The Routledge Handbook of Teaching English to Young Learners

Sue Garton, Fiona Copland

Assessment of young English language learners

Publication details
https://www.routledgehandbooks.com/doi/10.4324/9781315623672-25
Szilvia Papp
Published online on: 10 Sep 2018

PLEASE SCROLL DOWN FOR DOCUMENT

<div align="right">24</div>

Assessment of young English language learners

<div align="right">*Szilvia Papp*</div>

Introduction and definitions

Assessment of young learners' English as a second or foreign language has 'come of age' and matured into a field of enquiry with its own identity and integrity (Rixon 2016). It has its own questions, concerns and methodologies relevant to a wide range of highly involved stakeholders: ministries of education, assessment providers, publishers, school leaders, teacher trainers, teachers, parents and students.

The field has moved on from an assessment of general language proficiency to include the assessment of the language of schooling (academic English) and the integrated assessment of content and language learning (Bailey and Huang 2011; Inbar-Lourie and Shohamy 2009; Nikolov 2016). Several approaches and methods populate the field, reflecting variation in constructs and their measurement. What type of assessment is most effective and beneficial for young learners depends on their age, context of instruction, amount and type of exposure to English, purpose of assessment and use of results.

This chapter reviews some of the current research on instruments used in classroom assessment and large-scale national and international tests of English developed for young learners. Knowledge about alternative approaches to assessing young learners' English language competence may help stakeholders (parents, teachers, policy makers) make informed decisions on what assessment is appropriate in their context for the young people in their charge and how to use results generated by such assessments to make sound decisions.

Who are young learners of English?

The label 'young English language learners' has been used mainly for primary/elementary school age children who learn English as a second or foreign language. However, it sometimes encompasses adolescents in lower secondary/middle school contexts, as well as very young learners in early years or kindergarten settings. As a result, the age range designated by the label of 'young learners' may vary between three and 16. Within this wide age range, there are milestones in cognitive, linguistic, social and emotional development, which stages of schooling tend to recognise and build on. However, large variations

in educational systems exist in terms of start of compulsory schooling and start and nature of English language instruction. Even wider differences exist among stakeholders' views on whether and how young learners' English language development should be assessed. In this chapter, we take the widest definition of 'younger learners' (within the range of 3–16 years of age) and will bring examples of assessment from pre-school, primary and lower secondary school contexts.

What is assessment?

Educational assessment is used to identify levels of ability within a target population in a particular learning domain such as English language competence in order to distinguish between strong and weak performance. Masters (2014) argues that there is only one purpose of assessment: to find out individual learners' current standing in a learning domain. However, the results may be put to various use, such as readiness checking, diagnosis, screening, placement, selection or certification. The ultimate aim in generating assessment information is to answer important and well-articulated questions about young learners' English language learning in order to make sound educational decisions. Assessment thus involves gathering and analysing evidence to make valid inferences about learning and teaching. Importantly, the evidence collected is always a sample of all the information that could be collected, and decision making will need to take this into account. Typical questions for which assessment data is used include:

- Which learner is ready to proceed, which needs help/support and which needs additional challenge?
- What are the levels of progression on the learning ladder?
- What are achievable targets for young learners in various contexts?
- What is the minimum standard for a particular purpose, such as learning content (maths, science, history, geography) through the medium of English?
- What are the skills profiles of young learners?
- Where are the achievement gaps?
- What does each learner/class/school/region/country need in terms of appropriate next steps?
- How does this assessment predict future achievement (e.g., performance in further assessments or in future study)?
- Which school needs improvement? Which one needs additional challenge to excel further?
- What are the conditions of success as demonstrated by best practice in learning and teaching?

Assessment data at the individual level is used to find out about readiness to learn or proceed, current achievement, rate of progress or potential future performance of individual children. At group level, assessment data is sometimes used for establishing a baseline to develop a new strategy or for benchmarking against an external (national or international) standard. Historical assessment data is used for information on learning gains, growth and trends over time.

A traditional dichotomy is usually made between internal versus external assessment in terms of purposes and use of results. It is usually assumed that classroom assessment is carried out by teachers to make sound pedagogical decisions, while large-scale tests are usually

used for policy decisions such as accountability or gatekeeping purposes. Whether internal or external, the differences in the intended use of assessment results will have implications for reporting: whether diagnostic feedback on strengths and weaknesses at the task/skill level is reported or information on overall proficiency is provided. Assessment results can be used to monitor progress, plan future action or predict future performance of individuals and groups (usually referred to as formative use of assessment data) or evaluate the effectiveness of teaching and learning in programmes by individual learners, teachers, classes, schools, regions or nations (usually termed summative use of assessment data).

It follows that assessment design and reporting will vary according to whether the intended use has more formative and developmental aspects that require a domain, skill or task-centred interpretation with domain-, skill- or task-specific criteria (achievement) or whether the intended use is a construct-centred summative interpretation in terms of generic criteria and standards (communicative competence or proficiency). Reporting should match the purpose of the use of assessment results (e.g., readiness checking, diagnosis, screening, placement, monitoring progress, selection, certification) (Moss 2015).

In interpreting the results, the frame of reference can be the performance of other students of the same age, in the same class or other classes, schools, regions or nations (norm referencing). Alternatively, the frame of reference can be external standards, benchmarking frameworks or curriculum expectations specifying certain learning outcomes in terms of target knowledge, skills and abilities (criterion referencing) or the student themselves in their earlier performance (ipsative referencing). Whether the purpose of using the results is instructional, evaluative or predictive, the overarching aim of all assessment should be to improve learning outcomes by increasing student motivation and to ensure positive impact. It follows that the stakes of all assessment with young people are always very high.

Teachers use a variety of information (e.g., observation of learner classroom performance, periodic teacher-made or textbook tests at the end of unit of learning, term or year, portfolio of classwork, homework) to monitor students' progress in English language learning. Once recorded, these informal assessments can be used to check if a set of learning outcomes have been achieved and to provide feedback to inform subsequent teaching and learning. Teacher-based assessment is also used in some national curriculum testing regimes for summative purposes, such as in the UK. In large-scale international tests, trained examiners assess the performance of candidates against specific criteria and standards. Learning aims and assessment criteria can also be used for self- and peer-assessment for formative purposes.

Historical perspectives

In the last 60 years or so, a wide variation in instructional contexts and educational purposes have emerged in response to different English language learner needs that require a range of approaches to teaching and assessment. Learning objectives and assessment outcomes depend on the role English language plays in instruction: whether English is used as a vehicle for learning other subjects, as in full immersion, in various types of bilingual/trilingual schooling or in content and language integrated learning (CLIL); or whether English is a subject to be learnt as in traditional modern foreign language classes (Bailey and Huang 2011, Bailey et al. 2014, Inbar-Lourie and Shohamy 2009, Murphy 2014). In the following we discuss assessment instruments used in contexts where students learn English as a foreign language and also contexts where learners receive all or part of their education in English.

Szilvia Papp

Assessment in English as foreign language (EFL) contexts

While national tests of general English language proficiency have existed for a long time as secondary school leaving exams, large-scale international assessments for younger English language learners have appeared only in the last 35 years.

Pearson's Test of English Young Learners formerly known as London Test of English for Children has been in existence since 1982. The Cambridge English Young Learners tests were launched in 1997, the for Schools version of Cambridge Key and Preliminary was introduced in 2009 and Cambridge First for Schools followed in 2011. The Pearson and Cambridge tests measure general English language competence. Both Pearson and Cambridge tests have explicit exam syllabuses specifying what vocabulary and grammatical structures learners need to have mastered for successful achievement.

In the USA, Educational Testing Service (ETS) developed TOEFL Junior in 2011 and TOEFL Primary in 2012 to expand their family of assessments. With no clearly specified exam syllabus, TOEFL Primary provides 'information about the English proficiency of young English learners in countries where English is not typically used in daily life' (ETS 2015a, 2015b). On the other hand, TOEFL Junior measures English language proficiency needed in English language instructional contexts (So et al. 2015), which is reflected in the construct and content of the tests. Existing international tests for under six-year-olds are Trinity Stars and Anglia's First Step.

In these assessments, children's oral skills are assessed face-to-face (individually, in pairs of groups) with an examiner (or two) and using traditional paper-and-pencil tests for listening, reading and writing. Alternatively, there are computer-based versions, such as those for all Cambridge English exams for children and teenagers, the Oxford Young Learners Placement Test or British Council's Aptis for Teens. The mode of delivery will define the construct and inevitably have a backwash effect on teaching. In these tests, the stakes vary at different levels, with the pedagogical motivational purpose at the lower levels overtaken by arguably higher stakes of certification at higher levels.

Tests for young learners tend to measure the four skills of listening, speaking, reading and writing and contain items and task types (closed/selected and open/constructed response items) similar to tests developed for adults. Well-designed restricted response items (multiple choice questions, short answer questions, matching) can measure knowledge at word and phrase levels but also some higher order skills at sentence, text and discourse levels. Their use in assessments is not out of line with modern theories of learning, as there is a place for logical reasoning, use of analogy and elimination and even informed guessing from context in current cognitive psychological theories. Single best answer questions used in medical education are an example of how well-constructed MCQs can measure higher order thinking skills such as problem solving and application of knowledge. However, authentic direct tests of performance have the highest fidelity to the real world and therefore greater relevance to learners' lives. The respective weighting of restricted and constructed response items in a test for young learners should be dictated by the purposes for which test results are used.

Assessment in English as a second language/ content-based instruction

In English as a second language (ESL) contexts (e.g., USA, Australia, Canada), assessment of English language competence is part of the national standards-based educational system. In these contexts, minimum standards are specified in English language development/

proficiency frameworks of competences (e.g., WIDA 2012, McKay et al. 2007). In the USA, assessment of both content knowledge and language development is carried out in substantive ways. Schools are held accountable for reaching the minimum standards; therefore assessments are subjected to close scrutiny (Bailey and Carroll 2015). Typically, in these frameworks, the learning domain is mapped and progression is charted using a learning ladder. The WIDA framework integrates age/level of schooling with language and academic content in core subject areas (maths, science, social studies and language arts). These standards help teachers understand and assess the required academic language skills for each core subject at each level of the curriculum. If such a map is based on empirically validated developmental sequences and learning trajectories, it can chart out critical paths for high achievers. In the UK, the document linked to the previous National Curriculum entitled 'National Curriculum 2000 A Language in Common: Assessing English as an Additional Language' (QCA 2000) dealt with the four skills but not the academic language requirements in primary and secondary schools. A national framework to be used with EAL learners in the UK has recently been developed (Evans, Jones et al. 2016a) to more accurately assess the needs of English language learners and guide their teaching in both language and content areas (Arnot et al. 2014, Evans et al. 2016b).

Assessment in content and language integrated learning (CLIL)

Content and language integrated instruction (CLIL) is an increasingly popular methodology within Europe and beyond, typically at primary and secondary levels, where the L2, usually English, apart from being a target language to be learnt as a school subject, is also used as a vehicle to deliver content knowledge and target domain-specific skills. CLIL describes any learning activity where an additional second or foreign language is used as a tool to develop new learning in a subject area or theme. CLIL can range from total immersion to a single subject or topic taught in the L2. It may help maintain motivation of children who start learning English as foreign language at a young age (Nikolov 2016). CLIL theory has clear links with general education theory and modern cognitive and assessment theories (e.g., reference to lower/higher order thinking skills, balancing of cognitive challenge with linguistic support, differentiation in assessment and continuous assessment as integral part of instruction providing feedback to inform subsequent teaching and learning). However, true integration of language and content is still rare and remains a major issue in both instruction and assessment (Massler et al. 2014). It is recognised that assessment is fundamental to the success of CLIL. However, how teachers assess progress and attainment in CLIL is still something of a 'blind spot' (Massler et al. 2014, p. 137). This is due to lack of clear policy decisions on assessment in CLIL and the scarcity of assessment tools. Successful CLIL implementation calls for continuity across school levels; thus coherent assessment principles and procedures are needed that bridge educational stages (Stotz and Megías 2010). Llinares et al. (2012, p. 280) point out that to be useful and beneficial, assessment has to be an integral and indispensable part of instruction that should be planned before any teaching takes place.

In CLIL, especially at the beginning stages and at lower levels of language and academic proficiency, short-term learning goals should be set and assessed to build student confidence. It is necessary to design a variety of cognitively appropriate instructional tasks with clear assessment criteria for their achievement: challenging but not too frustrating for the learners, graded in terms of difficulty. Individual feedback needs to be provided on whether the student has achieved the outcomes or not.

Also, in CLIL, a range and variety of assessment tools is recommended. This should give the students greater confidence and provide more reliable data, because it can measure individual progress and check a wider range of competences and desirable learning outcomes. A range of assessment tools can consist, among others, of graded mini-assessments of each building block for each learning outcome, visual representations of students' understanding of content (tables, graphs, visual organisers), self- and peer-correction, self- and peer-assessment and portfolios. If feedback is provided on all these, assessment can usefully guide learning of both content and language. It is very important that the scores on each of these assessments are not averaged to derive a final score, as that would break all the principles of formative developmental assessment. A student's final assessment could be based on their progression in tackling similar content and tasks aimed at one learning goal or closely related set of leaning goals, and their ability to take feedback on board, reflecting their final achievement of the learning objectives.

Surveys

With this wide array of instructional contexts and purposes, a number of questionnaires and empirical surveys have been initiated to find out the state of play in young learners' language assessment worldwide. Early on, Rea-Dickins and Rixon (1999) found that primary teachers used internal assessment by way of paper-and-pencil tests in spite of the universal declaration that speaking and listening were teaching priorities. The focus on written assessment in primary schools mirrored the tradition in secondary schools where instruction and assessment has traditionally emphasised formal language study and reading and writing skills. Performance assessment of communicative language ability has been a recent addition in a lot of assessment regimes (Rixon 2013).

Since the advent of standards-based assessment, many countries have set explicit target attainment levels for the end of primary and secondary schooling. Selecting or developing instruments to assess whether targets have been reached has been the focus of intense effort in many countries with widely differing assessment cultures ranging from a largely egalitarian view (e.g., Norway, Carlssen 2008) to much more competitive examination-focused cultures (e.g., Butler and Lee 2010; Carless and Lam 2014). Some South American and European countries have developed their own national EFL examinations for young learners (e.g., in Uruguay, Fleurquin 2003; Norway, Hasselgreen 2005b; Germany, Rupp et al. 2008; Slovenia, Pizorn 2009; Switzerland, Haenni Hoti et al. 2009; Hungary, Nikolov and Szabo 2012; Poland, Szpotowicz and Campfield 2016).

Recently, two large scale empirical surveys have been carried out on attainment levels at the end of primary and secondary schooling in Europe (Enever 2011, European Commission 2012). These studies combine assessment results with questionnaire data in order to identify variables that contribute to high attainment in foreign language learning.

Critical issues and topics

Ethics

As the review above indicates, young English language learners have become the focus of intense attention, and the stakes in assessing them have become higher. Therefore, ethical considerations should be at the forefront of all assessment activity involving young learners. Ethics of assessment is a branch of philosophy dealing with issues of right and wrong

decisions and actions; it is a synonym for morality. Codes of conduct offer ethical guidelines about professional responsibilities and accountabilities (e.g., BERA 2011). Assessment should only be carried out to do good and for the benefit of learners, not for surveillance or the exercise of power that may harm young learners either directly or indirectly through their effect on teachers, schools, curriculum, educational systems or society. When conducting research with or assessment on children, children's rights must be observed (United Nations 1989). Criteria to evaluate the ethics of assessments for young learners include (a) whether assessment is in the children's best interests, (b) whether it is universal in that it allows equal opportunities to learning and access to assessment, (c) whether it attends to matters of diversity and individual difference and (d) whether it allows children's voices to be heard (Elwood 2013, Pinter 2011, 2014).

Desirable test qualities

One of the major responsibilities of assessment providers is to make sure tests for young learners have desirable test qualities, validity, positive impact, reliability.

Validity

Based on a review of the literature on validity, including the work of scholars (e.g., Messick 1989, Frederiksen and Collins 1989, Kane 2013), the *Standards for Psychological and Educational Testing* (AERA et al. 2014) and documents developed by assessment providers (e.g., Cambridge English 2013, SQA 2015), an assessment can be said to be valid when it fits the following criteria:

- Is appropriate for its purpose.
- Is a catalyst for curricular, instructional change and improves learning.
- Allows candidates to show that they have the required knowledge, understanding and skills to demonstrate the assessment outcomes, assessment standards or performance criteria.
- Allows all assessors to make reliable assessment decisions.
- Allows the interpretation and inferences which can be drawn from the scores/grades to be meaningful, useful, appropriate and justifiable.

A valid use of an assessment or its outcomes is when decisions made are sound and follow-up actions are justified, closely linked to the original intended purpose and supported by the results.

Impact

Impact of an assessment on learners can be gauged by investigating fitness for purpose, e.g., how well it motivates learners to learn English and/or how well it prepares them for the next level of study. Motivation is a central critical issue in the reporting of results in young learners' assessment, as children and adolescents tend to suffer from test anxiety and may get demotivated by assessment results. Tests can have devastatingly negative impact on learning and young learners' future prospects (Carreira 2012, Kim and Seo 2012). It is very important that assessments do not alienate them from taking tests in the future and ultimately from learning English.

Szilvia Papp

In addition, assessments, when misused or abused, can have negative effects on teachers, educational systems and society. Two recent large-scale global surveys carried out by Cambridge English (Papp et al. 2011, Papp and McElwee 2015) found that there is uncertainty, concern, fear or even distrust of tests developed for young learners among some stakeholders. Those who oppose testing young learners warn of the danger of over-testing children. Some are concerned that increasing standards of achievement may be demanded of young learners, which will entail more pressure in terms of competition that might generate a fear of failure, especially among the weaker learners. There is a distrust of large-scale testing as, if used for accountability purposes, it may foster test-orientedness (teaching to the test) among teachers, which might be linked with a loss of enjoyment and interest in teaching and learning. Some test users fear that results may not reflect the true ability of learners. Some see a risk that standards might ultimately be lowered as a result of a focus on accountability. Diagnostic testing is felt to be more appropriate in order to cater to young learners' individual needs and for the provision of feedback to improve learning and teaching among them. Assessment providers should be committed to carrying out ongoing research on the consequences of using assessments in young learners' education.

Reliability

Reliability relates to how much confidence users of tests can have in the results, in terms of the accuracy of test scores or consistency of classification.

It is commonly believed that with the fundamental requirement to be motivational, young learners' tests may not be psychometrically optimal (Jones 2002). Lower reliability may be due to the high facility of these tests, which means that they are designed so that most candidates can answer most items correctly, resulting in a skewed distribution. In addition, data generated by young learners may not be reliable as children are easily distracted and affected by physical and mental variations. It is very easy to get children to fail items by using the wrong assessment type or method in research or assessment. For instance, young learners may not be familiar with the content of the test, choose responses for idiosyncratic reasons or may be confused by test instructions. This is another reason why it is good practice to collect data from young learners at several different occasions and in several different ways from several sources. It is imperative that assessment of young learners' performance does not punish unexpected but clearly ingenious responses because that will stifle children's creativity (Cameron and McKay 2010). This principle should be reflected in scoring and assessment criteria.

Studies containing technical qualities or measurement properties of tests developed for young learners are rare. The same applies for many national tests designed for primary and secondary school aged learners. Even less information is available on teacher-made and textbook tests.

Current contributions and research

Alignment with external frameworks, benchmarks and standards

International standards such as the Common European Framework of Reference (CEFR, Council of Europe 2001) offer a way to set attainment targets and establish and compare standards among various assessment instruments. The CEFR's positive 'can do' approach is well suited to the principles of assessment of young learners. International tests aligned

with external frameworks such as the CEFR can be used as an aid to increase standardisation of learning and teaching, which may lead to improved teaching and higher levels of proficiency.

Major international large-scale tests of English for young learners, such as Cambridge English Young Learners and for Schools exams, Pearson Test of English Young Learners, TOEFL Primary and Junior, Oxford Young Learners Placement Test and British Council APTIS for Teens, all claim alignment to the CEFR levels, with some reporting formal alignment procedures (Papp and Salamoura 2009, Baron and Papageorgiou 2014). However, until the CEFR is adapted with young learners' needs and development in mind, all such linkage is tentative (Hasselgreen 2005a, McKay 2006, Papp and Salamoura 2009, Enever 2011). Nevertheless, in most countries, CEFR A1 and A2 levels have been set as a target of primary schooling (Rixon 2013). One of the authors of the CEFR, John Trim, has stated that in EFL settings in Europe:

'As a very rough guide,

- A1 (Breakthrough) is appropriate to progress in the first foreign language at the 10 or 11 year primary/secondary interface,
- A2 (Waystage) to around 14,
- B1 (Threshold) to 16+, the lower secondary goal,
- B2 (Vantage) to 18+, the completion of upper secondary education, and
- C1 and C2 to specialist university level'. (Trim 2005, p. 4)

In the face of the diversity inherent in international language learning, Jones and Saville (2009, p. 37) have argued, 'our default expectation must be that different countries' interpretations will be culturally determined (in a broad sense) and therefore may differ'. Indeed, Trim (2001/2009, p. 6) recalls how, when working on the Breakthrough/CEFR A1 specifications with groups of educators from a range of European countries, 'it became clear from different specimen descriptions that very different interpretations of words like "simple", "basic", "familiar", etc. were possible'.

How far young learners can progress in L2 proficiency due to linguistic, cognitive, emotional, social and literacy development was investigated by Hasselgreen and Caudwell (2016) for the British Council. In their analysis, the highest potential levels attainable by different age groups are as shown in Table 24.1. These represent much more ambitious targets than Trim's estimates above.

It is important to note that both of these sets of targets were not based on empirical research on young learners' language development in specific contexts. As Trim argued, *'I say a rough guide since the speed of learning depends greatly on such factors as the learner's age and aptitude, the curricular time available, extra-curricular contact, the relation of*

Table 24.1 Target CEFR attainment levels for different age groups (Hasselgreen and Caudwell 2016, p. 34)

Age groups	Typical limits of CEFR levels potentially attainable
Young children (roughly between 5/6 years and 8/9 years)	A2
Older children (roughly between 8/9 years and 12/13 years)	B1
Teenagers (roughly between 13 and 17 years)	B2
Exceptional older teenagers	C1

Szilvia Papp

Table 24.2 Cambridge English guidance on learning hours

CEFR Level	Cambridge English Exam	Number of Hours (approximate)
C2	Cambridge English: Proficiency (CPE)	1,000–1,200
C1	Cambridge English: Advanced (CAE)	700–800
B2	Cambridge English: First (FCE)	500–600
B1	Cambridge English: Preliminary (PET)	350–400
A2	Cambridge English: Key (KET)	180–200

L1 to L2, etc.' Trim (2005, p. 4). How long it takes to go from one level to the next depends on the factors pointed out by Trim as well as the intensity and quality of instruction, learner motivation, etc. However, a very general rule of thumb is that 180 hours are required to move within the A levels, 200 hours at B levels and 220+ at C levels (see Table 24.2).

Little (2007) pointed out that above the B-levels the CEFR attainment levels are probably not achievable for learners below the age of 16, as they require high levels of educational experience with the associated tasks that most young learners have not had exposure to. There is consensus that learners below 16 lack cognitive and social maturity that the tasks in the C-levels require (Hasselgreen and Caudwell 2016, Goodier and Szabo 2017).

The Finnish National Certificates of Language Proficiency (2011) uses the CEFR to identify targets for English and Finnish and Swedish as L2. The proficiency scale in the Finnish framework uses the levels of the CEFR and targets are set for both English and national languages. For instance, the University of Helsinki website states

> the foreign language requirement for a lower university degree is B2 in English/B1 in other languages. The minimum requirement for Finnish/Swedish as a second national language is B1.

Härmälä et al. (2015) found:

- Students at the age of 12/13 are required to have a minimum level of language competence mirroring B1 in all skills in order to succeed in history/mathematics.
- 15/16-year-old students need a B2 competence in the same skills/subjects.

These ambitious but realistic targets are corroborated by assessments created for L2 learners who live and learn in an environment where English is spoken. Cambridge Lower Secondary English as L2 by Cambridge International Examinations (CIE) specifies high B1/low B2 targets for 11–14 year-olds. Shaw and Imam (2013) also suggest that CEFR B2 represents a critical level for 16-year-old learners who are assessed through the medium of English in subjects such as history. They found that linguistic range and accuracy at B2 level are essential, but some C1-level skills provide added advantage. Especially influential are the written cognitive-academic skills (e.g., Thematic Development, Propositional Precision, Coherence and Cohesion, Overall Written Production, Text Processing). However, Goodier and Szabo (2017, p. 16) point out that written skills in a foreign language environment need specific support:

> The treatment of descriptors relating to written reception, production and integrated skills therefore should take a 'bias for best' approach, assuming what is reasonable/ possible for the age range given optimum literacy support.

Various sets of level descriptors have been collected by assessment boards, either reflecting typical or likely performance of candidates at a target level (Papp 2009), from teachers (Pearson Education 2015, Benigno and de Jong 2016) or based on specific exam content (So et al. 2015). Each of these sets of can-do statements are context specific in the way they were developed and validated. With the extended set of descriptors for young learners (aged 7–10 and 11–15) collated by the Council of Europe (Goodier and Szabo 2017), it is now easier and more meaningful to align young learners assessments with the CEFR and set attainment targets more directly relevant for each age group.

Accountability

Assessment of young learners is increasingly used as a policy instrument for accountability purposes or to evaluate educational reform. One such use of assessment is to establish a baseline against which achievable targets can be set, growth can be measured and evaluated and standards maintained. For plans on introducing such a baseline assessment in the UK among four- and five-year-old children, see Bradbury and Roberts-Holmes (2016). It is clear from the reaction to these plans, that to be useful and beneficial, young learners' assessments must be very well conceived and should not be used as surveillance, judgement or to exert power. Their use should be with the express aim to bring about improvement in learning. It should be ensured through reporting, use of results, decision making and follow-up action that the effect of assessment is transformational, productive and empowering (Earl 1999, Klenowski and Wyatt-Smith 2014). Assessment providers have a moral obligation to work toward this aim.

Examination boards are responding to calls for transparency and accountability. For instance, Cambridge English is about to publish an account of the Cambridge approach to assessing young learners aged between six and 16 within the school contexts (Papp, and Rixon Field forthcoming). The volume, within the Studies in Language Testing series, will set out the theoretical foundations, language competence model, development and validation framework within Weir's (2005) sociocognitive model, and test specifications to provide evidence for the validity of Cambridge English's range of assessments for children and teenagers. Cambridge English also works with ministries and governments on various education reform projects across the school sectors (see Cambridge English Case Studies 2015).

In the USA, there is a similar response to a need to show accountability. ETS has set out an extensive research programme on the development of the TOEFL Primary and Junior tests, providing systematic evidence for their validity within Kane's (2013) interpretive/assessment use argument. On the other hand, Pearson has been developing the Global Scale of English Learning Objectives for Young Learners (Pearson 2015). These developments prove that the field of young learners English language assessment has truly come of age.

Recommendations for practise

What is good assessment for young learners?

Good assessment measures 'critically explored and clearly defined constructs' (Daugherty 2012). Test developers for young learners must justify what knowledge, skills and abilities are assessed and clearly state the intended purpose of test use. Test score interpretations should be made very clear so that test users can make the right inferences from the results. There should be recommendation for legitimate uses of test results and some examples

Szilvia Papp

of illegitimate interpretations of score data and use of test results. Test developers should encourage informed and responsible use of results among stakeholders.

Since learning is central for children, teaching and assessment should foster meaning making through language, both in their L1 and L2 English. Therefore, the centrality of concept formation, critical thinking and problem solving through language should take priority in young learners' assessment (Butler 2016). This is a fundamental principle in CLIL and has also been emphasised in current thinking on the future of assessment. However, focus on form should not fall prey and traditional standards of accuracy and fluency should also be promoted.

Good assessments for young learners manage to balance two seemingly opposing requirements: cognitive challenge with the right amount of support. Tests that take into account young learners' cognitive, social and emotional development are learner friendly and offer a positive experience to candidates. Useful assessments help learners prepare for the 'real world', either in terms of promoting real-life abilities or general learning to learn skills or soft interpersonal or intrapersonal skills.

When it comes to the ethics of assessing young learners, best practice strives for fairness, equity and equal access. Quality large-scale tests should be equally accessible to all candidates – geographically, financially and in terms of special needs.

Assessment literacy

Teachers play a crucial role as decision makers, users of results and developers of various types of assessment. This makes assessment literacy one of the most important aspects of teacher training and professional development. Increasing assessment literacy among teachers of primary and lower secondary school learners is a capacity-building exercise. The aim is to build confidence among teachers in designing and/or selecting assessments that are valid, reliable, fit for purpose and have positive consequences for young English language learners, so that teachers can make sound decisions based on test results.

In addition, to help teachers make valid and reliable judgements of language use, it would be useful to ask them to act as examiners for large-scale tests in order to familiarise themselves with assessment criteria and standards, and in order for them not to be biased and underestimate some groups' achievement (SEN, disabilities, some ethnic minorities) (Harlen 2004, Campbell 2013). This would also enhance their 'diagnostic competence' (Edelenbos and Kubanek-German 2004).

According to Swaffield and Dudley (2003/2014), Popham (2009) and Taylor (2009, 2013) increasing assessment literacy may help tackle negative emotions, views and attitudes towards assessment among language teachers and the general public. Taylor (2013) proposes that language teachers need an understanding of the purposes and social role of assessment in education; an awareness of test consequences (impact and washback), accountability, ethics and of the responsibilities of stakeholders. Teachers need to have an understanding of the link between various assessment purposes, tools or instruments, methods and the curriculum. They need to be equipped with knowledge of the principles of sound assessment: an ability to identify and evaluate, develop and analyse a quality test. They should ideally have an in-depth knowledge of language competence: the trait to be measured, the link between learners' cognitive, social, emotional development and language learning for social and academic purposes. They should also have some understanding of how to use statistical information from classroom and large-scale test data: a basic grasp of numbers and measurement and an ability to extract data and interpret results

for various purposes. And finally, they need to have the wisdom to apply assessment information to inform decision making. This is an ability to make sensible decisions and critical choices, the know-how required to use effective assessment to maximise learning and minimise negative consequences and the wisdom to integrate assessment into the overall teaching and decision-making process.

Coherent educational framework

The challenge for policy makers and assessment providers is to create a coherent assessment framework with achievable targets and appropriate reporting at each transition point (between pre-school and primary, between primary and junior secondary and between junior and senior secondary education). There should be tight coherence in curriculum, assessment and teacher professional development. Without a clearly articulated progression in all three areas, support for teachers and provision of resources, there is a risk of overestimating feasible attainment levels within a given timeframe. If assessment information on children's attainment is not used in an informed and responsible way at various transition points in the education process, there is a 'continued danger that the achievements at primary school will be undervalued and underexploited at secondary school. This has serious consequences for ultimate attainment' (Rixon 2013, p. 40).

Future directions

The importance accorded to English language learning by parents, teachers and education authorities is likely to lead to a growing demand for English language instruction and an expansion of assessment among children and teenagers. Tasks ahead include the dissemination of the CEFR's extended set of descriptors for young learners in school contexts, to cater for communicative situations that fit better with young learners lives and experiences (BICS). In addition to this, there is a growing need to develop assessments to address and measure children's language learning needs related to academic achievement, both general language of schooling/academic English (CALP) and subject literacy in CLIL and other English as a medium of instruction contexts.

Future research is still due on the following:

- Attainable targets/standards of achievement by age groups in various contexts.
- Young learners' progression in social and instructional target language use domains.
- Technical qualities of young learners' tests.
- In-depth impact studies in specific contexts to investigate issues relating to:

  - Learner and teacher motivation.
  - The link between educational aims, curricula, teaching and assessment.

Technology

Technology has already produced computer-based and computer adaptive tests (e.g., Papp and Walczak 2016). Current developments point towards a revolution in assessment. Item-level data from large-item banks can be put to best use in adaptive assessments. Adaptive tests can be taken when ready, offer the right level of challenge and support and provide instant diagnostic feedback to inform learning and teaching. This is the promise of

next-generation instructional design and learning-oriented assessment. Automated assessment of open constructed response items is already firmly on the research and development agenda of major assessment boards (e.g., Evanini et al. 2015). Technology can also be exploited for the marking of speaking performances and writing scripts by human raters. Markers can be asked to make paired comparisons/comparative judgements (Jones 2016, McMahon and Jones 2015). This would lead to the creation of a reliable scale of quality by making holistic judgements about pairs of performances or scripts, making marking criteria redundant.

Assessing valued outcomes

The view on what desirable outcomes should be measured in educational assessments for young learners is constantly changing. 'Understanding of what makes up effective performance in [. . .] languages [. . .] is constantly changing; as are the content of educational programmes, societal expectations, and requirements in the economy' (Cambridge Assessment 2009, p. 8). Apart from communicative competence, a wider range of competences have been identified for assessment in the future. Some are relevant for the use of English as an international language: e.g., collaborative problem solving, creativity, concept formation, learning to learn skills and computer literacy (Masters 2013, The Gordon Commission 2013, Hill and Barber 2014, AQA 2015). Assessment of these broad desirable dispositions and competences can build on existing knowledge in the field of second language assessment among young learners.

Further reading

Bailey, A. L., and Carroll, P. (2015). Assessment of English language learners in the era of new academic content standards. *Review of Research in Education*, 39, 253–294.
    A discussion of academic English, and the validity of assessments of English language learners who are learning content through the medium of English in the USA.

Butler, Y. G. (2016). Assessing young learners. In Tsagari, D., and Banerjee, J. (eds.) *Handbook of second language assessment*. Berlin: Mouton de Gruyter, 359–376.
    A recent overview of the characteristics of young learners, issues related to construct, i.e., communicative language ability and academic English, age-appropriate tasks and assessment formats and impact of assessment.

Nikolov, M. (Ed.). (2016). *Assessing young learners of English: Global and local perspectives*. Dordrecht: Springer.
    A collection of recent research on assessing young learners by major international examination boards, and national examination providers as well as academic research on individual differences contributing to success, self and peer assessment.

Swaffield, S., and Dudley, P. (2003/2014). *Assessment for wise decisions*, 4th ed. Association of Teachers and Lecturers. London. https://www.atl.org.uk/Images/Assessment_and_literacy_for_wise_decisions_May_2015.pdf
    A useful publication originally published in 2003 now in its fourth edition, used as part of teacher training for assessment literacy in the UK.

Related topics

Differentiation, syllabus, grammar, vocabulary, speaking and listening, reading and writing.

References

AERA, APA, NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

AQA. (2015). *The Future of Assessment, 2025 and Beyond*. AQA.

Arnot, M., Schneider, C., Evans, M., Liu, Y., Welpy, O., and Davis-Tutt, D. (2014). *School Approaches to the Education of EAL Students: Language development, social integration and achievement*. Cambridge: The Bell Foundation.

Bailey, A. L. Heritage, M., and Butler, F. A. (2014). Developmental considerations and curricular contexts in the assessment of young language learners. In Kunnan, A. J. (ed.). *The Companion to Language Assessment*. Hoboken, NJ: Wiley-Blackwell.

Bailey, A. L., and Huang, B. H. (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, 28, 343–365.

Baron, P. A., and Papageorgiou, S. (2014). *Mapping the TOEFL® Primary™ Test onto the Common European Framework of Reference (TOEFL Research Memorandum ETS RM 14–05)*. Princeton, NJ: Educational Testing Service.

Baron, P. A., and Tannenbaum, R. (2011). *Mapping the TOEFL® Junior™ Test onto the Common European Framework of Reference (TOEFL Research Memorandum ETS RM-11–07)*. Princeton, NJ: Educational Testing Service.

Beacco, J.-C. (2010). *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Teaching/Learning History (End of Compulsory Education)*. Strasbourg: Council of Europe.

Beacco, J.-C., Byram, M., Cavalli, M., Coste, D., Egli Cuenat, M., Goullier, F., and Panthier, J. (2016). *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education*. Strasbourg: Council of Europe.

Beacco, J.-C., Coste, D., van de Ven, P.-H., and Vollmer, H. (2010). *Language and School Subjects – Linguistic dimension of knowledge building in school children*. Strasbourg: Council of Europe.

Benigno, V., and de Jong, J. (2016). A CEFR-based inventory of YL descriptors: Principles and challenges. In Nikolov, M. (ed.). *Assessing Young Learners of English: Global and local perspectives, educational linguistics volume 25*. New York: Springer, 43–64.

Bradbury, A., and Roberts-Holmes, G. (2016). The introduction of reception baseline assessment. "They are children . . . not robots, not machines" ATL, NUT, UCL: London. http://www.teachers.org.uk/files/baseline-assessment--final-10404.pdf

British Educational Research Association (BERA) (2011) *Ethical guidelines for educational research*. London. http://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-Ethical-Guidelines-2011.pdf

Butler, Y. G. (2009). Issues in the assessment and evaluation of English language education at the Elementary school level: Implications for policies in South Korea, Taiwan, and Japan. *The Journal of Asia TEFL*, 6(2), 1–31.

Butler, Y. G., and Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27(1), 5–31.

Cambridge Assessment. (2009). *The Cambridge Approach: Principles for designing, administering and evaluating assessment*. Cambridge: Cambridge Assessment Available online.

Cambridge English Language Assessment. (2013). *Principles of Good Practice: Quality management and validation in language assessment*. Cambridge. http://www.cambridgeenglish.org/images/22695-principles-of-good-practice.pdf

Cameron, L., and McKay, P. (2010). *Bringing Creative Teaching into the Young Learner Classroom*. Oxford: Oxford University Press.

Campbell, T. (2013). *Stereotyped at Seven? Biases in teacher judgements of pupils' ability and attainment* (CLS Working Paper 2013/8). London: University of London, Institute of Education, Centre for Longitudinal Studies.

Carless, D., and Lam, R. (2014). The examined life: Perspectives of lower primary school students in Hong Kong. *Education 3–13*, 42(3), 313–329.

Carlssen, C. (2008). The role of testing in an egalitarian society. *Research Notes*, 34, 2–6.

Carreira, J. M. (2012). Motivational orientations and psychological needs in EFL learning among elementary school students in Japan. *System*, 40, 191–202.

Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., and Hauck, M. C. (2016). *Designing the TOEFL® PrimaryTM tests (Research Memorandum No. RM-16–02)*. Princeton, NJ: Educational Testing Service.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Daugherty, R. (2012). Alternative perspectives on learning outcomes: Challenges for assessment. In Gardner, J. (ed.). *Assessment and Learning*. London: Sage, 76–78.

Earl, L. M. (1999). Assessment and accountability in education: Improvement or surveillance? *Education Canada*, 39 (3), 4–47.

Edelenbos, P., and Kubanek-German, A. (2004). Teacher assessment: The concept of "diagnostic competence". *Language Testing*, 21(3), 259–283.

Elwood, J. (2013). Educational assessment policy and practice: A matter of ethics. *Assessment in Education: Principles, Policy & Practice*, 20(2), 205–220.

Enever, J. (ed.) (2011). *ELLiE: Early language learning in Europe*. London: British Council.

ETS. (2015a). *Understanding TOEFL primary reading and listening score reports* (PDF). Educational Testing Service. https://www.ets.org/s/toefl_primary/pdf/understand_reading_listening_score_reports.pdf

ETS. (2015b). *Understanding TOEFL primary speaking score reports* (PDF). Educational Testing Service. https://www.ets.org/s/toefl_primary/pdf/understand_speaking_score_reports.pdf

European Commission. (2012). *First European survey on language competences*. Final Report, Technical Report, Executive Summary.

Evanini, K., Heilman, M., Wang, X., and Blanchard, D. (2015). *Automated scoring for the TOEFL Junior® Comprehensive Writing and Speaking Test (ETS Research Report No. RR-15–09)*. Princeton, NJ: Educational Testing Service.

Evans, M., Jones, N., Leung, C., and Liu, Y. (2016a). *EAL Assessment Framework for Schools*. Cambridge: The Bell Foundation.

Evans, M., Schneider, C., Arnot, M., Fisher, L., Forbes, K., Hu, M., and Liu, Y. (2016b). *Language Development and School Achievement: Opportunities and challenges in the education of EAL students*. Cambridge: The Bell Foundation.

Fleurquin, F. (2003). *Development of a Standardized Test for Young EFL Learners: Spaan Fellow Working Papers in Second or Foreign Language Assessment, Volume 1*. English Language Institute, University of Michigan.

Fredericksen, J., and Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27–32.

Goodier, T., and Szabo, T. (2017). The CEFR, ELP and lifelong learning – relating a collated sample of ELP descriptors for Young Learners to the updated CEFR illustrative scales. *4th Meeting of the EALTA CEFR Special Interest Group*, London, 11 February 2017. Eurocentres.

Haenni Hoti, A., Heinzmann, S., and Müller, M. (2009). "I can you help?": Assessing speaking skills and interaction strategies of young learners. In Nikolov, M. (ed.) *The Age Factor and Early Language Learning*. Berlin and New York: Mouton de Gruyter.

Harlen, W. (2004). *A Systematic Review of the Evidence of Reliability and Validity of Assessment by Teachers Used for Summative Purposes*. Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education*, 20(3), 245–270.

Härmälä, M., Lee Kristmanson, P., Moe, E., Pascoal, J., and Ramonienè, M. (2015). Language skills for successful subject learning. CEFR-linked descriptors for mathematics and history/civics (Language

Descriptors). https://www.ecml.at/Portals/1/Language%20Descriptors/Moe%20&%20Härmälä%
20-%20EALTA%20-%20Language%20descriptors%20for%20history%20and%20maths.pdf

Hasselgreen, A. (2005)a. Assessing the language of young learners. *Language Testing*, 22(3), 337–354.

Hasselgreen, A. (2005)b. *Testing the Spoken English of Young Norwegians: A study of testing validity and the role of 'smallwords' in contributing to pupils' fluency*. Cambridge: Cambridge University Press.

Hasselgreen, A., and Caudwell, G. (2016). *Assessing the Language of Young Learners*. Equinox: British Council.

Hill, P., and Barber, M. (2014). *Preparing for a Renaissance in Assessment*. London: Pearson.

Hönig, I. (2010). *Assessment in CLIL: Theoretical and empirical research*. Saarbrücken: VDM Verlag.

Inbar-Lourie, O., and Shohamy, E. (2009). Assessing young language learners: What is the construct? In Nikolov, M. (ed.), *The Age Factor and Early Language Learning*. Berlin and Germany: Mouton de Gruyter, 83–96.

Jones, N. (2002). Linking YLE levels into a single framework. *Research Notes*, 10, 14–15.

Jones, N. (2016). "No more marking": An online tool for comparative judgment. *Research Notes*, 63, 12–

Jones, N., and Saville, N. (2009). European language policy: assessment, learning and the CEFR. *Annual Review of Applied Linguistics*, 29, 51–63.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.

Kim, T.-Y., and Seo, H.-S. (2012). Elementary school students' foreign language learning demotivation: A mixed methods study of Korean EFL context. *The Asia Pacific Education Researcher*, 21(1).

Klenowski, V., and Wyatt-Smith, C. (2014). *Assessment for Education*. California: Sage, 101–103.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91, 645–655.

Little, D. (2011). The Common European Framework of Reference for Languages: A research agenda. *Language Teaching*, 44, 381–393.

Llinares, A., Morton, T., and Whittaker, R. (2012). *The Roles of Language in CLIL*. Cambridge: Cambridge University Press.

Massler, U. (2010). Assessment in CLIL learning. In Ioannou-Georgiou, S. and Pavlou, P. (eds.) *Guidelines for CLIL Implementation in Primary and Pre-primary Education*. Cyprus Pedagogical Institute, pp. 115–137. http://arbeitsplattform.bildung.hessen.de/fach/bilingual/Magazin/mat_aufsaetze/clilimplementation.pdf

Massler, U., Stotz, D., and Queisser, C. (2014). Assessment instruments for primary CLIL: The conceptualisation and evaluation of test tasks. *The Language Learning Journal*, 42(2), 137–150.

Masters, G. N. (2013). *Reforming Educational Assessment Imperative, Principles and Challenges*. Camberwell, Victoria: Australian Council for Educational Research.

Masters, G. N. (2014). Assessment: Getting to the essence. In *Designing the Future*, Issue 1, Centre for Assessment Reform and Innovation. www.acer.edu.au/files/uploads/Assessment_Getting_to_the_essence.pdf

McKay, P. (2006). *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

McKay, P., Hudson, C., Newton, M., and Guse, J. (2007). The NLLIA ESL bandscales Version 2. In McKay, P. (ed.) *Assessing, Monitoring and Understanding English as a Second Language in Schools*. Brisbane: Queensland University of Technology and Independent Schools Queensland.

McMahon, S., and Jones, I. (2015). A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy and Practice*, 22(3), 368–389.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.

Murphy, V. A. (2014). *Second Language Learning in the Early School Years: Trends and contexts*. Oxford: Oxford University Press.

Nikolov, M., and Szabó, G. (2012). Developing diagnostic tests for young learners of EFL in grades 1 to 6. In Galaczi, E. D., and Weir, C. J. (eds.). *Voices in Language Assessment: Exploring the impact of*

language frameworks on learning, teaching and assessment – Policies, procedures and challenges. Cambridge: UCLES/Cambridge University Press, 347–363.

Papp, S. (2009). Development of can do statements for KET and PET for schools. *Research Notes*, 36, 8–12.

Papp, S., Chambers, L., Galaczi, E., and Howden, D. (2011). *Results of Cambridge ESOL 2010 Survey on Young Learners Assessment*. Research and Validation Report Number: VR1310. Cambridge: Cambridge ESOL internal report.

Papp, S., and McElwee, S. (2015). *Development and Trialling of a New Test for Very Young Learners aged 4–6. Research and Validation Report Number: VR1550*. Cambridge: Cambridge English Language Assessment internal report.

Papp, S., Rixon, S. with Field, J. (forthcoming). *Examining Young Language Learners: The Cambridge English approach to assessing children and teenagers in schools*. Cambridge: Cambridge University Press.

Papp, S., and Salamoura, A. (2009). An exploratory study into linking young learners' examinations to the CEFR. *Research Notes*, 37, 15–22. Cambridge: Cambridge ESOL.

Papp, S., and Walczak, A. (2016). The development and validation of a computer-based test of English for children: the computer-based Cambridge English: Young Learners tests. In Nikolov, M. (ed.). *Assessing Young Learners of English: Global and local perspectives*. New York: Springer, 139–190.

Pearson Education. (2015). *GSE learning objectives for young learners*. https://dbxmrk3ash14o.cloud-front.net/englishfiles/GSE_LO_YoungLearners_beta_Dec15.pdf

Pinter, A. (2011). *Children Learning Second Languages*. Basingstoke: Palgrave Macmillan.

Pinter, A. (2014). Child participant roles in applied linguistics. *Research Applied Linguistics*, 35 (2), 168–183.

Pizorn, K. (2009). Designing proficiency levels for English for primary and secondary school students and the impact of the CEFR. In Figueras, N., and Noijons, J. (eds.) *Linking to the CEFR Levels: Research perspectives*. Arnhem: Cito/EALTA, pp 87–100.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48, 4–11.

QCA (2000). *National curriculum 2000 a language in common: Assessing English as an additional language*. Qualifications and Curriculum Authority. www.naldic.org.uk/Resources/NALDIC/Teaching%20and%20Learning/1847210732.pdf

Rea-Dickins, P., and Rixon, S. (1999). Assessment of young learners' English: Reasons and means. In Rixon, S. (ed.) *Young Learners of English: Some research perspectives*. Harlow: Addison Wesley Longman/The British Council, 89–101.

Rixon, S. (2013). *Survey of Policy and Practice in Primary English Language Teaching Worldwide*. London: British Council.

Rixon, S. (2016). Do developments in assessment represent the "coming of age" of young learners English language teaching initiatives? The international picture. In Nikolov, M. (ed.) *Assessing Young Learners of English: Global and local perspectives*. New York: Springer, 19–41.

Rupp, A., Vock, M., Harsch, C., and Köller, O. (2008). *Developing Standards-based Assessment Tasks for English as a Foreign Language: Context, processes, and outcomes in Germany*. Münster: Waxmann.

Shaw, S., and Imam, H. (2013). Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations. *Language Assessment Quarterly*, 10(4), 452–475.

Short, D. J. (1993). Assessing integrated language and content instruction. *TESOL Quarterly*, 27(4), 627–656.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Tumposky, D., and Wang, L. (2015). *TOEFL Junior design framework (TOEFL Young Students Research Report No. TOEFL Jr-02)*. Princeton, NJ: Educational Testing Service.

SQA. (2015). *Guide to Assessment. Scottish Qualification Agency*. www.sqa.org.uk/files_ccc/25Guide ToAssessment.pdf

Stotz, D., and Megías, R. M. (2010). Moving on, broadening out. A European perspective on the transition between primary and secondary foreign language learning. *Babylonia*, 1, 30–34.

Szpotowicz, M., and Campfield, D. (2016). Developing and piloting proficiency tests for Polish young learners. In Nikolov, M. (ed.). *Assessing Young Learners of English: Global and local perspectives*, pp. 109–137.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21–36.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, 30(3), 403–412.

The Gordon Commission. (2013). The Gordon Commission Final Report: To Assess, To Teach, To Learn: A Vision for the Future of Assessment, Technical report.

United Nations. (1989). *Convention on the Rights of Child*. www.unicef.org/rightsite/files/uncrcchilld friendlylanguage.pdf

Vollmer, H. (2010). *Items for a Description of Linguistic Competence in the Language of Schooling Necessary for Teaching/learning Sciences (End of Compulsory Education)*. Strasbourg: Council of Europe.

Weir, C. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, 22(3), 281–300.

WIDA. (2012). *English Language Proficiency Standards for English Language Learners in Kindergarten through Grade 12*. Board of Regents of the University of Wisconsin System, on behalf of the WIDA Consortium. www.wida.us