Measures (PROMs): User Manual'. https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Mokkink, L. B., Terwee, C. B., Patrick, D. L., Alonso, J., Stratford, P. W., Knol, D. L., . . . de Vet, H. C. (2010) 'The COSMIN Checklist for Assessing the Methodological Quality of Studies on Measurement Properties of Health Status Measurement Instruments: An International Delphi Study', *Quality of Life Research*, 19 (4), pp. 539–549. DOI: 10.1007/s11136-010-9606-8

Newton, P. and Shaw, S. (2014) *Validity in Educational and Psychological Assessment*, London: Sage.

Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R. and Miller-Friedmann, J. (2017) 'Review of SES and Science Learning in Formal Educational Settings: A Report Prepared for the EEF and the Royal Society', London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/public/files/Review_of_SES_and_Science_Learning_in_Formal_Educational_Settings.pdf

OECD (2019) 'PISA 2018 Assessment and Analytical Framework'. https://www.oecd-ilibrary.org/education/pisa-2018-assessment-and-analytical-framework_b25efab8-en

Renaissance (2019a) 'Star Assessments for Maths: Technical Manual', London: Renaissance Learning, Inc.

Renaissance (2019b) 'Star Assessments for Reading: Technical Manual', London: Renaissance Learning, Inc.

Roberts, N. (2020) 'Assessment and Testing in Primary Education (England)', House of Commons Library, Briefing Paper CBP 07980. https://commonslibrary.parliament.uk/research-briefings/cbp-7980/

Ruttle, K., Lallaway, M., Bennett, M., Pepper, L., Kilburn, V., Swift, J., . . . McCarty, C. (2020) 'New PiRA Primary: Progress in Reading Assessment, Key Stage One & Key Stage Two Interim Test Guidance', Abingdon: RS Assessment from Hodder Education.

Standards and Testing Agency (2020) 'Understanding Scaled Scores at Key Stage 2'. https://www.gov.uk/guidance/understanding-scaled-scores-at-key-stage-2

Støle, H., Mangen, A. and Schwippert, K. (2020) 'Assessing Children's Reading Comprehension on Paper and Screen: A Mode-Effect Study', *Computers and Education*, 151. DOI: 10.1016/j.compedu.2020.103861

Walker, A. J., Batchelor, J. and Shores, A. (2009) 'Effects of Education and Cultural Background on Performance on WAIS-III, WMS-III, WAIS-R and WMS-R Measures: Systematic Review', *Australian Psychologist*, 44 (4), pp. 216–223. DOI: 10.1080/00050060902833469

Wechsler, D. (2018) 'Wechsler Individual Achievement Test, Third UK Edition for Teachers (WIAT-IIIUK-T): Examiner's Manual', London: Pearson Assessment.

Appendix 1: Methodology and Search Terms

The systematic review protocol (Breadmore & Carroll, 2020) was developed by considering and combining several different sources of information: the recommendations from our advisory panel (see Table 10); the COSMIN checklist (Consensus-based Standards for the selection of health status Measurement Instruments - Mokkink et al., 2010); and the EFPA (European Federation of Psychologists' Associations) revised review model for the evaluation of tests (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013). The protocol was independently peer reviewed by the Education Endowment Foundation prior to commencing searches. A PRISMA diagram is provided in Figure 1.

In the COSMIN study, a four-round Delphi method was used to develop a taxonomy and checklist to evaluate the methodological and measurement quality of health-related patient-reported surveys (see https://www.cosmin.nl/). The COSMIN taxonomy (Mokkink et al., 2010) provides a useful summary of the importance and utility of measures of reliability and validity, which we apply to the evaluation of the psychometric properties of assessments of attainment using the EFPA review model (described further below). In particular, we applied the principles from the COSMIN risk of bias checklist (Mokkink, de Vet, et al., 2018) to select the EFPA review questions used to evaluate the tests. We also used the principles from this checklist to combine reliability and validity information from multiple sources (e.g., from the administration/technical manuals for assessments and peer reviewed journal articles) and account for risk of bias in these studies. In this review there were very few tests where it was necessary to combine information from multiple sources, a dearth of information was a far more common concern. The COSMIN checklists are not applicable to evaluation of quantitatively measured standardised assessments (having been developed to evaluate qualitative self-reported measures of patient health outcomes).

The EFPA review model was developed by the Board of Assessment (http://www.efpa.eu/professional-development/assessment) for the description and evaluation of psychological and educational tests. This review model is more applicable to evaluation of quantitative measures such as tests of attainment, and similarly highlights the need to evaluate the psychometric properties of tests (reliability and validity), but also highlights the importance of providing qualitative evaluation of implementation factors. The EFPA review model informed inclusion and exclusions criteria in this review, and selected questions from "Part 2 Evaluation of the Instrument" were used to evaluate tests in the final stage of the review (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013). Further detail about how this review model was applied to the present review is provided in the following sections.

Inclusion and exclusion criteria for the review

The PRISMA diagram in Figure 1 illustrates the process of systematic inclusion and exclusion of sources of information about assessments.

Test identification was achieved using the search strategy described below to identify a long list of tests for consideration. Duplicates were removed and tests were then screened for relevance to the review using information that was publicly available from publisher and distributor websites (see Table 2 for criteria).

After screening tests, full text copies of the relevant administration, scoring and technical manuals were obtained. This information was supplemented with systematic database searches to identify peer reviewed publications that provided information about the psychometric properties of each test. Peer reviewed publications were screened for relevance to the review (see Table 13 for criteria). Information from peer reviewed articles and test manuals was then combined, and tests were subjected to eligibility checks to identify whether essential information needed to evaluate the psychometric properties of the tests was available to review (see Table 3 for criteria). These assessments formed the short-list for full evaluation which was included in this written synthesis.

Finally, short-listed tests were evaluated, summarising key implementation factors (see Table 5) and systematically evaluating the psychometric properties of tests (see Table 6).

Search strategy: test and publication identification

Test identification

Initial searches created a "long list" of tests, by finding the name and acronyms of tests of literacy, mathematics and science that are available in the UK. Note that an a priori decision was made that national tests and qualifications (such as Key Stage assessments, GCSEs, PISA tests) were not included in the database, because the content and norming varies over time. The database will contain minimal information about all tests on the long list, as indicated in Table 1.

Cultural and educational background is well documented to influence performance on standardised assessments (e.g., Walker, Batchelor, & Shores, 2009 reviews). Further, norm-referenced tests are only suitable for use with individuals who are demographically similar to the normative sample. There is a well-established trend for cohorts to score higher than similar cohorts ten or twenty years ago (Flynn, 2012). Hence, tests are only included in this review if they have been recently published and normed with a relevant sample. At this stage, "relevance" is loosely defined to allow identification of older tests that have re-normed, hence we included all tests published in or since 2000 (see also Denman, et al., 2017 who followed the same procedure). However, to be included on the long list, tests must be relevant to and suitable for the target subjects – 6 to 18-year-old pupils in the UK.

Search criteria to identify tests include an initial screen to ensure that measures are;

- Used to assess literacy, mathematics or science attainment.
- Published in or since 2000 (see also Denman et al., 2017).
- Suitable for English-speaking 6 to 18-year-olds.

If it is not initially clear whether a test fulfils these criteria, the test will be included in the long list but may be filtered out during screening and/or eligibility checks.

Tests were identified by

- Comprehensive hand searches of publisher and distributor websites, indicated in Table 11. This list of 18 websites was identified by the advisory panel, who were asked to identify any websites that they used to access assessments, or that they knew teachers or researchers commonly used.
- Search of the ERIC database using search terms based on recommendations from the COSMIN review[2]. The ERIC database then provides a list of tests included in these papers, which were added to the database. Further, the publications were added to the list of publications for review.
    - Search terms: (Assessment: Literacy OR Assessment: Math* OR Assessment: Scien*) AND (Measure* OR Test* OR Assess* OR Screen*) AND (Psychometr* OR Reliability OR Validity) AND (educationlevel: Elementary Education OR educationlevel: Secondary Education OR educationlevel: Elementary Secondary Education OR educationlevel: Middle Schools OR educationlevel: High Schools OR educationlevel: Junior High Schools OR educationlevel: Primary Education).
    - Limitations: Peer reviewed only, Location: United Kingdom.
- Other sources, including
    - Outcomes measures used in EEF trials and a list of recommended measures (provided by the EEF in personal communication, 29/01/2020)
    - Recommendations from the advisory panel, who were asked to provide a list of any tests commonly used in UK schools (for teaching or research purposes).

[2] As the literature searches were considered supplementary to the publisher searches, some further searches of other databases were deemed unfeasible due to the quantity of information likely to be yielded, and the limited likelihood of these searches returning meaningful information for the purposes of this review. For example, an equivalent search of PsycInfo returned 13,250 articles. As you will see later in this review, a very small proportion of sources found through database searches yielded useful information to the evaluation of tests.

- Using an iterative approach to identification of tests, supplementing the long list with any publicly available assessments identified through the review process that fulfil the search, screening and eligibility criteria. For example, additional tests could be encountered when checking version history, during publication identification, or while reviewing concurrent validity. In which case, initial checks were conducted to ensure that these tests meet the search criteria above and further publication searches would be conducted. Tests were then subject to screening and eligibility checks before inclusion in the qualitative synthesis.

An initial screen checked that measures identified in these ways met the initial search criteria. This is indicated in the identification phase of the PRISMA diagram by the difference between the numerator and denominator. The denominator indicates all measures identified by a source. The numerator indicates the number of these measures that met initial search criteria.

Table 11: Websites that were hand searched

Publisher/distributer name	Website
Pearson: Pearson Clinical (including The Psychological Corporation)	www.pearsonclinical.co.uk
Pearson: Pearson Schools and FE Colleges	www.pearsonschoolsandfecolleges.co.uk
Pearson: Pro-Ed	https://www.proedinc.com/
GL Assessment	www.gl-assessment.co.uk
NFER	www.nfer.ac.uk
Hodder Education	www.hoddereducation.co.uk/rsassessment
Hodder: Rising stars	www.risingstars-uk.com/subjects/ assessment
Centre for Evaluation and Monitoring (CEM)	www.cem.org
Hogrefe	www.hogrefe.co.uk
Ann Arbor Publishers	www.annarbor.co.uk/
Oxford University Press	https://global.oup.com/education/content/primary/key-issues/assessment/?region=uk
Cambridge Assessment	https://www.cambridgeassessment.org.uk/about-us/what-we-do/assessment/
Collins	https://collins.co.uk/pages/collins-assessment
Renaissance Star Assessments	www.renlearn.co.uk
Dyslexia action shop	http://dyslexiaactionshop.co.uk
Psychological Assessment Resources (PAR) Inc	www.parinc.com
SEN books	www.SENbooks.co.uk

All tests, in all subjects, were appraised using the same criteria. The next step was to establish whether each test should be fully evaluated.

Screening tests

Minimal assessment information was included for all tests identified through the searches, as outlined in Table **1** and Table **2**. Following the recommendations of the EFPA review model, this information was provided by publishers (Evers, Hagemeister, et al., 2013). A brief description of the test will be obtained directly from the publisher website. Reasons

for screening a test are summarised in the "Exclusion criteria" column of Table **2** and is included in the measures database

Publication identification

Subsequent searches identified information needed to evaluate the psychometric properties of the measure. This included information provided to users in administration and/or technical manuals supplied with the test, and information available in the academic (or other) literature (Evers, Hagemeister, et al., 2013). Publishers may hold further information that is not publicly available (Evers, Hagemeister, et al., 2013). However, to ensure that the content of this review is reliable and replicable, here we evaluate tests using information sourced from;

(a) Standard administration and technical manuals provided to test users obtained from our own test library, subject librarians, publishers, distributors, and test authors.

(b) Peer reviewed publications identified through systematic database searches. Search terms and limitations are described in Table 12, and are based on those recommended in the revised COSMIN recommendations (Mokkink, Prinsen, et al., 2018).

Table 12: Search terms and limitations for publication identification – information about assessments.

Database	Search terms	Limitations
PsycInfo	tests and measures: (Name of test OR acronym of test) AND (Child*) AND (Measure* OR Test* OR Assess* OR Screen*) AND (Psychometr* OR Reliability OR Validity)	Search mode: Find all my search terms. Turn off Apply equivalent subjects. English AND Language: English AND Age group: School age (6-12 years; Adolescence (13-17 years) AND publication date in or after year of assessment publication AND peer reviewed journal AND peer reviewed AND Document Type: Journal Article AND exclude dissertations

Following searches, further criteria must be met for publications (including manuals) to be included in the review. The abstracts of peer reviewed publications were initially be screened using the criteria in Table 13, based on the revised COSMIN recommendations (Mokkink, Prinsen, et al., 2018). Exclusion criteria are recorded to indicate why publications were screened. Note, that we do not assess the methodological quality of the studies at this point.

Table 13: Screening criteria for publications about tests

Criteria for inclusion	Exclusion criteria
Relate to at least one test on the long list.	Test screened.
Study aims to Evaluate one or more psychometric property (i.e., reliability and/or validity) of the test. Develop a new test. Evaluate interpretability of the test. Study <u>does not</u> merely use the test as an outcome measure. The following studies should be excluded Randomised controlled trials. Studies where the test is used to validate another test.	Publication does not contribute to psychometric evaluation.
Include typically developing English speaking British children aged 6 to 18-years.	Sample is not relevant to review.

Content or sampling differs from the information provided elsewhere (i.e., does not duplicate the manual/other publications).	Publication does not contribute novel information.

Eligibility criteria

Following screening, all manuals and full-text publications will be reviewed to establish whether enough information is available about a test to evaluate the psychometric quality of that test in line with the recommendations from the COSMIN study (Mokkink et al., 2010) and EFPA Review Model (Version 4.2.6, (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013).

For a test to be eligible for evaluation, manuals and/or full-text publications must present at least one measure of reliability and at least one measure of validity. See Table 3 for the minimal information to be included in the database, and a summary of terms that will be accepted as measures of reliability and validity. Note that evaluation of responsiveness and interpretability (also recommended by COSMIN) is beyond the scope of this review.

If tests were excluded because information needed for full evaluation could not be obtained or did not pass screening or eligibility checks, no further evaluation took place[3]. Note, for example, criterion referenced assessments identified through this process will be documented in the long list database but will be screened and therefore will not be evaluated. The database indicates why a test was excluded from full evaluation, using the 'exclusion criteria' indicated in Table 2 and Table 3. It is essential for both the evidence synthesis and the online interface of the database to make it clear that assessments excluded in the filter are not of low psychometric quality, but that systematic searches did not identify enough information for evaluation.

Evaluation and appraisal of tests

The data collected at this point forms the criteria for evaluation of tests. This includes more detailed information about implementation from the test manual (see Table 5) which will enable users to filter and short-list the measures, and an evaluation of the psychometric properties of the test using a broader range of sources (see Table **6**).

Implementation factors

Information about implementation has been gathered from test manuals and provided as descriptors, consistent with recommendations from the EFPA review model (Evers, Muñiz, et al., 2013). This information could be used to search or filter the measures database. Implementation is not, however, rated in the evaluation of tests. The implementation factors evaluated in Table **5** were selected in consultation with our advisory panel and largely align to Part 1 of the EFPA review model "Description of the instrument"[4]. All terms will be defined in the written synthesis.

Evaluation of psychometric properties

Evaluation of the psychometric properties (validity, reliability and quality of norms) of a test was conducted using selected questions from the EFPA review model (Evers, Hagemeister, et al., 2013). First, we reviewed all sources of validity and reliability (i.e., each publication) independently, before combining into an overall evaluation for each assessment using methodology based on the COSMIN risk of bias checklist (Mokkink, Prinsen, et al., 2018). This enabled us to effectively and objectively combine information gathered from both manuals and academic sources. This information was summarised in the measures database as indicated in Table **6**.

- Construct validity questions culminated in an overall construct validity score from 0-4. This is an overall judgement rather than a simple average of scores across different questions model (Evers, Hagemeister, et al., 2013). A score of 0 indicates that validity cannot be rated because of lack of information, 1 is inadequate validity, 2 is adequate validity, 3 is good validity and 4 is excellent

[3] On the whole, these criteria match the first filter align with those used during development of the early years measures database (Dockrell et al., 2017).

[4] Note, however, that we have not evaluate computer generated reports or supply costs. These implementation factors are beyond the scope of this review and are likely to be subject to change over time.

validity. Scores of three or above were translated to a star in the measures database. In addition, we consider to what extent the assessments reflect the multi-dimensionality of the target construct (structural validity)? See p25 for further description of how these ratings were calculated.

- Criterion validity questions culminated in an overall criterion validity score from 0-4 using the same scale as construct validity model (Evers, Hagemeister, et al., 2013). Scores of three or above were translated to a star in the measures database. See p26 for further description of how these ratings were calculated.
- Reliability questions culminated in an overall score of 0-4 using the same scale as construct and criterion validity model (Evers, Hagemeister, et al., 2013). Scores of three of more received a star in the measures database. See p28 for further description of how these ratings were calculated.
- The EFPA review model does not provide an overall score for evaluating the quality of available norms model (Evers, Hagemeister, et al., 2013), hence in line with these recommendations we note any biases in norming.

Data synthesis

The database of assessments contained all information indicated in the tables above. This will be shared with the EEF in an excel spreadsheet. The EEF will implement the database on their website, supported by a user testing group including members of the research team and advisory panel.

Protocol deviations

The timeline for conducting the systematic review was delayed and extended due to the COVID-19 pandemic impacting on the resources available to the research team and their ability to contact publishers.

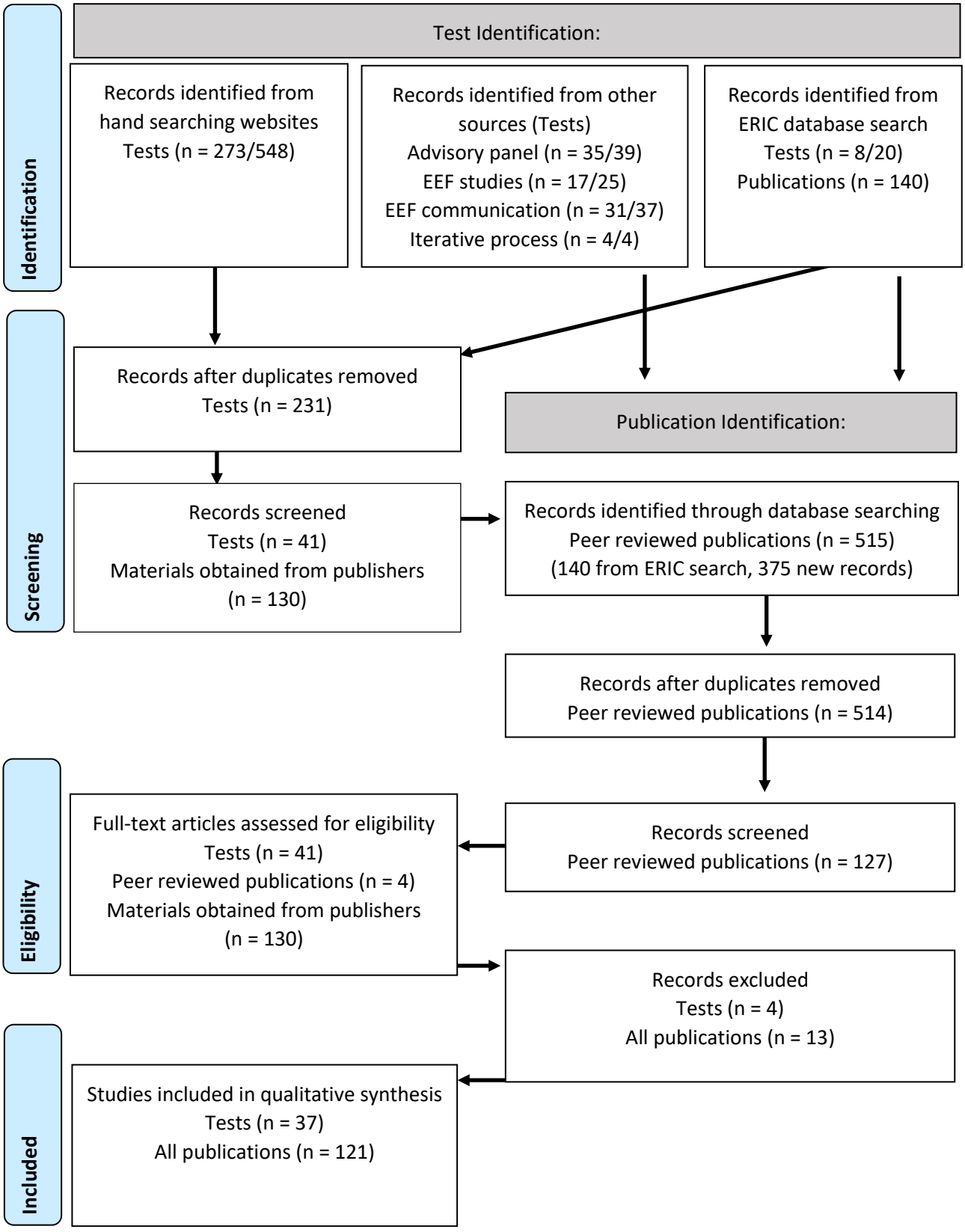
The review team added an additional free text field to the summary of construct validity in the evaluation stage, given the variation in quality of evidence for face validity combined with the dearth in statistical measurement of validity we felt it was important to be able to describe how our subjective judgements of construct validity were determined.

Timeline

Dates	Activity	Staff responsible/ leading
January - March 2020	Advisory Panel Meeting Development of systematic review protocol. Two drafts delivered to EEF and reviewed.	HB
May 2020	Systematic Review evaluation protocol approved by EEF	EEF
January - June 2020	Identification of assessments through hand searches.	Research assistants
July 2020	ERIC database search Initial screening of assessments. Systematic searches for peer reviewed publications.	HB
August - September 2020	Gathering publisher's materials for evaluation	Research assistants
October - November 2020	Evaluation of assessments. Analysis of results. Delivery of draft review and draft database content to EEF.	HB and JC
December 2020	Independent peer review.	EEF
January 2021	Revised review and database content delivered to EEF.	HB and JC

Appendix 2: PRISMA flow diagram

Figure 1: Flowchart of selection criteria according to PRISMA (Moher, Liberati, Tetzlaff, Altman, & Group, 2009). In the test identification phase, the numerator indicates all eligible records identified and the denominator indicates all records leniently included at first, but which on further investigation were revealed as ineligible for inclusion.



Supplementary Materials 1

Summary information for each test in the long list database is presented in Supplementary Materials 1.

Supplementary Materials 2

Summary information for each test in the short-listed database is presented in Supplementary Materials 2.

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn