MAY 14, 2020

PROTOCOL FOR A SYSTEMATIC REVIEW OF MEASURES OF ATTAINMENT IN LITERACY, MATHEMATICS AND SCIENCE

VERSION 4 (AFTER REVIEW)

DR HELEN L BREADMORE, PROF JULIA M. CARROLL
CENTRE FOR GLOBAL LEARNING: EDUCATION AND ATTAINMENT, COVENTRY UNIVERSITY
Priory Street, Coventry, CV5 8LB

Protocol for A Systematic Review of Measures of Attainment in Literacy, Mathematics and Science

Principal investigator(s): Dr Helen L. Breadmore, Professor Julia M. Carroll

Table of contents

Protocol for A Systematic Review of Measures of Attainment in Literacy, Mathematics and Science

Principal investigator(s): Dr Helen L. Breadmore, Professor Julia M. Carroll

Background and review rationale

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.

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.
- Evaluate the effectiveness of interventions.

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

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

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

Currently, there are few sources of impartial guidance and information to find and compare assessments literacy, mathematics and science attainment in school age children. The Education Endowment Foundation SPECTRUM database (https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/measuring-essential-skills/spectrum-database/) and Early Years Measures Database (https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/early-years-measure-database/early-years-measures-database/) evaluate assessments for other constructs/populations, but there is not a comparable resource for literacy, mathematics and science attainment in school age children. Indeed, much of the information that users needed to make an informed choice is within assessment 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 assessments available. Selected questions from the European Federation of Psychologists' Associations (EFPA) review model for the description and evaluation of psychological and educational tests (Evers, Hagemeister, et al., 2013) will be used to evaluate assessments. This information will be summarised within a searchable database, and accompanied by a written synthesis outlining the systematic review methodology used to form the database.

The database will be somewhat comparable to the Early Years measures database, but will include additional information and filters. A rating system based on the psychometric properties of the test will transparently indicate the quality of each assessment. 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, researchers) this is important, ensuring that implementation factors are considered a preference, and are not misinterpreted as relating to the quality of an assessment.

The written synthesis will include definitions of terminology used to evaluate the psychometric properties (reliability, validity, norms), how to interpret this information when selecting assessments, description of the systematic review methodology, summaries of assessments subjected to evaluation including the proportion of assessments evaluated as 4*/3*/2*/1*/0* in each subject.

Objectives

This review will provide much-needed guidance to support selection of measures of attainment in literacy, mathematics and science. Our approach will focus on assessments of particular relevance to educators and evaluators in the UK, who wish to measure the attainment of children and adolescents aged 6 to 18-years. The evidence will be summarised in a written synthesis and presented in a searchable database. The Education Endowment Foundation (EEF) will publish these outputs on their website.

The research questions are:

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

Protocol for A Systematic Review of Measures of Attainment in Literacy, Mathematics and Science

Principal investigator(s): Dr Helen L. Breadmore, Professor Julia M. Carroll

How can teachers and evaluators assess attainment and progress in literacy, mathematics and science in the UK?

Before reviewing measures of attainment, we will first define what we mean by attainment in literacy, mathematics and science. To do so, we have consulted recent evidence reviews commissioned by the EEF (Breadmore, Vardy, Cunningham, Kwok, & Carroll, 2019; Hodgen, Foster, Marks, & Brown, 2018; Nunes et al., 2017), policy documents (DfE, 2014; OECD, 2019) and worked with key experts in each field and in assessment more broadly, through our advisory panel (see p22). The key concepts in each subject are described briefly below and will be defined more elaborately within the narrative of the evidence synthesis. The purpose of this narrative is to describe what assessments in Literacy, Mathematics and Science might measure. It is clear that attainment in all three subjects depend upon multi-faceted sources of knowledge and skills. Hence, we should distinguish between general and specific measures of attainment.

Literacy

Accurate and fluent reading and writing are not only essential means of communication in modern society, but also underpin learning and assessment throughout the curriculum. Shortly after children learn to read, they begin to read to learn (Chall, 1983). Literacy includes word-level knowledge (word reading and spelling) and text and language level knowledge (reading comprehension and writing composition). Reading and writing occur in multiple modalities – reading can be silent or oral, writing can be handwritten or typed (Breadmore et al., 2019). Literacy draws upon a wide range of cognitive and linguistic skills, but for the purposes of reviewing measures of attainment, we focus on key concepts in literacy attainment rather than the underpinning skills. We have worked with our advisory panel to define key concepts in literacy attainment as;

- Word reading – including regular and irregular words which assess use of grapheme-phoneme correspondence and orthographic knowledge[1].
- Reading fluency – the combination of accuracy and rate (and prosody in oral reading), which can be measured at the level of the word, sentence and prose.
- Reading comprehension – which involves understanding of grammar and syntax, use of literal and elaborative inferences, as well as knowledge of narrative and genre.[2]
- Spelling – similarly to word reading, including regular and irregular words.
- Writing fluency – can be measured at the level of word, sentence and prose. Includes assessment of handwriting and typing fluency, in addition to writing rate.
- Writing composition – includes correct use of grammar and punctuation during writing, knowledge of narrative structure and genre, understanding of audience and wider context.
- Omnibus literacy tasks (e.g., sentence completion, which demands use of both reading and writing processes).

[1] Note that we do not include nonword reading in word reading. Nonword reading is a measure of grapheme-phoneme decoding, which is an underlying skill that contributes to word reading, but is not a measure of literacy attainment in of itself.

[2] Note that we do not include language comprehension or background reading in reading comprehension. These skills are crucial for success in reading comprehension, but are not measures of literacy attainment.

Mathematics

Mathematical proficiency and fluency is also essential to academic success throughout the curriculum and crucial to everyday life. The (US) National Research Council (2001, p. 116) described the following five strands or components of mathematical proficiency;

- Conceptual understanding—comprehension of mathematical concepts, operations, and relations.
- Procedural fluency—skill in carrying out procedures flexibly, accurately, efficiently, and appropriately.
- Strategic competence—ability to formulate, represent, and solve mathematical problems.
- Adaptive reasoning—capacity for logical thought, reflection, explanation, and justification.
- Productive disposition—habitual inclination to see mathematics as sensible, useful, and worthwhile, coupled with a belief in diligence and one's own efficacy.

These components of mathematical knowledge are not independent from one another, they are "interwoven and interdependent". However, it is sometimes possible to arrive at a correct answer without proficiency in all components. As a concrete example, one can calculate the length of a side of triangle by applying Pythagoras Theorem by using procedural fluency. Getting to the correct answer does not necessarily necessitate conceptual understanding of why the equation is true.

For the purposes of reviewing measures of attainment in mathematics, the advisory board also defined the content areas of mathematics as;

- Calculus.
- Statistics and probability.
- Number, quantity and arithmetic, including subitization, factors, comparison, fractions, symbolic and non-symbolic magnitude, exponents/roots, ordering (cardinality and ordinality), mental maths, and number sense.
- Ratio and Proportion.
- Shape, space and measures, including shape, special relations/reasoning, geometry, length, area and volume.
- Generalisation, including algebra.
- Proof.

Science

Scientific knowledge enables us to use what is known from experiments and scientific theory to explain what is going on in the world. Scientific literacy demands a combination of content, procedural and epistemic knowledge (OECD, 2019).

Similar to mathematics, in science there is a key distinction between what is known, and how it is known. For example, Harlen et al. (2010); (2015) distinguished between the ten big ideas *of* science and four ideas *about* science. The ideas of science are the content knowledge about specific scientific concepts and theories about the natural world and technology. These ideas of science provide a useful summary of key content areas for attainment in science;

1. All material in the Universe is made of very small particles. [Atoms, compounds and mixtures, matter – solids, liquids, gases]

2. Objects can affect other objects at a distance. [Light, sound, electrostatic forces, magnetism, gravity]
3. Changing the movement of an object requires a net force to be acting on it. [Forces]
4. The total amount of energy in the Universe is always the same but energy can be transformed when things change or are made to happen. [Energy]
5. The composition of the Earth and its atmosphere and the processes occurring within them shape the Earth's surface and its climate. [Earth science]
6. The solar system is a very small part of one of millions of galaxies in the Universe. [Space – sun, planet, universe]
7. Organisms are organised on a cellular basis. [Organisms and Cells]
8. Organisms require a supply of energy and materials for which they are often dependent on or in competition with other organisms. [ecosystems, photosynthesis, respiration]
9. Genetic information is passed down from one generation of organisms to another. [Genetics]
10. The diversity of organisms, living and extinct, is the result of evolution. [Evolution]

Distinct from this specific content knowledge, knowledge *about* science involves understanding scientific methods – how scientific enquiry creates scientific knowledge, by using scientific reasoning. Scientific reasoning and inquiry includes "questioning and generating hypotheses, experimenting, designing, and planning, predicting, modeling/ visualizing, observing and data collection, analyzing data, interpreting and explaining, developing/evaluating/arguing, reaching conclusions, and communicating findings." (Donnelly, Linn, & Ludvigsen, 2014, p. 573). It is using scientific ideas and processes to answer questions or solve problems. This is closely linked to epistemic knowledge, which refers to understanding of the rationale and justification for using these procedures (OECD, 2019). Harlen et al. (2010); (2015) summarise the following big ideas *about* science;

1. Science is about finding the cause or causes of phenomena in the natural world.
2. Scientific explanations, theories and models are those that best fit the facts known at a particular time.
3. The knowledge produced by science is used in some technologies to create products.
4. Applications of science often have ethical, social, economic and political implications.

In summary, literacy, mathematics and science are core subjects in the curriculum because these generalisable skills are essential for learning across disciplines and are therefore necessary to succeed in education more broadly. Indeed, skill and fluency in literacy, mathematics and science is essential in order to function effectively in modern society. Attainment in these subjects depends on a broad range of skills and knowledge. Literacy, mathematics and science are complex multi-dimensional constructs. Further, key concepts of attainment change over the course of development as these skills develop. Searching for measures of attainment in each key concept is beyond the scope of this review. Here, we seek to evaluate measures of overall attainment in each subject. Nonetheless, it will be important to distinguish between specific and general measures of attainment. Specific measures of attainment in literacy, mathematics or science 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.

What is the psychometric quality and implementation utility of the assessments identified through this review to measure attainment in students aged 6 to 18-years?

Systematic searches (see p9) will result in a list of existing assessments that measure attainment in these core subjects. We will then systematically review and evaluate each assessment by interrogating the following research questions:

1) Is the assessment a specific or general (multi-dimensional) measure of attainment in the target subject?
2) Is the assessment appropriate for the target population (UK, aged 6-18 years)?
3) Does the assessment conform to minimum psychometric properties (relating to reliability, validity and the quality of the standardised norms)?
4) Are there any special considerations for use of the assessment (implementation factors)?

Methodology

This protocol was developed by the recommendations from our advisory panel (see Table 9, p23), the COSMIN study (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 (Version 4.2.6 - Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013). A template PRISMA diagram is provided in Figure 1, and will be elaborated in the written synthesis. This protocol has also been independently peer reviewed by the Education Endowment Foundation.

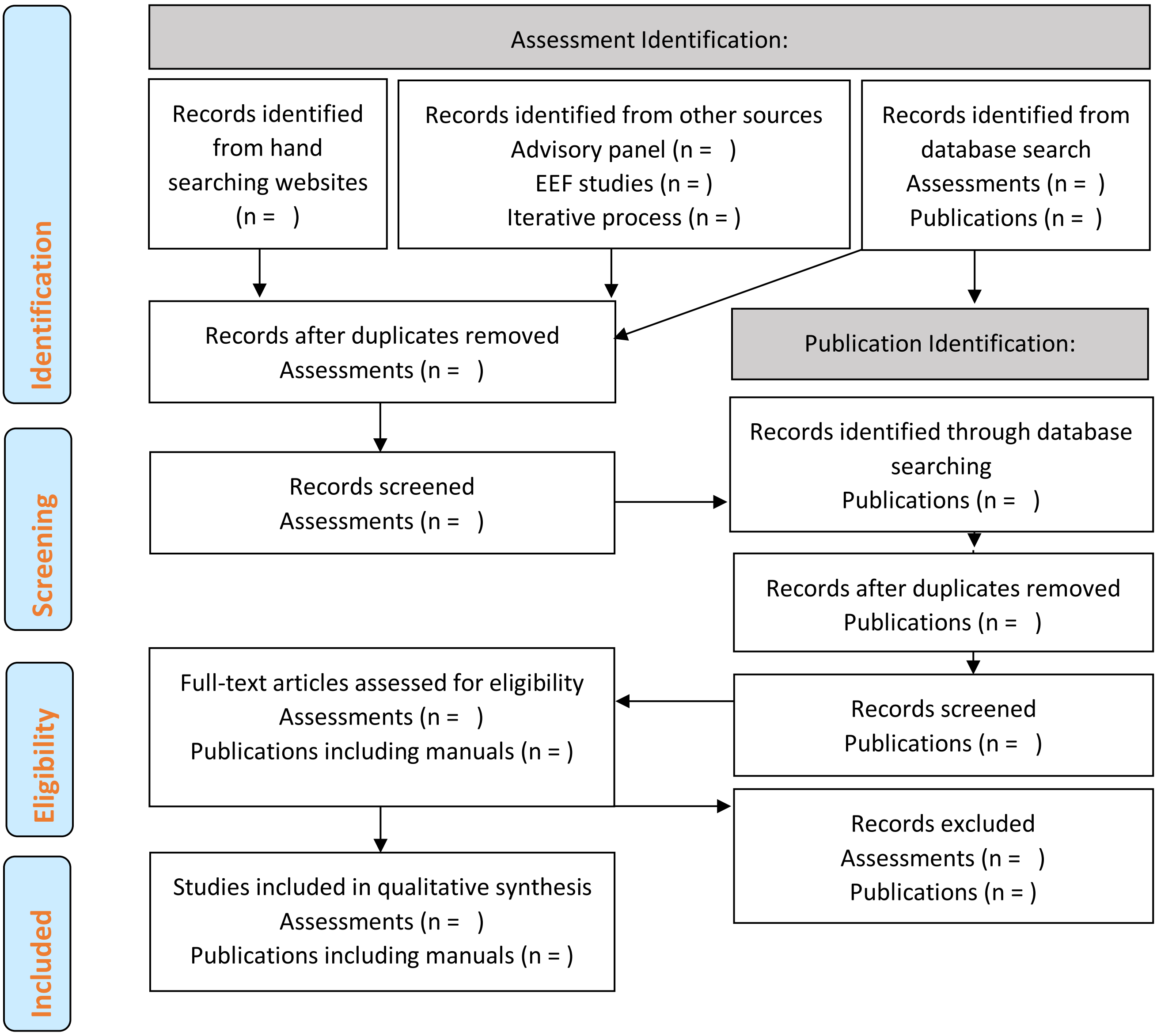
In the COSMIN study, a four-round Delphi method was used to develop a taxonomy and checklist to evaluate the methodological and measurement quality of health-related patient-reported surveys (see https://www.cosmin.nl/). The COSMIN taxonomy (Mokkink et al., 2010) provides a useful summary of the importance and utility of measures of reliability and validity, which we apply to the evaluation of the psychometric properties of assessments of attainment. The COSMIN risk of bias checklist (Mokkink, de Vet, et al., 2018) will be applied to combine reliability and validity information from multiple sources (e.g., from the administration/technical manuals for assessments and peer reviewed journal articles).

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 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, and selected questions from "Part 2 Evaluation of the Instrument" will be used to evaluate assessments in the final stage of the review (Evers, Hagemeister, et al., 2013; Evers, Muñiz, et al., 2013).

Inclusion and exclusion criteria for the review

A template PRISMA diagram for the systematic inclusion and exclusion of sources of information about assessments is provided in Figure 1, and will be elaborated in the written synthesis. The search strategy will identify a long list of assessments for inclusion. These assessments will be screened for relevance to the review (see Table 3 for criteria). After screening assessments, systematic database searches will identify peer reviewed publications about the psychometric properties of the assessments. Peer reviewed publications are then screened for relevance to the review (see Table 5 for criteria). Information from both test manuals and publications is then combined, and assessments are subjected to eligibility checks to identify whether essential information about the psychometric properties of the assessment is available to review (see Table 6 for criteria). Finally, we summarise key implementation factors (see Table 7) and systematically evaluate the psychometric properties of assessments (see Table 8).

Figure 1: Flowchart of selection criteria according to PRISMA 2009


Assessment Identification:
Records identified from hand searching websites (n =)
Records identified from other sources Advisory panel (n =) EEF studies (n =) Iterative process (n =)
Records identified from database search Assessments (n =) Publications (n =)
Records after duplicates removed Assessments (n =)
Publication Identification:
Records identified through database searching Publications (n =)
Records screened Assessments (n =)
Records after duplicates removed Publications (n =)
Full-text articles assessed for eligibility Assessments (n =) Publications including manuals (n =)
Records screened Publications (n =)
Records excluded Assessments (n =) Publications (n =)
Studies included in qualitative synthesis Assessments (n =) Publications including manuals (n =)
Identification
Screening
Eligibility
Included

Search strategy: assessment and publication identification

Assessment identification

Initial searches aim to create a “long list” of assessments, finding the name and acronyms of publicly available assessments of literacy, mathematics and science that are available in the UK. Note that national tests and qualifications (such as Key Stage assessments or GCSEs) will not be included in the database, because the content and norming varies over time. The database will contain minimal information about all assessments on the long list, as indicated in Table 1.

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

Criterion	Minimal information to include in the database	Exclusion criteria
Basic assessment information.	Name of assessment. Current version/Edition number. Name and acronym of previous/original version(s) of the assessment (if applicable). Subject (literacy/mathematics/science/generic).	**Does not meet search criteria.**

Assessments on the long list must be relevant to and suitable for the target subjects – 6 to 18-year-old pupils in the UK. Cultural and educational background is well documented to influence performance on standardised assessments (e.g., Walker, Batchelor, & Shores, 2009 reviews). Further, norm-referenced assessments are only suitable for use with individuals who are demographically similar to the normative sample. Hence, assessments are only included if they have been recently published and normed with a relevant sample. At this stage, “relevance” is loosely defined to allow identification of older assessments that have re-normed.

Search criteria to identify assessments 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 an assessment fulfils these criteria, the assessment will be included in the long list but may be filtered out during screening and/or eligibility checks.

Assessments will be identified by

- Comprehensive hand searches of publisher and distributor websites, indicated in Table 2. 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[3].
  - 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:

[3] Further searches of other databases were deemed unfeasible due to the quantity of information likely to be yielded. For example, an equivalent search of PsycInfo returned 13,250 articles.

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 (provided by the EEF in personal communication, 29/01/2020)
    - Recommendations from the advisory panel, who were asked to provide a list of any assessments commonly used in UK schools (for teaching or research purposes).
    - Using an iterative approach to identification of assessments, 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 assessments could be encountered when checking version history, during publication identification, or while reviewing concurrent validity. In which case, initial checks will be conducted to ensure that these assessments meet the search criteria above and further publication searches would be conducted. Assessments would then be subject to screening and eligibility checks before inclusion in the qualitative synthesis.

Table 2: Websites to hand search.

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 assessments, in all subjects, will be appraised using the same criteria. The next step is to establish whether sufficient information is available to subject assessments to evaluation. The database will include basic information about all assessments identified in the long list, but further information and evaluation will only be provided for those that pass the screening and eligibility criteria.

Screening Assessments

Minimal assessment information will be included for all assessments identified through the searches outlined on p11 (see Table 3). Following the recommendations of the EFPA review model, this information should be provided by publishers (Evers, Hagemeister, et al., 2013). A brief description of the test will be obtained directly from the publisher website (if available). At this point, no attempt will be made to rephrase the information provided nor will there be any attempt to identify key concepts included in the assessment. Reasons for screening an assessment are summarised in the "Exclusion criteria" column of Table 3 will be entered into the measures database.

Table 3: Screening criteria for assessments, minimal additional assessment information and summary exclusion criteria included in the database.

Criterion	Minimal information to include in the database	Exclusion criteria
Basic assessment information (additional information added during screening).	List of subscales (if applicable). Additional references/hyperlinks for other sources of information about the assessment (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 assessment[4]. Administration guidelines not available.	**Assessment is not available for review.**
Norm-referenced scores.		**Criterion-referenced.**
Suitable for age range (6 to 18-years).	Specific population and age range that publisher states the assessment is intended/suitable for. Key Stage(s) applicable to.	Assessment is not applicable to sample.
UK standardisation sample.	Yes/No.	**No UK standardisation available.**
Published or re-normed since 2010.	Publication date. Date of re-norming (if applicable).	**No recent norms available.**

*Note: Criteria highlighted in **bold** are new exclusionary criteria.*

[4] The hyperlink enables users to obtain additional information that may change over time, such as the cost of materials required for administration.

Publication identification

Subsequent searches aim to identify information needed to evaluate the psychometric properties of the measure. This includes information provided to users in administration and/or technical manuals supplied with the assessment, 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 assessments using information sourced from;

(a) Standard assessment and technical manuals provided to assessment users obtained from our own test library, subject librarians, publishers, distributors, and assessment authors
(b) Peer reviewed publications identified through systematic database searches. Search terms and limitations are described in Table 4, and are based on those recommended in the revised COSMIN recommendations (Mokkink, Prinsen, et al., 2018).

Table 4: Search terms and limitations for publication identification – information about assessments.

Database	Search terms	Limitations
PsycInfo	tests and measures: (Name of assessment OR acronym of assessment) 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 will initially be screened using the criteria in Table 5, based on the revised COSMIN recommendations (Mokkink, Prinsen, et al., 2018). Exclusion criteria will be recorded to indicate why publications were screened. Note, that we do not assess the methodological quality of the studies at this point.

Table 5: Screening criteria for publications about assessments

Criteria for inclusion	Exclusion criteria
Relate to at least one assessment on the long list.	Assessment screened.
Study aims to i) Evaluate one or more psychometric property (i.e., reliability and/or validity) of the assessment. ii) Develop a new assessment. iii) Evaluate interpretability of the assessment. Study <u>does not</u> merely use the assessment as an outcome measure. The following studies should be excluded iv) Randomised controlled trials. v) Studies where the assessment is used to validate another assessment.	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 an assessment to evaluate the psychometric quality of that assessment 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 an assessment 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 6 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 assessments are excluded because information needed for full evaluation cannot be obtained or is below threshold (i.e., removed during screening or eligibility checks), no further evaluation will take place[5]. Note, for example, criterion referenced assessments identified through this process will be documented in the database, but will be screened and therefore will not be evaluated[6]. The database will include why an assessment is excluded from full evaluation, using the 'exclusion criteria' indicated in Table 3 and Table 6. It is essential for both the evidence synthesis and the online

[5] 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).

[6] This may change if at a later stage it is felt that they should be evaluated. This could occur, for example, because criterion-referenced assessments dominate a subject.

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.

Table 6: Assessment 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-rate/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	

Evaluation and appraisal of assessments

The data collected at this point forms the criteria for evaluation of assessments. This includes more detailed information about implementation from the test manual (see Table 7) which will enable users to filter and short-list the measures, and an evaluation of the psychometric properties of the assessment using a broader range of sources (see Table 8).

Implementation factors

Information about implementation will be 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 assessments. Preference over implementation is variable and should be determined by the user. The implementation factors evaluated in Table 7 were selected in consultation with our advisory panel and largely align to Part 1 of the EFPA review model "Description of the instrument"[7]. All terms will be defined in the written synthesis.

Table 7: Evaluation stage – additional implementation information included in the database (not rated).

Criterion	Minimal information to include in the database	Response options
Basic assessment 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 assessment	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 assessment	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)

[7] Note, however, that we will 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.

Criterion	Minimal information to include in the database	Response options
	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
	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

Evaluation of psychometric properties

Evaluation of the psychometric properties (validity, reliability and quality of norms) of an assessment will be conducted using selected questions from the EFPA review model (Evers, Muñiz, et al., 2013). First, we will review 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, de Vet, et al., 2018). This enables us to effectively and objectively combine information gathered from both manuals and academic sources. This information will be summarised in the measures database as indicated in Table 8.

To evaluate the validity of an assessment, we will consider both construct and criterion validity.

- Construct validity examines the extent to which the assessment is an adequate measure of literacy, mathematics and science. This enables us to evaluate evidence for the theoretical underpinnings of the assessment, as well as the quality of and extent to which statistical evidence supports the view that the assessments measures the construct that it intends to measure. Questions from the EFPA review model culminate in an overall construct validity score from 0-4, which is an overall judgement rather than a simple average of scores (Evers, Muñiz, et al., 2013). A score of 0 indicates that it cannot be rated because of lack of information, 1 is inadequate, 2 is adequate, 3 is good and 4 is excellent. Hence, scores of three or above will be translated to a star in the measures database. In addition, we will consider to what extent does the assessments reflect the multi-dimensionality of the target construct (structural validity)?
- Criterion validity considers the extent to which assessment scores are related to scores on other established assessments of the construct. Of particular note are comparisons against national key stage tests. We will consider the nature of measures of validity (predictive, concurrent, post-dictive), the quality of the evidence and the strength of the relationship.

Questions from the EFPA review model culminate in an overall criterion validity score from 0-4 using the same scale as construct validity (Evers, Muñiz, et al., 2013). Scores of three or above will be translated to a star in the measures database.

Reliability explains the degree to which the assessment is free from measurement error. There are many different measures of reliability. The COSMIN taxonomy (Mokkink et al., 2010) summarises these as internal consistency, reliability and measurement error. Questions from the EFPA review model enable us to objectively combine and evaluate the quality of the available evidence of an assessment's reliability, resulting in a score of 0-4 using the same scale as construct and criterion validity (Evers, Muñiz, et al., 2013). Scores of three of more will receive a star in the measures database.

- Internal consistency refers to the interrelatedness of items in the assessment. Measures of internal consistency include internal reliability (the consistency of results across items within a test), and content or item sampling (the consistency of results subsets of items).
- Reliability refers to the proportion of total variance in performance on the assessment which is due to "true" differences between individuals. Measures include inter-rater/inter-scorer reliability (comparing scores by different people on same occasion), intra-rater/scorer reliability (comparing scores by the same person on different occasions), test-retest reliability/temporal stability (comparing scores after a short/long duration between testing), and parallel/equivalent forms reliability (comparing performance by the same person on different test versions on same occasion).
- Measurement error refers to the amount of systematic and random error in an individual's score which is not due to true changes in the underlying construct. Measures include the standard error of measurement (the spread of observed scores around true score) and the smallest/minimal detectable change (the amount of change in score that is meaningful and not simply due to chance).

Finally, evaluation of the quality of the available norms includes consideration of the sampling and representativeness of the norm-derived population (including sample size) to examine whether the norms are appropriate and free from bias. The EFPA review model does not provide an overall score (Evers, Muñiz, et al., 2013), hence in line with these recommendations we will note any biases in norming.

Table 8: 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	
Criterion validity	Predictive/Concurrent/Post-dictive validity: Does test performance adequately correlate with later, current or past performance?	0-4	Overall criterion validity score ≥ 3/4 = Star
	Summarise available comparisons [e.g., specify the measures compared to assess concurrent validity] and correlation [Value as reported in the test manual or from academic searches , with citation].	Free text	
Reliability	Is test performance reliable?	0-4	Overall reliability score ≥ 3/4 = Star
	Summarise available comparisons [e.g., specify the measures used to assess reliability] and correlation [Value as reported in the test manual or from academic sources, 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	

Dealing with missing data

If the number of assessments that enter the evaluation and appraisal stage is disproportionately low in a given subject (literacy, mathematics, or science), we will include implementation information (not rated, see Table 7) for excluded assessments, where possible.

If reliability and/or validity cannot be evaluated adequately because of a lack of data, a note will be made in the database to indicate that the absence of a star is a result of missing data.

Data synthesis

The database of assessments will contain 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.

The database will be accompanied by a narrative synthesis, which will include an executive summary, an introduction summarising the nature of the key concepts of attainment in literacy, mathematics and science, definitions of terminology used to evaluate the psychometric properties of the assessments, the methodology used to form the database and in evaluation of assessments, and a summary of the types of the assessments that were found, including a discussion of gaps in the availability of assessments.

Reporting

The written synthesis will include;

- Executive summary
- Introduction
    - Defining attainment, general comments on the role of assessment.
    - Review of attainment in each subject (literacy, mathematics, science) – definitions and description of key models.
    - Definitions of terminology used to evaluate the psychometric properties of the assessments (e.g., description of different measures of reliability and validity, how to interpret, why it is important to consider during test selection).
- Methodology of the systematic review and creating the database
    - Inclusion and exclusion criteria
        - Initial searches
        - First filter
        - Evaluation stage
    - Flow diagram indicating the number of assessments included and excluded at each stage for each subject.
- Summaries of assessments subjected to full evaluation – those that entered the appraisal and evaluation stage (as Dockrell et al., 2017, p. 35).
    - Separated by subject, and whether suitable for primary/secondary school use.
    - Proportion of measures evaluated as 4*/3*/2*/1*/0* for psychometric properties in each subject.
- Discussion.

    - Including identification of gaps in availability of assessments, and differences in psychometric properties across subjects.
    - Reflections on differences between subjects in the nature of assessments – availability and utility of specific verses general (multi-dimensional) assessments, and mapping key concepts onto assessments.
- Conclusion.
- Appendix: including as table listing all measures identified through the review process.

Personnel

Dr Helen Breadmore (PI, helen.breadmore@coventry.ac.uk), Associate Professor (Research) in Equity and Attainment, Coventry University will oversee the project and lead the narrative synthesis, ensuring realisation of the recommendations from the advisory panel throughout the review process.

Professor Julia Carroll (Co-I, julia.carroll@coventry.ac.uk), Professor in Equity and Attainment, Coventry University will co-author the narrative synthesis, and review information coded within the searchable database.

A research assistant will work closely with the PI and Co-I, following the procedures outlined in this protocol to gather the information necessary for Dr Breadmore and Professor Carroll to review and evaluate the measures.

The advisory panel (see Table 9) is formed of experts in the fields of literacy, mathematics, science, assessment design and evaluation. The advisory panel supported the development of the protocol.

Table 9: Advisory panel members.

Name	Job title	Affiliation
Katie Baker	Specialist mathematics teacher, PhD (submitted) evaluating 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.

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

Timeline

Dates	Activity	Staff responsible/ leading
21/02/2020	Advisory panel review first draft of protocol, advisory panel meetings conducted, second draft of protocol delivered to EEF for review.	Helen Breadmore
06/03/2020	EEF return second draft of protocol with comments.	Diotima Rapp
27/03/2020	Final draft of protocol (with further definitions of key concepts in science) delivered to EEF.	Helen Breadmore
26/06/2020	Draft Evidence Synthesis sent to EEF for review.	Helen Breadmore
24/07/2020	Database content sent to EEF for review.	Helen Breadmore
31/08/2020	EEF return comments on evidence synthesis and database content.	Diotima Rapp
30/09/2020	Finalised evidence and measures database delivered.	Helen Breadmore

Reference list

Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). *Properties of commercial tests in the EEF database*. Retrieved from London, UK: https://educationendowmentfoundation.org.uk/public/files/Support/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf

Breadmore, H. L., Vardy, E. J., Cunningham, A. J., Kwok, R. K. W., & Carroll, J. M. (2019). *Literacy Development: Evidence Review*. Retrieved from London: https://educationendowmentfoundation.org.uk/public/files/Literacy_Development_Evidence_Review.pdf

Chall, J. S. (1983). *Stages of Reading Development*. New York: McGraw-Hill.

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y. W., & Cordier, R. (2017). Psychometric Properties of Language Assessments for Children Aged 4-12 Years: A Systematic Review. *Front Psychol, 8*, 1515. doi:10.3389/fpsyg.2017.01515

DfE. (2014). *The national curriculum in England: Framework document*. London, UK Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/381344/Master_final_national_curriculum_28_Nov.pdf

Dockrell, J., Llaurado, A., Hurry, J., Cowan, R., Flouri, E., & Dawson, A. (2017). *Review of assessment measures in the early years: Language and literacy, numeracy and mental health*. Retrieved from London, UK: https://educationendowmentfoundation.org.uk/public/files/Review_of_assessment_measures_in_the_early_years.pdf

Donnelly, D. F., Linn, M. C., & Ludvigsen, S. (2014). Impacts and Characteristics of Computer-Based Science Inquiry Learning Environments for Precollege Students. *Review of Educational Research, 84*(4), 572-608. doi:10.3102/0034654314546954

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P., Muñiz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests: Test review form and notes for reviewers*. Retrieved from Brussels:

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: revision of the EFPA review model. *Psicothema, 25*(3), 283-291. doi:10.7334/psicothema2013.97

Harlen, W., Bell, D., Devés, R., Dyasi, H., Franández de la Garza, G., Léna, P., . . . Yu, W. (2010). *Principles and big ideas of science education*. Retrieved from https://www.ase.org.uk/bigideas

Harlen, W., Bell, D., Devés, R., Dyasi, H., Franández de la Garza, G., Léna, P., . . . Yu, W. (2015). *Working with big ideas of science education*. Retrieved from https://www.ase.org.uk/bigideas

Hodgen, J., Foster, C., Marks, R., & Brown, M. (2018). *Improving mathematics in Key Stages 2 and 3: Evidence Review*. Retrieved from London, UK: https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/improving-mathematics-in-key-stages-two-and-three/

Mokkink, L. B., de Vet, H. C. W., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. *Qual Life Res, 27*(5), 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., & Terwee, C. B. (2018). *COSMIN methodology for systematic reviews of Patient—Reported Outcome Measures (PROMs): User manual*. Retrieved from 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. *Qual Life Res, 19*(4), 539-549. doi:10.1007/s11136-010-9606-8

National Research Council. (2001). *Adding It Up: Helping Children Learn Mathematics*. Washington, DC, USA: National Academy Press.

Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R., & Miller-Friedmann, J. (2017). *Review of SES and Science Learning in Formal Educational Settings: A report prepared for the EEF and the Royal Society* Retrieved from London, UK: 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*. Retrieved from https://www.oecd-ilibrary.org/education/pisa-2018-assessment-and-analytical-framework_b25efab8-en

Terwee, C. B., Prinsen, C. A. C., Chiarotto, A., Westerman, M. J., Patrick, D. L., Alonso, J., . . . Mokkink, L. B. (2018). COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res, 27*(5), 1159-1170. doi:10.1007/s11136-018-1829-0

Walker, A. J., Batchelor, J., & 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), 216-223. doi:10.1080/00050060902833469