Language Teaching (2021), **54**, 1–37
doi:10.1017/S0261444820000294

CAMBRIDGE UNIVERSITY PRESS

STATE-OF-THE-ART ARTICLE

Assessing young learners' foreign language abilities

Marianne Nikolov[1]* and Veronika Timpe-Laughlin[2]

[1]University of Pécs, Hungary and [2]Educational Testing Service, Princeton, USA
*Corresponding author. Email: nikolov.marianne@pte.hu

Abstract

Given the exponential growth in the popularity of early foreign language programs, coupled with an emphasis of evidence-based instruction, assessing young learners' (YLs) foreign language abilities has moved to center stage. This article canvasses how the field of assessing young learners of foreign languages has evolved over the past two decades. The review offers insights into how and why the field has developed, how constructs have been defined and operationalized, what language proficiency frameworks have been used, why children were assessed, what aspects of their foreign language proficiency have been assessed, who was involved in the assessment, and how the results have been used. By surveying trends in foreign language (FL) and content-based language learning programs involving children between the ages of 3 and 14, the article highlights research into assessment OF and FOR learning, and critically discusses areas such as large-scale assessments and proficiency examinations, comparative and experimental studies, the impact of assessment, teachers' beliefs and assessment practices, young learners' test-taking strategies, age-appropriate tasks, alternative and technology-mediated assessment, as well as game-based assessments. The final section of the article highlights where more research is needed, thus outlining potential future directions for the field.

1. Introduction

The widespread implementation of early teaching and learning of FLs/L2s (second languages), particularly English, has possibly become 'the world's biggest policy development in education' (Johnstone, 2009, p. 33). Historically, interest in early FL learning dates back to the late 1960s; since then, the development of FL programs for YLs has advanced globally in three noticeable waves (Johnstone, 2009), all of which were followed by a subsequent loss of enthusiasm toward an early start due to discouraging results. Currently, we are experiencing the fourth wave characterized by three trends in the exponential spread of early FL programs. These trends include (1) an emphasis on assessment for accountability and quality assurance, (2) assessment not only of YLs in the first years of schooling but also of very young learners of pre-school age, and (3) an increase in content-based FL teaching, thus adding to the broad range of early FL programs.

As shown in Table 1, early FL education programs vary substantially in their foci and amount of time allocated in the curriculum to learning the FL (Edelenbos, Johnstone, & Kubanek, 2006; Johnstone, 2010). With regard to focus, the approaches to teaching FLs can be placed along a continuum between language and content (Inbar-Lourie & Shohamy, 2009), ranging from (1) target language (TL) as subjects and (2) language and content-embedded approaches that aim to develop competence in the FL by borrowing topics from the curriculum (e.g., science, geography etc.) to (3) content and language integrated learning (CLIL), a popular term for content-based approaches (Cenoz, Genesee, & Gorter, 2014) covering a wide range of practices, all the way to (4) immersion programs teaching multiple school subjects in the L2.[1] The time allocated to the TL in the four

[1]Note that we focus on the first three approaches, whereas immersion programs and language awareness programs with no achievement targets in the FL are beyond the scope of this article.

© The Author(s), 2020. Published by Cambridge University Press

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Table 1. Models of formal early FL education programs

	Category[a]	Time allocated to the FL	Description
1.	TL as subjects	5–10%	A set amount of time per day or week of formalized FL instruction (e.g., a few minutes per day; one hour per week), oftentimes based on a textbook that provides the curriculum.
2.	TL as subject (embedded)	5–10%	Similar to 1. with a more flexible approach insofar as the TL instruction is included in the teaching of other content or subject areas such as Science, Geography or current events etc.
3.	CLIL	20–30%	Although used in different ways, CLIL generally refers to learning content through an additional language. It denotes teaching both the subject and the TL.
4.	Immersion programs	50–90%	Multiple school subjects are taught in the TL.

Note: [a]The labels and descriptions are partly adapted from Johnstone, 2010, p. 16f.

approaches gradually increases in line with the content focus from modest in the cases of (1) and (2), to significant in (3), and to substantial in (4) (Johnstone, 2009, 2010).

Although in reality, the various programs and classrooms may be much more multifaceted, this broad categorization highlights two key aspects: (a) the diversity of the early FL education scene and (b) the variety of 'language-related outcomes [that] are strongly dependent on the particular model of language education curriculum which is adopted' (Edelenbos et al., 2006, p. 10). Hence, just as teaching and learning contexts vary substantially, so do the contents and goals of early FL learning.

Across this diverse landscape of early FL education programs and learning contexts, the fourth wave is characterized by an emphasis on assessment of YLs as part of a policy shift toward evidence-based instruction (e.g., Johnstone, 2003). As a result, we have witnessed an increase in investigations—both in educational and social sciences (e.g., Haskins, 2018)—into how assessment in early language learning programs impacts children's overall development and their teachers' work. Hence, the popularity of early FL programs, coupled with the emphasis of evidence-based instruction, has resulted in the 'coming of age' (Rixon, 2016) of YL assessment—a field as diverse as the early FL education landscape itself.

In this review, we explore the main trends in assessing YLs' FL abilities over the past two decades, offering a critical overview of the most important publications. In particular, we offer insights into how and why the field of assessing YLs has evolved, how constructs have been defined and operationalized, what frameworks have been used, why YLs were assessed, what aspects of FL proficiency have been assessed and who was involved in the assessment, and how the results have been used. By mapping meaningful trends in the field, we want to indicate where more research is needed, thus outlining potential future directions for the field.

1.1 Criteria for choosing studies

We identified a body of relevant publications using a number of criteria for both inclusion and exclusion (Table 2). First, we followed Rixon and Prošić-Santovac's (2019, p. 1) definition of assessment as 'principled ways of collecting and using evidence on the quality and quantity of people's learning'. Accordingly, our review includes a consideration of both summative and formative approaches to eliciting YLs' knowledge and performances for the sake of informing classroom-based teaching and learning. Thus, this review encompasses research on formative assessment, including alternative assessments such as observations, self-assessment, peer assessment, and portfolio assessment, as well as studies on large-scale summative assessment projects and proficiency examinations.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Table 2. Criteria for inclusion and exclusion of studies

	Inclusion criteria	Exclusion criteria
Age of research participants	3–14 years	All other age groups
Language of publication	English	All other languages
Target languages	English, French, German, Japanese	All other languages
Type of program	FL as subject; content-embedded; content-based	Language awareness; immersion
Type of publication	• Refereed flagship journals • Chapters in edited volumes • Monographs and books • White papers • Language policy document • Research reports	• Theses • Conference presentations • Testing materials published by large-scale assessment providers

Second, in the larger area of YL assessment, the term 'young learners' is broadly used to denote students younger than those entering college, but young learners are far from being homogenous in terms of age and cognitive, emotional, social, and language development (McKay, 2006; Hasselgreen & Caudwell, 2016; Pinter, 2006/2017). In this review, we included publications in English on various TLs (not only English) that focused on YLs in pre-school (ISCED 0), lower-primary or primary (ISCED 1), and upper-primary or lower-secondary (ISCED 2; UNESCO Institute for Statistics, 2012) programs, ranging from age 3 to 14. Additionally, we focused on contexts where the TL is an FL, not an official language. Although we are aware that the status of a target language should be seen more like a continuum than a dichotomy of FL or L2, we excluded studies from our database that were conducted in L2 contexts, as well as those in language awareness and (bilingual) immersion programs. Hence, we reviewed studies conducted in FL contexts while taking into account the recent shift toward content-based instruction.

Third, as the body of research has substantially grown over the past two decades, we reviewed publications starting from the first discussions by Rea-Dickins and Rixon (1997) and the seminal 2000 special issue of *Language Testing*. Our aim was to explore how the field has changed since then. We included relevant language policy documents, frameworks, and assessment projects published in a range of venues (Table 1) in order to analyze (1) how the assessment constructs have been defined, (2) what specific language proficiency models have been proposed and tested, (3) what principles test developers followed, and (4) how predictions and actual tests have been used, (5) how they have worked, and (6) how the results have been utilized.

Given these inclusion criteria and the primary focus on studies in which the proficiency of YLs in an FL is assessed, this review has a few limitations. For example, it does not discuss in detail phenomena that are related to early language learning. Accordingly, the review excludes the following aspects: (a) how children's attitudes toward their FL, their speakers and their cultures, and toward language learning in general are shaped, (b) how early learning of an FL evokes and maintains YLs' language learning motivation, self-confidence, willingness to communicate, low anxiety, and growth mindset, or (c) other aspects of individual differences, including learner strategies and sociocultural background. It discusses, however, how these aspects have been found to impact YL's performance on assessments in studies which used them for triangulation purposes either to test models of early language learning or to complement qualitative data in mixed methods research. As will be seen, most studies have been conducted in European countries—an aspect that will be discussed in the future research section below.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

2. The larger picture: What is the construct and how has it changed over time?

In this section, we focus on constructs and frameworks underlying the assessment of YLs' FL abilities and trace how these have developed over time. In particular, we show how the field has gradually identified those language skills relevant for young FL learners, YLs' potential linguistic goals and FL achievements, as well as (meta)cognitive and affective variables that impact FL development and assessment.

2.1 First steps in early FL assessment

Initial studies provided some insights into approaches taken to evaluate what YLs are able to do in the FL, thus revealing glimpses into the local contexts and underlying constructs that were assessed. Mainly conducted in the context of type (1) and (2) FL education programs designated in the Introduction, initial assessment studies included both summative and formative assessments.

From a summative perspective, early YL assessment studies used data obtained in the context of national assessments that were administered at the end of primary education in various countries (e.g., Edelenbos & Vinjé, 2000; Johnstone, 2000; Zangl, 2000). For example, Edelenbos and Vinjé (2000) analyzed data collected in a National Assessment Programme in Education that aimed at measuring YLs' achievements in English as a foreign language (EFL) at the end of their primary education in the Netherlands. Assessments were administered in paper-and-pencil format (listening, reading, receptive word knowledge, use of a bilingual wordlist) and in individual face-to-face sessions with an interlocutor. The latter included a focus on speaking elicited through a discussion with English-speaking partners, pronunciation gauged by means of reading aloud sentences, and productive word knowledge tested by providing students with pictures that they had to label or describe. Edelenbos and Vinjé (2000) found that overall students performed well in listening, but not in reading. In particular for reading, they noted better outcomes if teachers used a communicative approach instead of a grammar-oriented one, while the latter approach tended to result in higher scores in word knowledge. Results were mediated by learners' socioeconomic backgrounds, the amount of EFL instruction, the types of teaching materials used, and in particular, the training and proficiency of EFL teachers.

Similar to Edelenbos and Vinjé (2000), Johnstone (2000) described efforts with regard to the assessment of YLs' French as an FL attainments in primary schools in Scotland. He emphasized diversity and variability in teaching contexts as the main challenge in the assessment of YLs' proficiency. Therefore, early assessments for reading, listening, and writing administered in Scotland were developed locally at the different schools, while the research team designed content-free speaking tasks that were deployed across learning contexts. As 'vessels into which pupils could pour whatever language they were able to' (Johnstone, 2000, p. 134), the content-free speaking assessments included three types of data elicitations: (1) systematic classroom observations of student-teacher and student-student interactions, (2) vocabulary retrieval tasks that were based on free word association tasks in which children were asked to say 'whatever words or phrases came into their head in relation to topics with which they were familiar' (Johnstone, 2000, p. 135), and (3) paired speaking tasks administered in face-to-face sessions that mirrored classroom interactions familiar to the students. Especially the latter served as the main oral assessment to gauge YLs' pronunciation, intonation, grammar control, and range of structures—all of which were rated on a three-point scale. In his conclusion, Johnstone (2000) highlighted the need to explore in more detail the construct of FL proficiency at the early age and promote formative assessments in primary FL education—areas that were further investigated by Zangl (2000) and Hasselgreen (2000).

Exploring in more detail YLs' FL competences, Zangl (2000) introduced the assessments deployed in a FL education program in primary schools in Austria in which English was taught for 20 minutes three times per week. Children were assessed at the end of primary school by means of observations of classroom interactions (e.g., students' group work, role-plays with puppets, and student-teacher interactions centered on certain topics such as holidays, favorite books, or hobbies), semi-structured

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

interviews conducted with individual learners to gauge spontaneous speech, and oral tests that aimed to elicit specific structures in the areas of morphology, syntax, and lexis/semantics. Similar to Johnstone (2000), Zangl (2000) wanted to obtain a comprehensive picture of learners' English language proficiency in the areas of social language use and developing discourse strategies, in particular with regard to spontaneous speech, pragmatics (the use of language in discourse and context) and specific structures in morphosyntax and lexis. Conducting a multi-component analysis, Zangl highlighted specific aspects of YL's FL development, including its non-linear nature and the interconnectedness of the different language components. For example, she found that a learner may first produce the correct morphosyntactic form and a little later, may produce an incorrect form. Thus, it may seem that the learner regresses. However, Zangl referred to this phenomenon as 'covert progression' (p. 256), arguing that learners tend to initially memorize a correct form which may then vanish as they begin to produce language more actively and freely, thus applying rules and occasionally overgeneralizing them—a phenomenon commonly referred to as U-shaped development. With regard to interconnectedness of language phenomena, Zangl found that a student's increasing ability to form questions positively impacted their ability to take turns and to participate actively in student-teacher interactions. She highlighted that these insights into the development of learners' L2 acquisition provide important pieces of information for test developers who should adapt assessment materials to students' age, cognitive and linguistic abilities, their interests, and attention span in order to provide assessments that 'reflect the *learning* process' (p. 257; italics in the original).

In addition to summative national assessment projects, early investigations also explored formative approaches to assessing YLs (e.g., Gattullo, 2000; Hasselgreen, 2000). Hasselgreen (2000), for example, described an assessment battery developed for primary-level classrooms in Norway—a context in which the curriculum was underspecified in terms of content and outcomes of early EFL education. As a first step, Hasselgreen conducted a survey with 19 teachers in Bergen, thus identifying four components of communicative language ability taught in early EFL education: (1) vocabulary, morphology, syntax, and phonology, (2) textual ability (e.g., cohesion), (3) pragmatics (i.e., how language is used in TL contexts), and (4) strategic ability (i.e., how to cope with communicative breakdown and difficulties in communication). Based on these four areas, she then developed an assessment battery with tests for reading, listening, speaking, and writing. Each test included integrated tasks that featured topics, situations, and texts YLs were believed to be familiar with through classroom activity. For instance, reading was assessed by means of matching tasks with pictures, true-false choice tasks, and gap-filling activities. During the listening test, students would listen to a mystery radio play and were asked to identify specific aspects in pictures (listen for detail). The writing test would then build on the mystery play with students being asked to write a diary entry or response letter to a character in the play. Finally, the speaking assessment was administered in pairs with picture-based prompts. Additionally, classroom-based observations were added to provide further insights about the learners' speaking skills. Thus, Hasselgreen was among the first to develop systematically an initial assessment battery that was supposed to be used regularly in EFL classrooms in order to promote metalinguistic awareness and assessment literacy among EFL teachers and YLs.

To summarize, early, largely descriptive accounts of summative and formative approaches to YL assessment provide insights into the diversity of early FL teaching contexts. In particular, they reveal the lack of a consensus with regards to what proficiency in the FL means for YLs and how the putative language abilities were supposed to be assessed. That is, the construct definition of what was assessed varied widely across contexts and included to various degrees the four macro-skills as well as different components of the language system such as lexis/semantics, grammar, pragmatics, intonation, cohesion and/or pronunciation (see Table 3 for an overview)—features that are strongly reminiscent of constructs proposed in the context of adult L2 learning and assessment. Additionally, many assessments used at this early stage were still based on rather traditional, paper-and-pencil formats such as multiple-choice items (Edelenbos & Vinjé, 2000).

However, despite the large diversity and at times traditional approaches to FL assessment, certain trends appeared to develop that emphasized the particular needs and individual differences of YLs. For

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Table 3. Language components in early assessment studies aiming to measure young learners' FL abilities

						Grammar						
	Speaking	Listening	Reading	Writing	Search skills	Morphology	Syntax	Vocabulary	Discourse strategies	Pragmatics	Pronunciation/ intonation	Cohesion
Edelenbos and Vinjé (2000)	x	x	x	—	x	—	—	x	—	—	—	—
Johnstone (2000)	x	x	x	—	—	x	x	x	—	—	x	—
Zangl (2000)	x	—	—	—	—	x	x	x	x	x	—	—
Gattullo (2000)	x	x	—	—	—	x	x	x	—	—	—	—
Hasselgreen (2000)	x	x	x	x		x	x	x	—	x	x	x

example, all early assessment studies focused on speaking and interaction, thus accounting for the oral dimension of early FL learning. Moreover, most of the studies administered assessments in contexts that are familiar to YLs. Interviews or paired speaking assessments (Johnstone, 2000; Zangl, 2000) were carried out in settings that were meant to resemble familiar face-to-face classroom interactions, while several assessments even highlighted the use of classroom observations as a means of gauging insights into how YLs could use the FL in context and interaction. Also, administrators deployed integrated tasks based on materials that were regarded age-appropriate, familiar, and engaging to YLs, primarily using visual and tactile stimuli (Hasselgreen, 2000). Finally, in all early assessment studies, researchers have highlighted the need to gain a detailed understanding of YLs' FL development and proficiency in order to inform the design and construct underlying YL assessments—a call that was pursued in more depth in assessment-oriented research in the early 2000s.

2.2 Learning to walk: The evolution of frameworks for early FL assessment

While the first approaches to YL assessment largely prioritized '*fun and ease*' in terms of providing anxiety-free, positive testing experiences (Nikolov, 2016a, p. 4; italic in the original), the field soon faced a call for more accountability and thus the need to determine realistic, age-appropriate achievement targets (Johnstone, 2009; Nikolov, 2016a). As a result, researchers began to put forth principles, models, and frameworks intended to guide YL FL teaching and assessment (e.g., Cameron, 2003; McKay, 2006). Among the first, Cameron (2003, p. 109) proposed a multidimensional 'model of the construct "language" for child foreign language learning' that is aligned with children's first language (L1) acquisition and based on three fundamental principles that focus on: (a) meaning, (b) oral communication, and (c) development that is sensitive to children's emerging L1 literacy. Rooted in these principles, this framework distinguished between 'oral' and 'written' language, identifying in particular vocabulary (i.e., the comprehension and production of single words, phrases, and chunks) and discourse (i.e., understanding, recalling, and producing 'extended stretches of talk' including songs, rhymes, and stories) as key dimensions of oral and meaning-oriented language use (Cameron, 2003, p. 109). Discourse is further subdivided into conversation and extended talk as means of using the FL in communicative interaction. Grammar is argued to constitute more of an implicit element in YL instruction insofar as it is needed to develop a sense for patterns underlying the FL. Cameron (2003) argued that teaching and assessment need to foreground meaning-oriented, oral communication—a step forward if we consider the relatively strong emphasis on grammar in the early YL FL assessments (see Table 3).

Building on principles included in Cameron (2003)and Nikolov (2016a), in the context of developing a diagnostic assessment for YLs in Hungary, proposed a framework that provides a more comprehensive picture of how primary-level learners develop their EFL proficiency. In contrast to earlier frameworks that primarily drew upon insights from L1 acquisition, development of ESL learners' academic language proficiency, and adult L2 ability, Nikolov also included findings from longitudinal projects in second language acquisition that investigated YLs FL development over a period of time relative to factors such as age (García Mayo & García Lecumberri, 2003; Muñoz, 2003, 2006), cognitive and socioaffective development (e.g., Mihaljević Djigunović, 2006; Kiss, 2009), learning strategies (e.g., Csapó & Nikolov, 2009), and the quality of FL instruction (e.g., Nikolov & Curtain, 2000). Based upon that research, she put forth the following high-level principles:

- The younger learners are, the more similar their FL development is to their L1 acquisition.
- Children tend to learn implicitly, based on memory, and only gradually develop the ability to rely on rule-based, explicit learning strategies, which becomes more prominent in their approach to learning around puberty.
- Children develop fairly similarly in terms of aural and oral skills (see Cummins, 2000), while more individual differences, related to YLs' L1 abilities, aptitude, cognitive abilities, and parents' socioeconomic status, can be found in their literacy development.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

- Learning and assessment should (a) focus on children's aural and oral skills (i.e., listening comprehension and speaking abilities), while 'reading comprehension and writing should be introduced gradually when they are ready for them' (Nikolov, 2016b, p. 75) and (b) always build upon what YLs know and can do in terms of their world knowledge, comprehension, and L1 abilities, thus promoting a positive attitude toward FL learning.
- Learning and assessment tasks need to be age-appropriate insofar as they recycle familiar language while offering opportunities to learn in a way that is 'intrinsically motivating and cognitively challenging' (Nikolov, 2016b, p. 71).

Discussing each principle in detail, Nikolov highlighted that overall 'achievement targets in [YLs'] L2 tend to be modest' (Nikolov, 2016a, p. 7) as children move from unanalyzed chunks to more analyzed language use (Johnstone, 2009). Moreover, she captured the diversity with regard to FL educational contexts, learners' individual differences, and developmental paths (for overviews, see Nikolov & Mihaljević Djigunović, 2006, 2011)—aspects that need to be considered in order to align assessment constructs and outcome expectations with specific learner groups in the given local educational contexts.

In addition to frameworks and principles, there was a need, in particular for national and international assessments, to provide accounts of quantifiable targets that describe in detail what children are expected to do at certain stages in their FL development. This need resulted in studies, mostly across Europe, that adapted the language descriptors included in the English Language Portfolio (ELP) and Common European Framework of Reference (CEFR) for young learners of FLs (e.g., Hasselgreen, 2003, 2005; Curtain, 2009; Papp & Salamoura, 2009; Pižorn, 2009; Baron & Papageorgiou, 2014; Benigno & de Jong, 2016; Papp & Walczak, 2016; Szabo, 2018a,b). Benigno and de Jong (2016), for example, described the first phase of a multiyear project to develop a 'CEFR-based descriptor set targeting young learners' (p. 60) between 6 and 14 years of age. After identifying 120 learning objectives for reading, listening, speaking, and writing from English language teaching textbooks, curricula, and the ELP, they assigned proficiency level ratings to the objectives in standard-setting exercises with teachers, expert raters, and psychometricians who calibrated and scaled the objectives relative to the CEFR descriptors. Although Benigno and de Jong argued that they were able to adapt the CEFR descriptors from below A1 to B2 for YLs and align them with Pearson's continuous Global Scale of English, it remains unclear how key variables such as age, learning contexts, developing cognitive skills and L1 literacy, and empirical data on YLs' test results feature into the rather generic descriptors (for the complete set of descriptors, see https://online.flippingbook.com/view/872842/).

In an attempt to account for individual differences with regard to social and cognitive development and provide a reference document for YL educators, Szabo (2018a,b) presented a collation of CEFR descriptors of language competences for YLs between 7 and 10 as well as 11 and 15 years, respectively. She iteratively reviewed ELPs for YLs from 15 European countries and mapped the self-assessment statements to the CEFR descriptors, while rating the CEFR descriptors and the ELP can do statement with regard to perceived relevance for primary-level learners. For the age range of 7–11-year-old learners, for instance, she included CEFR levels ranging from pre-A1 to B2 across all language skills (C1 and C2 levels were excluded due to their limited relevance to the age group), thus outlining can do descriptors for reception, production, interaction, and mediation abilities in an effort to provide 'the basis of language examination benchmarking to the CEFR' (Szabo, 2018a, p. 10). Although a very thorough attempt to establish a potential benchmark, Szabo acknowledged a '"bias for best" approach' (p. 15) with regards to the hypothetical learning context. In other words, whether or to what extent the B2 level can do descriptors constitute a realistic achievement target for primary-level FL education is questionable.

To summarize, the field of young learner assessment has put forth frameworks aimed at defining in more detail a construct of language for child FL learning to account for achievement targets both at local and more global levels. The framework descriptions proposed primarily in Europe (Cameron,

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

2003; Nikolov, 2016a) and the United States (see e.g., Curtain, 2009) tend to foreground aural and oral FL abilities as opposed to language knowledge (e.g., grammar), while highlighting children's developing social, emotional, cognitive, and literacy skills. Additionally, more globally oriented frameworks such as the CEFR collation suggest a somewhat wider construct by including listening, speaking, writing, reading, and interaction skills ranging from pre-A1 to B2—skills that would also be mediated by differences in developing cognitive, literacy, affective, and socioeconomic aspects.

2.3 On firmer empirical footing: Investigating frameworks and variables of early FL education

To explore the considerable variation in achievements of young FL learners from similar backgrounds in similar learning contexts, researchers have increasingly focused on aspects related to YLs' FL development such as cognitive, affective, and sociocultural variables. In particular, assessment studies have focused on aptitude and cognition (Kiss & Nikolov, 2005; Alexiou, 2009; Kiss, 2009), affect, motivation, anxiety, and learning difficulties (Mihaljević Djigunović, 2016; Kormos, 2017; Pfenninger & Singleton, 2017), socioeconomic background (Bacsa & Csíkos, 2016; Butler & Le, 2018; Butler, Sayer, & Huang, 2018; Nikolov & Csapó, 2018), learning strategies, emerging L1 and L2 literacy skills and how learners' languages interact in order to explore what it means for children to learn additional languages and how these factors impact L2 assessment.

As a learner characteristic that is considered responsible for much of the variation in FL achievements, aptitude for language learning is generally viewed as a predisposition or natural ability to acquire additional languages in a fast and easy manner (Kiss & Nikolov, 2005; Kiss, 2009). While aptitude is relatively well researched in adult L2 learners, Kiss and Nikolov (2005) were among the first to report on the development and psychometric performance of an aptitude test for YLs (specifically, 12-year-old L1 Hungarian learners of English). Based on earlier models and aptitude tests for adult L2 learners, they conceptualized aptitude as consisting of four traits. Accordingly, they included four tests in their larger aptitude test battery for YLs (Kiss & Nikolov, 2005, p. 120):

1. Hidden sounds: Associating sounds with written symbols
2. Words in sentences: Identifying semantic and syntactic functions
3. Language analysis: Recognizing structural patterns
4. Vocabulary learning: Memorizing lexical items (short-term memory).

They administered the aptitude test battery, an English language proficiency test with listening, reading, and writing sections based on the local curriculum, and a motivation questionnaire to 398 sixth graders from ten elementary schools in Hungary. Although they could not account for the children's oral abilities in English, Kiss and Nikolov found that the aptitude test exhibited evidence of construct validity with results indicating four relatively independent abilities that all showed strong relationships with students' performance on the English proficiency measure. Overall, aptitude scores explained 22% and motivation explained 8% of variation in the English scores.

To help a primary school select children for a new dual-language teaching program, Kiss (2009) administered a slightly adapted version of the same aptitude test to 92 eight-year-old Hungarian students in second grade. Additionally, students completed a five-minute oral interview and engaged in an oral spot-the difference task. Kiss was able to confirm the good performance of the aptitude test insofar as the test identified students with the higher FL oral performance. Additionally, she found short-term working memory ability to be quite distinct from other traits. Also, when comparing the results with the 8-year-olds in the earlier study (Kiss & Nikolov, 2005), she found that the 12-year-olds performed much better. She speculated that this was most likely because at about eight years of age, children did not have exposure to vocabulary memorization and thus, had not developed memorization strategies yet.

Additionally, studies began to investigate aptitude relative to specific language skills such as YLs' vocabulary development in the FL (Alexiou, 2009) and listening comprehension (Bacsa & Csíkos,

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

2016). Alexiou (2009) investigated YLs' aptitude and vocabulary development in English with five–nine-year-old L1 Greek students ($n = 191$). Using non-language measures as some of her test takers were not yet literate, Alexiou administered an aptitude measure consisting of memory and analytic tasks as well as receptive and productive vocabulary tests that featured words selected from the learners' academic curriculum. She found rather moderate, yet statistically significant relationships between YLs aptitude and vocabulary development in English, hypothesizing that at the beginning YLs favor phonological vocabulary and only later does the orthographic recognition exceed phonological learning. Unfortunately, Alexiou's analysis did not account for differences in age. However, she still argued that aptitude appears to progress with age as cognitive skills evolved and potentially reached its peak when children become cognitively mature—a hypothesis that was further examined by Bacsa and Csíkos (2016).

Over the course of six months, Bacsa and Csíkos (2016) investigated the listening comprehension and aptitude of 150 fifth and sixth graders (age 11–12 years) in ten school classes in Hungary. After training teachers on how to add diagnostic listening tests to their current syllabi, they administered listening assessments at the beginning and end of the six-month period as pre- and posttests. In addition, Kiss and Nikolov's (2005) aptitude test and questionnaires on motivation and anxiety were administered. Deploying correlations, regression analysis, cluster analysis, and path analysis, they focused their analysis on six variables, including parents' education (or socioeconomic background), aptitude, language learning strategies, beliefs, attitudes, motivation, and anxiety. They found students' achievements in both grades to be considerably higher on the posttest. The largest percentage of variation in children's listening comprehension was explained by YLs' aptitude and parents' education (28.3% and 4.4%, respectively). Additionally, learners' beliefs about difficulties in language learning (6.8%), anxiety about unknown words (3.2%), and difficulty of comprehension (5.7%) in listening tests also contributed to their listening scores. Overall, cognitive factors explained more of the variation in YLs' FL achievement than affective factors. Affective factors, however, changed consistently and seemed to depend upon the language learning context.

Similar findings were reported by Kormos (2017) who critically reviewed research at the intersection of L1 and L2 development, cognitive and affective factors, and YLs' specific learning difficulties (SLDs). With a particular focus on reading, she highlights that among the factors impacting YLs' reading abilities are processing speed, working memory capacity (storage and processing capacity in short-term memory), attention control, and ability to infer meaning. Identifying these as 'universal factors that influence the development of language and literacy skills in monolingual and multilingual children' (Kormos, 2017, p. 32), Kormos specifically pinpoints phonemic awareness and rapid automated naming as those phonological processing skills that play a key role in decoding and encoding processes across languages and that may create particular issues for YLs with SLDs. Nevertheless, Kormos advocates for the assessment of the FL abilities of YLs with and without SLDs, emphasizing in particular the need to develop appropriate assessments to explore how motivational, affective and cognitive factors, instructional environment, and personal contexts impact the literacy development of all students.

In sum, aptitude appears to be a significant predictor of FL achievements (Kiss & Nikolov, 2005; Kiss, 2009), with cognitive variables explaining a large portion of the variation in YLs' achievements (Csapó & Nikolov, 2009; Bacsa & Csíkos, 2016). Additionally, affective variables such as motivation and the perception of the learning environment which have only been examined indirectly in many studies (e.g., Kiss & Nikolov, 2005; Bacsa & Csíkos, 2016; Kiss & Nikolov, 2005) have been identified as predictors of YLs' FL achievement. For example, over seven months, Sun, Steinkrauss, Wieling, and de Bot (2018) assessed the development of English and Chinese vocabulary of 31 young Chinese EFL learners (age 3.2–6.2). Additionally, aptitude was assessed together with internal and external variables. Participants' vocabulary was tested by four tests: the Peabody Picture Vocabulary Test and the Expressive One-Word Picture Vocabulary Test, the depth of vocabulary by semantic fluency and word description tests both in English and in Chinese (translated versions) before and after the English program. Two aptitude measures tapped into the children's phonological short-term memory and non-verbal intelligence. The study reported stronger effects of external factors (e.g.,

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

exposure to English at school and at home) than aptitude, thus pointing to stronger effects of external factors over aptitude when it comes to Chinese YLs in an EFL context.

Overall, longitudinal research with larger samples would be desirable in order to gauge causal relationships and to account for the relative impact of these variables on language learning. While these variables constitute key factors in determining YLs' FL learning, they are not necessarily stable characteristics. Rather, they seem to change and evolve over time as children mature cognitively, become literate in their L1 and additional languages, and gather experiences related to formal language learning. Moreover, they are impacted by parents, teachers, and peers, for example, and their roles and impact change as YLs age (for an overview see Mihaljević Djigunović & Nikolov, 2019). Future research should focus on how memory-based learning, more typical of younger learners, shifts toward more rule-based learning. One would expect memory to be a better predictor of L2 learning for younger children than inductive or deductive abilities—a hypothesis which, if proven, could provide valuable information for the design and administration of assessments aimed to measure YLs' FL abilities.

3. Assessment of learning

Summative assessment, also referred to as assessment OF learning, serves the purpose of obtaining information regarding YLs' achievements at the end of a teaching or learning process (e.g., a task, a unit, a program etc.). In summative assessment, the aim is to measure to what extent YLs have mastered what they were taught, or in the case of proficiency tests, to what extent students have achieved the targets in the FL along certain criteria. Research reviewed in this section—including national assessment projects, validation projects, and examinations contrasting different types of learning contexts—pertains to this larger paradigm of assessment of learning.

Most of the reviewed studies were motivated by changes in policies regarding language learning and subsequent accountability needs, on the one hand, as well as researchers' interests in various aspects of early language learning, on the other. Quality assurance and accountability typically underlie national assessment projects and validation studies, as decision-makers want to know student learning outcomes relative to curricular goals or how YLs in programs starting earlier and later as well as FL and CLIL programs compare to one another. Additionally, validation projects aim to find evidence on how effective traditional and more innovative types of (large-scale) tests are for assessing YLs' language skills, thus also accounting for the quality of assessments. Finally, other projects—oftentimes small-scale experimental studies—reflect researchers' interests in various aspects of early language learning.

3.1 YLs' performance on large-scale national assessment projects

Over the past decade, as early FL programs have become the norm rather than the exception in many countries, attainment targets have been defined and national assessments have been implemented (Rixon, 2013, 2016). A recent publication by the European Commission (2017, pp. 121–128) offers insights into the larger picture in 37 countries along three characteristics of national curricula and assessments: (1) which of the four language skills are given priority, (2) what minimum levels of attainment are set for students' first and second FLs, and (3) what language proficiency levels the examinations target. Curricula in ten countries give priority to listening and speaking, and in two of these reading is also included. All four language skills are emphasized in 20 countries, whereas no specific skill was specified in 7 countries. Over half of the countries offer a second FL (L3) in lower secondary schools, yet no information is shared on L3 examinations. Overall though, attainment targets tend to be at a lower level than in the first FL. Out of 37 countries included in the survey, 19 European countries reported a national assessment for YLs in their first FL and defined expected learning outcomes along the CEFR levels. The levels specified in the examinations for YLs ranged between A1 and B1; A2 is targeted in six, whereas three levels (A1, A2, B1) are listed in another

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

six countries; B1 is targeted in 13 countries at the end of lower secondary education. There is no data on how many YLs achieve these levels, nor are any diagnostic insights provided relative to YLs' strengths and weaknesses.

Two examples that report national assessment projects provide additional details into YLs' development: one in a European (Csapó & Nikolov, 2009; Nikolov, 2009) and another one in an African context (Hsieh, Ionescu, & Ho, 2017). In Hungary, YLs' proficiency was assessed on nationally representative samples of about ten thousand participants in years 6 and 8 in 2000 and in 2002 to measure their levels of FL proficiency, analyze how their language skills changed after two years, and what roles individual differences and variables of provision (number of weekly classes and years of FL learning) played. A larger sample learned English and a smaller one learned German. YLs' listening, reading, and writing skills in the L2 and reading in L1, as well as their inductive reasoning were assessed, and their attitudes and goals were surveyed (Csapó & Nikolov, 2009). The research project was followed by a national assessment using the same test specifications in 2003 (Nikolov, 2009). At all three data points, English learners' achievements were significantly higher across all skills than their peers' scores in German. YLs of English were more motivated, set higher goals in terms of what examination they aimed to take, and their grades in other school subjects were also higher than those of German learners. The relationships between proficiency and number of years of learning English or German were weak in both languages. Based on the datasets from the first two years, students' test scores in year 6 were the best predictor of L2 skills in year 8 (Csapó & Nikolov, 2009). Additionally, relationships between L2 and L1 reading weakened and between L2 skills strengthened over the years.

In Kenya, the English proficiency of 4,768 YLs who spoke 8 different L1s was assessed in years 3–7 (Hsieh et al., 2017). Researchers administered the TOEFL Primary Reading and Listening test—a large-scale standardized English language assessment for children between 8 and 11 years of age—in 51 schools to find out if students were ready to start learning school subjects in English. Students obtained higher scores across school grades; however, scores varied considerably by region. In year 7, about two thirds of YLs were at A2 level of proficiency, and very few achieved B1 level, the threshold assumed to be necessary for English-medium instruction.

3.2 Test validation projects

Test validation is a key component of quality assurance in test development. Validation studies provide crucial evidence as to whether an assessment measures what it is intended to measure and whether test scores are interpreted properly relative to the intended test uses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Compared to the number of assessment projects pertaining to YLs, relatively few publications focus on validation (e.g., for a discussion of validating national assessments, see Pižorn & Moe, 2012). The majority of validation projects have been conducted in the context of international, large-scale assessments as a means of ensuring test quality. By contrast, the field has seen relatively few validation projects on locally administered, small-scale assessments. Among the limited validation research on small-scale assessments are studies of locally administered speaking assessments and C-tests.

3.2.1 International proficiency examinations for YLs

Most of the validation work on tests for YLs has been carried out in the context of large-scale, standardized English language proficiency assessments (e.g., Bailey, 2005; Wolf & Butler, 2017; Papp, Rixon & Field, 2018). Examples of these assessments offered primarily by testing companies such as Educational Testing Service, Cambridge Assessment English, Pearson, and Michigan English Assessment, include the TOEFL® Young Student Series, the Cambridge Young Learners English Tests, PTE Young Learners, as well as the Michigan Young Learners English tests. These assessments have been supported to varying degrees by empirical research, in particular with regard to the validity of the test scores.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

In the context of international proficiency tests, a popular way to support the validity of test scores is the argument-based approach to validation (Chapelle, 2008; Kane, 2011). The main objective of an argument-based approach to validation is to provide empirical support for claims about qualities of test scores and their intended uses. For example, a test developer may claim that test scores are consistent or reliable. To provide support for claims, a series of statements (or warrants) is put forth which then need to be backed up by theoretical and/or empirical evidence such as, for example, reliability estimates. In contrast to warrants, rebuttals constitute alternative hypotheses that challenge claims. Hence, research produces evidence that may either support claims about an assessment or undermine them (i.e., supporting rebuttals).

One of the most comprehensive applications of an argument-based approach to validation in the context of a large-scale proficiency test for YLs is the empirical validity evidence gathered for the assessments in the TOEFL® Young Student Series (YSS). Following Chapelle (2008), the interpretive argument approach was applied from the outset, driving test development as well as the subsequent research agenda. In the test design frameworks for the *TOEFL Junior* test and the *TOEFL Primary* tests, So et al. (2015) and Cho et al. (2016) specified the intended populations and uses of the test. They discussed in detail the test design considerations, the TL use domains, skills, and the knowledge components that are assessed and how these components are operationalized for the purpose of YL assessment. Additionally, the frameworks laid out the inferences, warrants, and types of research needed to provide empirical evidence for validating the different uses of the TOEFL YSS assessments (for a detailed overview, see So et al., 2015, p. 22). Over the years, research has been conducted for these assessments in order to support the validity of test scores and their uses. For instance, in the context of the *TOEFL Primary* tests, studies have investigated content representativeness of the tests (Hsieh, 2016), YLs' perceptions of critical components of the assessment (Cho & So, 2014; Getman, Cho, & Luce, 2016), YLs' test-taking strategies (Gu & So, 2017), the relationship between test performance and test taker characteristics (Lee & Winke, 2018), the use of the test for different purposes such as a measure of progress (Cho & Blood, in press), and standard-setting studies mapping scores to the CEFR levels (Papageorgiou & Baron, 2017).

In addition to argument-based approaches, other approaches to validation such as Weir's (2005) sociocognitive framework for test development and validation have been utilized in the context of large-scale YL assessment. An example of this type of approach is Papp et al. (2018), who used a set of guiding questions put forth by Weir (2005) when discussing constructs and research related to the Cambridge Young Learners English Tests. While placing the main focus on describing language skills, test taker characteristics, and the language development of YLs, they also critically discuss a few validity studies carried out with or in relation to the Cambridge Young Learners English Tests. Among the research discussed are pilot studies that investigated aspects such as test delivery modes of paper-based and digitally delivered tests (Papp, Khabbazbashi, & Miller, 2012; Papp & Walczak, 2016;), investigations of test administration for speaking assessments in trios vs. in pairs (Papp, Street, Galaczi, Khalifa, & French, 2010), washback studies (e.g., Tsagari, 2012), and candidate performance and scoring (e.g., Marshall & Gutteridge, 2002).

A critical component in the process of test validation—in particular when it comes to large-scale, standardized proficiency assessments for YLs—is collaboration with local users and stakeholders to ensure the fit of a given assessment for the local group of learners. For example, Timpe-Laughlin (2018) systematically examined the fit between the EFL curriculum mandated by the ministry of education in the state of Berlin, Germany and the competencies and language skills assessed by the *TOEFL Junior Standard* test. To gauge the fit, curricula were reviewed and activities in textbooks were coded systematically for competences and language skills. Additionally, interviews were conducted with teachers at different schools to take into account their perspectives. While results suggested that the *TOEFL Junior Standard* test would be an appropriate measure for EFL learners in Berlin, findings also revealed critical areas in need of further research such as the limited availability of diagnostic information on score reports.

Overall, regardless of the validation approach utilized, it is crucial for international proficiency tests to make available empirical evidence that supports all claims made about a given assessment in order

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

to help stakeholders in their decision-making processes. For example, when implementing a standardized assessment, it is important to consider the empirical research behind a test and whether an assessment company can back up all claims they make relative to an assessment. Paying close attention to the purpose of a large-scale assessment and potentially conducting (additional) research in the local context may provide insights into whether a large-scale assessment is a good fit, provides the intended information, and can support FL teaching and learning. Additionally, collaborative research may prevent unfounded uses of a test, promote dialog, and maybe mitigate the fact that 'language assessment is often perceived as a technical field that should be left to measurement "experts"' (Papageorgiou & Bailey, 2019, p. xi).

3.2.2 Small-scale speaking assessments for YLs

In addition to validity research carried out in the context of international large-scale assessments, validity investigations have been conducted in relation to different types of smaller-scale assessments, among them in particular speaking tests. Most if not all curricula for YLs aim to develop speaking (Cameron, 2003; Edelenbos et al., 2006; McKay, 2006); therefore, projects investigating the validity of speaking tests are of great importance. Testing YLs' speaking ability poses unique challenges, both in classrooms and in large-scale projects. For instance, assessing speaking skills is time consuming, and special training is necessary to make sure teachers and/or test administrators can elicit children's best performances. Therefore, it is key to explore how children perform on oral tasks, how their speaking abilities develop, and how they can be assessed. In what follows in this section, we review a number of publications that feature test validation projects conducted in different countries for speaking assessments with English, French, and Japanese as TLs. These assessments include locally administered interview tasks, as well as national and international examinations.

Kondo-Brown (2004), for instance, assessed the speaking skills of 30 American YLs of Japanese as a FL in fourth grade. The study explored how interviewers offered support, how the scaffolding provided by the interviewers impacted children's responses, negotiation of meaning processes, and students' scores. The oral tasks were based on the curriculum and teachers' classroom practices. Without clear guidance, interviewers offered inconsistent scaffolding, most often correction, and YLs had no opportunities to negotiate meaning. Supported performances tended to get better scores. This study and a similar project conducted with Greek pre-schoolers whose speaking performances were assessed with and without help from their teacher (Griva & Sivropoulou, 2009) raise important questions about scaffolding children's speaking skills: How should teachers assess what YLs can do with support today so that it will to lead to better performance without help tomorrow? Also, to what extent do we introduce construct-irrelevant variance when providing support during interviews when scaffolding is natural and authentic in oral interactions with children?

In a similar interview format, about 110 Swiss YLs of English were tested after one year of English in year 3 to gauge their oral interaction skills (Haenni Hoti, Heinzmann, & Müller, 2009). In two tasks children spoke with an adult, whereas in a role-play they interacted in pairs. Most of the YLs were able to do the tasks and fully or partially achieved A1.1 level in speaking, although considerable differences were found in their oral skills. For example, while most children used one-word utterances, a few high achievers produced utterances of nine or more words. Analyses of YLs' task achievement, interaction strategies, complexity of utterances, and range of vocabulary offered insights into the discourse they used. This strategy allowed the authors to fine-tune expectations and evaluate the assessments.

In Croatia, 24 high, average, and low ability YLs were selected from four schools to assess their speaking skills in years 5–8 (age 11–14) to map how they were related to their motivation and self-concept (Mihaljević Djigunović, 2016). The difficulty of the speaking tests (picture description tasks and interviews) increased over the four years to reflect curricular requirements, but the same assessment criteria were used. Children's test scores indicated slightly different trajectories along task achievement, vocabulary, accuracy, and fluency on the two oral tests. Their self-concept, an affective variable reflecting how good they thought they were as language learners, changed along similar lines resulting in a U-shape pattern, whereas their motivation showed an inverted U shape.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

3.2.3 C-tests for YLs

In addition to speaking assessments, the field is beginning to see validation research conducted on other types of YL assessment formats such as a C-test, a type of gap-filling test that is based on the principle of reduced redundancy and measures general language proficiency (Klein-Braley, 1997). For example, validating an integrated task was the aim of a study on 201 German fourth graders learning English. Porsch and Wilden (2017) designed four short C-tests based on adapted texts and analyzed relationships between YL's school grades in English and test-taker strategies. They found statistically significant relationships (.40–.50) between grades and scores, and frequency of strategy use conducive to reading comprehension. They did not investigate how much practice children needed to do C-tests or how scores compared to other reading comprehension tests. Additional research may want to investigate via think aloud or eye-tracking methodology how YLs approach and engage with C-tests.

3.3 Comparative YL assessment projects across ages, contexts, and types of programs

Much of assessment OF learning research has been conducted to compare achievements across different YL programs such as those in which students start at an earlier and later age. In addition, YLs' achievements have been compared across countries as well as across different types of YL education programs. In particular, outcomes of FL and CLIL programs have been investigated.

3.3.1 Comparative assessments of YLs in early and later start programs

As new FL programs were gradually introduced that targeted increasingly younger learners, it made sense to compare YLs' L2 skills on the basis of the age at which they began programs. Such a comparative research design could offer evidence in what domains YLs are better in the L2 and how implicit and explicit learning emerge. Thus far, we have witnessed a number of projects across Europe: in the 1990s, studies were conducted in Croatia (Mihaljević Djigunović & Vilke, 2000) and in Spain (García Mayo, 2003; Muñoz, 2006), followed by research in Germany (Wilden & Porsch, 2016; Jaekel, Schurig, Florian, & Ritter, 2017), Switzerland (Pfenninger & Singleton, 2017, 2018), and Denmark (Fenyvesi, Hansen, & Cadierno, 2018) in the 2000s.

In Croatia (Mihaljević Djigunović & Vilke, 2000), over 1,000 YLs (age 6–7) started to learn English, French, German, and Italian in their first grade in four or five hours a week in the first four years; then, from year 5, similarly to control groups who started in fourth grade (age 10–11), all YLs had three weekly classes. YLs were assessed after eight and five years, respectively, in their last year in primary school. Children in the early start cohort were significantly better at pronunciation, orthography, vocabulary tasks, and a C-test, and slightly better at reading comprehension. The control group outperformed their peers on a test of cultural elements. YLs' oral skills assessed by a single interview task were better in the early start groups, although significant variability was observed in all groups.

In Spain, two studies involved bilingual YLs who started English as their third language at the ages of 4, 8, and 11. A similar research design was used to compare 135 Basque-Spanish (Cenoz, 2003; García Mayo, 2003) and over 2,000 Catalan-Spanish (Muñoz, 2003, 2006) children in three cohorts after about 200, 400, and 700 hours of EFL instruction. In order to compare results of the different age groups, the same tests were used to assess participants' English speaking, listening, reading, and writing skills (Muñoz, 2003, p. 167). The key variables in both projects were starting age and amount of formal instruction. The tests were not based on the YLs' respective curricula, but they targeted what all groups were expected to be able to do. Some tests were meaning-focused (e.g., story-telling based on pictures, C-test on a well-known fairy tale, matching sentences in dialogs with pictures, letter writing to host family), whereas others focused on form (e.g., fill in blanks with 'auxiliaries, pronouns, quantifiers' and 'choose adverbs to describe eating habits' Cenoz, 2003, p. 84). Overall, the later the groups started learning English, the better they performed on the tests at each point of measurement. In both projects, lower levels of cognitive skills were identified as the main reason for the slower rate of progress among YLs (Cenoz, 2003; García Lecumberri & Gallardo del Puerto, 2003; García Mayo, 2003;

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Lasagabaster & Doiz, 2003; Muñoz, 2003). This outcome might be due to the lack of age-appropriate tests, as many assessments seemed to favor cognitively more mature age groups, thus potentially failing to tap into what YLs at earlier stages were able to do well. Unfortunately, no data were collected on factors that must also have impacted outcomes such as what was taught in the courses, how proficient the teachers were, and how much English they used for what purposes in the classroom.

In Germany, a recent change in language policy motivated a large-scale assessment project comparing the proficiency of YLs who started English in years 1 and 3 (Wilden & Porsch, 2016; Jaekel et al., 2017). The listening and reading comprehension skills of more than 5,000 YLs were tested in year 5 after learning English for two vs. three and a half years and in year 7 after two more years in grammar schools. In addition to testing participants' English development, the researchers also assessed students' literacy skills, their socioeconomic status (SES), and whether German was their L1 to examine what factors contributed to YLs' English scores. In year 5, listening comprehension was tested by multiple-choice items on picture recognition and sentence completion in German, whereas reading comprehension was assessed by multiple-choice and open items. In year 7, both the listening and reading comprehension tests included open and multiple-choice items, and some items were identical in years 5 and 7. In year 5, YLs in the earlier start cohort performed significantly better at the English tests than their peers starting later, and scores on German reading comprehension tests contributed to the outcomes indicating the importance of an underlying language ability. However, in year 7, late starters outperformed their peers, which cast serious doubts on the value of starting English early (Jaekel et al., 2017). Interestingly, test results in year 9 showed a different picture: early starters achieved significantly higher scores than late starters (Jaekel, p.c., July 17, 2019), making the outcome in year 7 hard to explain.

Lowering the starting age of mandatory EFL education triggered yet another comparative study on YLs starting at different times in Denmark. A total of 276 Danish YLs were assessed in two groups after learning English for a year in grades 1 (age 7–8) and 3 (age 9–10) (Fenyvesi et al., 2018) to assess how their development in receptive vocabulary and grammar interacted with individual differences and other variables. The Peabody Picture Vocabulary Test PPVT-4 and the Test for Reception of Grammar TROG-2 (Bishop, 2003) were used twice, following Unsworth, Persson, Prins, and de Bot (2014): at the beginning and the end of their first year of English. Both tests used pictures and YLs had to choose the correct word or sentence from four options. Children in the early start group achieved significantly lower scores on both tests at both points of measurement than their peers starting in year 3. Both groups achieved significantly better results after a year, but the rate of learning was not higher for the older group. Interestingly, older YLs at the beginning of formal EFL learning achieved similar scores to those in the early start group after a year of EFL. This result indicates how children benefit from extramural English activities (see a more detailed discussion in Section 3.5).

In one of the most comprehensive, longitudinal studies, over 600 Swiss YLs participated in a study to find out why older learners tend to perform better in classroom settings (Pfenninger & Singleton, 2017, p. 2018). In addition to age of onset (when YLs started learning English), they included the impact of YLs' bilingualism/biliteracy and their families' direct and indirect support. Participants were 325 early (age 8) and 311 later (age 13) starters; they were assessed after five years and six months of English, respectively. Early and later starters included subgroups of monolingual Swiss German children, simultaneous bilinguals (biliterate in one or two languages), and sequential bilinguals (illiterate in their L1; proficient in Swiss German). As the authors point out, Swiss curricula target B1 at the end of compulsory education, and they intended to use the same tests four years later. Therefore, the proficiency of all YLs was tested by (1) two listening comprehension tests at B2 level, (2) *Receptive Vocabulary Levels Test* (Schmitt, Schmitt, & Clapham, 2001); (3) *Productive Vocabulary Size Test* (Laufer & Nation, 1999); (4) an argumentative essay on talent shows; (5) two oral (retelling and spot-the-difference) tasks that were evaluated based on four criteria: lexical richness, syntactic complexity, fluency, and accuracy; and (6) a grammaticality judgment task. However, they did not use the listening and the productive vocabulary tests in the first round. Using multilevel modeling, they found that in neither written nor oral production did the early starters outperform the late starters.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Over a period of five years, late starters were able to catch up with early starters insofar as late starters needed only six years to achieve the same level as early starters had after eleven years—a result that Pfenninger and Singleton (2017) attributed to strategic learning and motivation. Additionally, findings showed that biliterate students scored higher than their monoliterate peers, and family involvement was always better than no involvement. A combination of biliteracy and family support was found to be particularly effective. Besides these factors, random effects made up much of the variance, indicating that it is almost impossible to integrate all factors into models.

In all of the comparative studies reviewed in Section 3.3, the research design impacted when and how YLs were tested. Findings indicate that the tests that were used were intended to tap into both implicit and explicit learning, but the exact emphasis on each is unclear. It is remarkable that hardly any of the tests were aligned with early FL curricula or the age-related characteristics of the YLs. Hence, it is quite likely that the tests were more appropriate for more mature learners, while failing to elicit the full potential YLs' FL achievements.

3.3.2 Comparative assessments of YLs in different educational contexts

In addition to comparisons with regard to age or age of onset, several studies that focused on the assessment OF learning were conducted to examine impacts of learning environments. In this section, we discuss assessment projects comparing YLs' achievements in early FL programs. In Section 3.3.3, we review research that focused on different types of early FL programs, in particular CLIL contexts.

The most ambitious comparative project, Early Language Learning in Europe (ELLiE), assessed YLs from seven European countries over three years. The project applied mixed methods and involved about 1,400 children from Croatia, England, Italy, Netherlands, Poland, Spain, and Sweden. In particular, Szpotowicz and Lindgren (2011) analyzed what level YLs achieved in the first years of learning an FL. In the first two years, YLs' listening and speaking skills were assessed, whereas in the third year, their reading comprehension was also tested. The number of items and the level of difficulty increased in the listening and speaking tests every year. Oral production was elicited by prompts in YLs' L1 to assess what they would say in a restaurant situation. The publication includes sample oral performances and graphic presentation of data, but it lacks any statistical analyses. The authors claimed that by age 11, average YLs made good progress toward achieving A1 level, the typical attainment target in YL curricula, but they emphasized high variability among learners.

In Croatia and Hungary, Mihaljević Djigunović, Nikolov, and Ottó (2008) compared Croatian and Hungarian YLs' performances on the same EFL tests in the last year of their primary studies (age 14) to find out how length of instruction in years, frequency of weekly classes, and size of group impacted YLs proficiency. They used ten tasks to assess all four language skills and pragmatics. Although Hungarian students started learning English earlier, in smaller groups, and in more hours overall, Croatian EFL learners were significantly better at listening and reading comprehension, most probably due to less variation in their curricula and more exposure to media in English (e.g., undubbed television).

Over the course of three years, Peng and Zheng (2016) compared two groups of young EFL learners from the same elementary school in Chongqing, China. Teachers used two different textbooks and corresponding assessments in the two learner groups. One group used the *PEP English* ($n = 304$) which tends to foreground vocabulary and grammar, the other used the *Oxford English* ($n = 194$), which was identified as having a stronger focus on communicative abilities. Students were assessed in years 4, 5, and 6 using the assessments that accompany the materials they learned from. Overall, scores declined slightly over the years. To triangulate the data, teachers were interviewed to reflect on the coursebooks, the tests, the results, and the difficulties they faced when assessing YLs. The authors offered valuable insights into how children's performances decreased over the years, indicating in particular motivational issues in the group that used *PEP English* with its focus on grammar and vocabulary.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

3.3.3 Comparative assessments of YLs in FL and content-based programs

As early FL programs and approaches to teaching FLs to YLs vary considerably (see Introduction), contents and goals of early FL learning do so as well. In FL programs, the achievement targets in knowledge and skills are defined in terms of the L2. In CLIL programs, by definition, the aims include knowledge and skills in the FL as well as in the subject (content area) studied in the FL. In short, there are considerable differences between the goals and contents of YL instruction in FL programs and CLIL programs.

From a measurement perspective, it is important that the construct of an assessment is in line with curricular goals. Therefore, summative assessments aimed at capturing YLs' achievements should operationalize aspects of the constructs that reflect the various curricular objectives. In FL programs, the construct should be operationalized in terms of FL learning, while in CLIL programs, both the FL proficiency AND the subject area domains should be accounted for in the test construct. However, this juxtaposition of FL and content learning is not necessarily reflected in assessment projects that focus on CLIL programs. For example, Agustín Llach (2015) compared two groups of Spanish fourth graders' (72 CLIL, Science; 68 non-CLIL) productive and receptive vocabulary profiles on an age-appropriate letter writing task (introduce yourself to a host family) and the 2k *Vocabulary Levels Test* (VLT; Schmitt et al., 2001). The VLT Receptive Vocabulary Test uses multiple matching items organized in ten groups of six words corresponding to three definitions. Scores ranged between 0 and 30. No significant differences were found between the two groups, despite 281 hours of CLIL in addition to 419 hours of EFL. On the writing task, YLs' vocabulary profiles were drawn up based on type/token ratios and lexical density. In both groups phonetic spelling was frequent and low scores on cognates were typical, indicating, in our view, low level of vocabulary knowledge and use of strategies. The author suggested that children's lack of cognitive skills must have been responsible for not benefiting from CLIL, although, most probably, the two selected tests must have played a role in not capturing vocabulary from YLs' CLIL classrooms.

In a longitudinal study, Agustín Llach and Canga Alonso (2016) assessed growth in receptive vocabulary of 58 CLIL and 49 non-CLIL Spanish learners of English in fourth, fifth and sixth grade by using the VLT test. After three years, the differences were modest, but vocabulary knowledge was significantly higher for CLIL learners. The rate of growth was quite similar in the two groups: 914 vs. 827 words, respectively, in year 6 after 944 vs. 629 hours of instruction. No tests were used to tap into the vocabulary taught in the CLIL classes.

A similar research design was used in Finland by Merikivi and Pietilä (2014) to compare CLIL (n = 75) and non-CLIL (n = 74) sixth graders' (age 13) English vocabulary. In this context, CLIL instruction was not preceded by or complemented by EFL learning. YLs in the CLIL group had 2,600 hours of English and in the non-CLIL group 330 hours. In addition to the VLT, the Productive VLT (PVLT) was also used (version 2; YLs fill in parts of missing words in sentences). CLIL learners' receptive and productive vocabulary scores were significantly higher (4,505; 1,853, respectively) than those of their non-CLIL peers (2,271; 788). Nevertheless, results on the VLT are directly comparable with those of the Spanish sixth-graders (Agustín Llach & Canga Alonso, 2016). Although Finnish belongs to a different (Finno-Ugric) language family, and English and Spanish are both Indo-European languages, Finnish sixth-graders achieved much higher scores than their Spanish peers not only in the CLIL group but also in the EFL group (2,271 vs. 827). These outcomes must have resulted from the quality of instruction and should be examined further.

A different approach was used by Tragant, Marsol, Serrano, and Llanes (2016) with third-graders in Spain. They assessed a group (n = 22) of eight–nine-year-old boys over two semesters. In the first semester, YLs learned EFL, whereas in the second one they studied Science in English. The study aimed to measure how much of the taught vocabulary was learned. The productive vocabulary test used before and after starting EFL and Science included 30 nouns taken from the course materials. YLs were asked to write the meanings of words next to small visuals; the initial letters of each word were given as prompts. Children's vocabulary developed in both programs, but they learned significantly more words in the EFL lessons. An analysis of the EFL and Science teaching materials

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

revealed more and more abstract and technical vocabulary in the CLIL materials. Classroom observations indicated extensive L1 use in CLIL classes, thus, pointing to important differences in teaching impacting the results in YLs' vocabulary.

In a large-scale comparative assessment, CLIL and non-CLIL fourth graders' (age 9–10) overall English proficiency was assessed (de Diezmas, 2016). All YLs (over 1,900 CLIL learners and 17,100 non-CLIL students) in a region of Spain took the same four tests to compare their four language skills. All participants learned English in a total 730 hours and the CLIL students had an additional 250 hours. In the listening test, YLs watched a short video twice about hygiene habits and answered six questions. In the oral test, they were given pictures of two bedrooms; they chose one, described it in writing, and then, in groups of two or three, they interacted orally and justified their choices. The reading test included a short email and six multiple-choice items, whereas the writing test comprised writing an outline and then the actual article with the help of a dictionary. The tests were designed to match the EFL curriculum (but not the subject domains learned in English), and all learners were assessed along the same criteria developed for the productive tasks. The only statistically significant difference between the two groups was found on the interactive oral task: CLIL students outperformed their non-CLIL peers.

Content-based (called dual-language) programs, in which some subjects are taught in English and German, are also popular in Hungarian primary schools. The government launched a high-stakes examination at the end of years 6 and 8 (age 12 and 14) to establish the ratio of YLs at A2 and B1 levels, respectively (Nikolov & Szabó, 2015; all tests for English and German used in 2014 and 2015 are available at https://www.iris-database.org). Unless 60% of the students achieve the prescribed levels for three years, schools must close their programs. In year six, 1,420 English and 402 German learners, and in year eight, 819 and 270 YLs, respectively, took the exams assessing their listening, reading, and writing skills in 2014. Test booklets at both levels included six tasks: two tests of ten items for listening and reading (multiple matching), and two short email writing tasks assessed along set criteria (communicative content, richness of vocabulary, grammatical accuracy). Significantly better results were found for English than for German in both years and large differences were found across schools. Overall, the majority achieved the required levels, although at a few schools, achievements were below expectations. Unfortunately, no data were collected on what subjects the participants learned, for how many years and in how many classes per week, and neither content knowledge nor speaking were assessed.

To summarize, out of the six studies on content-based programs (Table 4), three small-scale studies aimed to measure and compare YLs' English vocabulary in CLIL and non-CLIL groups in Spain (Agustín Llach, 2015; Agustín Llach & Canga Alonso, 2016) and in Finland (Merikivi & Pietilä, 2014), whereas another project on Spanish YLs (Tragant et al., 2016) collected data with multiple instruments and analyzed both EFL and Science coursebooks as well as classroom observation data. Two large-scale studies implemented in Spain and Hungary assessed multiple L2 skills (Nikolov & Szabó, 2015; de Diezmas, 2016). Although content-based English programs for YLs have gained ground in recent years, five of six CLIL studies assess gains in general proficiency, and four are limited to vocabulary and compare CLIL and non-CLIL groups. In other words, most of the CLIL-related assessment studies focused on vocabulary testing. However, the vocabulary tests used in the first three studies were not designed for young EFL learners and had little to do with the CLIL vocabulary the children had learned. In the Finnish context, the massive amount of exposure must have resulted in the CLIL learners' impressive scores. The overall findings of the only large-scale study comparing CLIL and non-CLIL cohorts of fourth-graders (de Diezmas, 2016) did not find evidence that early CLIL contributes to YLs' proficiency in important ways, although the impact may become measurable in later years. The other studies did not support the widespread popularity of and enthusiasm toward CLIL, while the outcomes of the other studies also revealed that the issues are more complex (what goes on in the classrooms) and the tests most probably fail to measure what YLs gained in the programs—issues that seem to be underpinned by a recent project that tapped into content learning (Fernández-Sanjurjo, Fernández-Costales, & Arias Blanco, 2017; Fernandez-Sanjurjo, Arias Blanco,

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Table 4. Language components in CLIL and non-CLIL program assessments for YLs

	Vocabulary	Writing	Listening	Reading	Speaking and interaction	Classroom observation
Agustín Llach (2015)	Receptive vocabulary test (VLT)	1 letter writing task	—	—	—	—
Agustín Llach and Canga Alonso (2016)	Receptive vocabulary test (VLT)	—	—	—	—	—
de Diezmas (2016)		1 task: write a short article	1 task on video	1 task on email	1 interactive oral task	—
Merikivi and Pietilä (2014)	Receptive and productive vocabulary (VLT; PVLT)	—	—	—	—	—
Nikolov and Szabó (2015)		2 writing tasks	2 listening tasks	2 reading tasks	—	—
Tragant et al. (2016)	Vocabulary tests on EFL and Science teaching materials	—	—	—	—	Interaction and discourses

& Fernandez-Costales, 2018). In this project, Spanish YLs' knowledge of Science was assessed in a study involving representative samples of 6th graders ($n = 709$) in English CLIL and non-CLIL groups. All participants took the same Science test based on the curriculum in their L1. Students learning content in Spanish achieved slightly but statistically significantly better results than their peers learning Science in English. Thus, the language used in the test may have influenced outcomes—an area for further research.

Despite these findings, thousands of YLs attend CLIL programs and no publications were found to offer insights into what young CLIL learners can do in English and the subjects they learn in English, how teachers assess them in the subjects taught in English, and how content learning interacts with FL learning and other variables. Data should be also collected on what achievement targets CLIL programs set, if they are defined separately or integrated into FL learning and the subject areas, and to what extent YLs perform at the expected levels on tests integrating the FL AND content.

3.4 Experimental studies assessing attainments of young FL learners, including pre-school children

In a few recent assessment studies, researchers focused on how certain teaching techniques and tasks work with YLs, and in particular what children learned by means of certain interventions. These publications tend to assume a straightforward relationship between what teachers do and what YLs achieve in a short time frame. In these studies, the most frequently assessed domain is YLs' L2 vocabulary, an exciting but highly problematic recent trend, especially if young children between two and three and six at pre-schools are investigated. For example, Coyle and Gómez Gracia (2014) involved 25 Spanish children (age 5) in three short English lessons to teach five nouns featured in a song and related activities. Children's receptive and productive knowledge of the target words was individually tested before and after the lessons, and then, five weeks later. The reported outcomes were minimal. For example, four children could name between one and five objects in the delayed posttest, whereas others could not recall any words. The results, however, were framed positively, highlighting 'a steady increase in the receptive vocabulary' (p. 280) recognized after three and five weeks indicated a mean of 1 and 1.72 words, respectively. While it is debatable whether this finding can be regarded as a steady increase, some ethical concerns also need to be highlighted. For example, the performance of one child who felt sleepy during the sessions was labeled as 'poor' (Coyle & Gómez Gracia, 2014, p. 283), —a highly inappropriate way of referring to the developing skills of very young learners.

In a similar intervention study, 64 Chinese children (age 4–5) participated in Davis and Fan's (2016) study to examine how songs and choral repetition contribute to learning vocabulary in 15 lessons of 40 minutes over 7 weeks. Children heard the same 15 short sentences in a song, repeat, and control treatment in a different sequence. Then, they were tested on a productive vocabulary test of the 15 items before and after the lessons. They were invited to say what they could see in 15 visual prompts. Results were reported in mean length of utterance, an indication that the authors expected relatively long answers. However, the song and choral repetition conditions resulted in similar outcomes: most children said either nothing or a single word. While the elicitation technique was not age-appropriate, the findings clearly indicate that most children were not ready to respond. That finding could be an artefact of the teaching approach which did not go beyond drills and included no meaning-making activities.

Three methods of vocabulary teaching through reading storybooks (explicit: rich, embedded, and incidental) were used by Yeung, Ng, and King (2016). Thirty Cantonese speaking children (age 5) participated in all conditions in three 30-minute sessions per week for three weeks. Three different storybooks were used to teach four target words each. An oral receptive vocabulary test (PPVT) with thirty-six items was used before the project and three tests were applied each time on the twelve words before and after listening to the stories, and eight weeks later. First, children were asked to explain the meaning of the twelve target words (Cantonese was accepted); then, their comprehension was measured; finally, they answered yes/no questions in one-on-one settings with their teacher. Children scored better in the first, rich method condition, but no difference was found between

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

embedded and incidental methods. Again, overall, the outcomes were minimal: the highest mean was 6 (out of 12 words), while some children recalled no words at all.

Seventeen two- and three-year-old Spanish children participated in a study conducted by Albaladejo Albaladejo, Coyle, and de Larios (2018) to find out how they learned English nouns via three types of activities. They listened to stories, songs, and both so that they heard each word between three and nine times. Children took a pre-test, a posttest, and a delayed posttest (three weeks after each condition) on five words (total = 15) they heard in each of the three conditions over three weeks. The authors reported that participants learned most words (between two and three words) from stories. However, three of the words included in the assessment were cognates, calling into question the validity of the test. In the song condition, four children recalled between one and four words, whereas others could not remember any.

A different research design was applied in a study covering a much longer period. Greek kindergarteners' achievements were tested by Griva and Sivropoulou (2009) in two groups ($n = 14$: age 4–5; $n = 18$: age 5–6) before and after an English course lasting eight months. The same three oral tests were used: children were asked to name twenty items on a poster, point to three actions their teacher described, and complete three sentences by looking at what the teacher pointed to. There was statistically significant improvement of the children's performance over time: for instance, the mean of the word recall test was 5 on the pre-test and 10.3 on the posttest. An innovative element was applied in scoring: the scores differed depending on speaking WITH or WITHOUT help. Some children could do the tasks before starting the course; others did not score on any of the posttests. Children's productive vocabulary was assessed by using many more items beyond their receptive skills (23 vs. 3), and it is unclear how being tested on many unfamiliar words impacted children achieving low or no scores.

Very young learners participated in treatment and control groups in a two-year longitudinal project in the Netherlands. Unsworth et al. (2014) assessed 168 Dutch pre-school learners' (mean age 4.4) receptive vocabulary and grammar with the PPVT-4 and the Test for Reception of Grammar, TROG-2 (Bishop, 2003) tests at three points: before starting EFL, after one, and again after two years. They aimed to find out how much children developed in English, and how learner- and teaching-related variables as well as the teachers' proficiency impacted their scores. Children performed statistically significantly better after two years on both tests and their scores depended more on whether their teacher's proficiency was at least at B2 level on the CEFR (irrespective of native and non-native speakers) than on the amount of their weekly exposure to English. Interestingly, some children knew quite a few words at the very beginning, potentially illustrating the impact of English in the lives of Dutch kindergarten children.

Although word reading is a key component of YLs' L2 reading skills, hardly any study focused on how it develops in FL programs. Some of these projects included receptive vocabulary tests, but authors failed to point out that they involve word reading in tests in which the printed word and its meaning are to be matched. A study that addressed the matter to a certain extent (Yeung, Siegel, & Chan, 2013) focused on how training in phonological awareness contributed to 72 Hong Kong pre-schoolers' English oral skills, word reading, and spelling skills. The language-enrichment group ($n = 38$, age 5) followed a special program, while the control group followed a typical holistic syllabus for 12 weeks in two to three 30-minute sessions per week. Multiple tests were used to assess outcomes: some tests tapped into what YLs learned (e.g., word reading, naming objects in pictures), while others were unrelated to the program (writing of unknown words, PPVT English Receptive Vocabulary Test). Additionally, five tests measured phonemic awareness (e.g., syllable deletion, rhyme detection). Although the focus of the experimental study was on phonological training, the authors emphasized oral skills and productive vocabulary as key predictors of success in word reading.

Overall, the studies involving very young children highlighted how challenging age-appropriate assessment can be. Although the research questions make sense, the validity of the tests deployed is problematic. The findings show that pre-school children's FL skills develop at a very slow rate, and there are important differences within the groups at this highly vulnerable age. None of the results

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

can be generalized to other YL groups and it is unclear how the outcomes can be reasonably applied. Research into age-appropriate teaching techniques should collect varied data from multiple sources, including observations and teachers' perceptions. Also, whether very young children should be subject to summative assessment can be questioned. Some children might be overwhelmed by formal testing. Additionally, feeling unsuccessful or being labeled 'slow' or 'poor' may negatively impact children's self-confidence, self-image, attitudes, and development over time.

3.5 Projects on extracurricular exposure to and interaction in the FL

An emerging trend in testing YLs focuses on incidental learning resulting from extramural activities, as many children know more than they are taught in school, mostly from exposure to media and gaming (Sundqvist & Sylvén, 2016). Two studies assessed YLs vocabulary profiles prior to instructed L2 learning and their findings are in line with results involving pre-school children (Unsworth et al., 2014).

Flemish learners' incidental vocabulary learning was assessed in two projects. De Wilde and Eyckmans (2017) tested 30 Flemish (12 Dutch L1; 18 bilingual) YLs in sixth grade before starting English as their third language. Participants' vocabulary was tested with the PPVT-4, whereas their overall proficiency by Cambridge English Test for Young Learners—Flyers. Two distinct profiles emerged in the test results: YLs with quite good English and those with hardly any English. On the receptive vocabulary test, 22 YLs knew over half of the words. On the listening test, 40% and on the other components between 10 and 25% of the YLs achieved A2 level with no formal English instruction mostly from gaming and watching subtitled programs.

In a larger-scale study, Puimège and Peters (2019) involved 560 Dutch L1 (including 24% multilingual) YLs in three age groups (age 10, 11, 12) prior to starting English in school. They tested YLs' meaning recognition and recall on the Picture Vocabulary Size Test (Anthony & Nation, 2017) and their Dutch vocabulary. Additionally, they included learner- and word-related variables. The impact of passive exposure (watching TV and songs) was statistically significant in receptive vocabulary scores; gaming and video streaming impacted meaning recall scores. At age 12, YLs' estimated receptive and productive vocabulary size was over 3,000 and 2,000, respectively. Cognateness and frequency of the vocabulary items were the best predictors of test scores.

3.6 Impact of YL assessment

Assessment at any age may impact participants in positive and unintended ways. YLs are particularly vulnerable (Rea-Dickins, 2000; McKay, 2006) and sensitive to criticism and failure; therefore, special care must be taken to avoid negative impact, as testing itself as well as its results may interact with how children see themselves and how they are seen as individuals. Papp et al. (2018) discuss the consequential validity of the Cambridge Young Learners English Tests. They point out that in many contexts around the globe where test results had an important role in gaining access to secondary schooling, the 'impact of English tests and exams could be very great' (p. 558). English test results determining at age 10 or 11 if the child can go on to secondary education will impact children's life chances. As summarized by Rixon (2013), a survey conducted in 64 countries by the British Council found that English test results played an important role in YLs' future education in about a quarter of the contexts, and in most countries private, extracurricular classes were perceived as offering better learning opportunities in smaller classes with better educated teachers than in state schools. Thus, equity is a serious issue related to YLs' learning and assessment, especially in the case of EFL.

Test results are often used for gate keeping purposes to allow high achievers to enter more intensive and good quality programs contributing to the Matthew effect (i.e., the rich get richer and the poor get poorer) and inequality in opportunities. For example, Chik and Besser (2011) analyzed the unanticipated social consequences of international examinations for YLs in Hong Kong. By asking all stakeholders (parents, principals, school administrators, YLs) they revealed that international test certificates empowered privileged YLs by ensuring access to better English-medium schools, whereas

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

less fortunate children whose parents could not afford such tests were disadvantaged. Thus, unequal access to language tests enhanced the inbuilt inequality in the education system.

Testing may empower and motivate children, as well as induce their anxiety, reduce their motivation, and threaten their self-esteem (Nikolov, 2016a). Over 100 children (age 6–8) participated in focus group interviews and drew pictures about their test-taking experiences in Hong Kong (Carless & Lam, 2014). The findings showed that although many children felt happy about their high achievements or relieved after having taken tests, negative emotions outweighed positive ones, as most children reported fear and anxiety. Additionally, Bacsa and Csíkos (2016) found that anxiety played an important role in YLs' EFL listening comprehension test results. Overall, these points have been discussed in an overview of the relationships between affect and assessment by Mihaljević Djigunović (2019), who pointed out that assessment may lead to demotivation, induce YLs' anxiety, and impact their self-concept negatively over time. She cited a Croatian 12-year-old: 'Each time our teacher announces a test, I panic. While preparing for the test at home, I feel nervous all the time.' (p. 25).

Some ethical issues also need to be pointed out. Tasks that are expected to be frustratingly difficult should be avoided. This is especially true in studies where authors aim to implement a publishable study, but they fail to bear in mind how the use of tests far beyond what children can do may impact them. Also, assessment is time consuming; many studies engage YLs in taking tests for a long time (often for hours). Thus, it is highly probable that some tests are beyond YLs' average attention span. In addition, testing children takes precious teaching time away from developing their FL.

Score reporting and use also concern impact; however, few studies discuss how test results are reported, how teachers, parents, and school administrators use them, and how they impact YLs' lives. International proficiency exams tend to devote some discussion to their score reporting (e.g., Papageorgiou, Xi, Morgan, & So, 2015; Papp et al., 2018, pp. 547–587), but generally, much more attention should be devoted to WHY YLs are assessed and WHAT happens afterwards. Overall, the main purpose of assessment OF learning should also be to make sure that children benefit from it.

4. Assessment for learning-oriented research

In this section we discuss studies with a potential for assessment FOR learning, a concept similar to learning-oriented, formative, and diagnostic assessment (Alderson, 2005; Black & Wiliam, 1998; Wiliam, 2011). In summative assessment the aim is to measure to what extent YLs have mastered what they were taught, or in the case of proficiency tests, to what extent they have achieved the targets in the FL along certain criteria. Diagnostic assessment, however, is geared toward identifying strengths and weaknesses so that challenging areas can be targeted and further practice provided. This latter approach is particularly important for YLs, as they need substantial encouragement in the form of frequent, immediate, and motivating feedback on where they are in their learning journey so that their vulnerable motivation can be maintained, while demotivation and anxiety can be avoided (Mihaljević Djigunović & Nikolov, 2019). Moreover, teachers need to know where YLs' strengths and weaknesses are so that they can tailor their instruction to their needs and thus facilitate learning (Nikolov, 2016b). In other words, appropriate diagnostic assessment is an integral part of good classroom practice.

In the following, we first review how teachers of YLs apply their competences when they assess their YLs and how they use diagnostic feedback to facilitate YLs' language development. Then, we will look at how alternative learning-oriented assessments are used, focusing on how learners have been involved in self- and peer assessment. Then we move on to how YLs apply test-taking strategies, and how certain task types promote learning.

4.1 Teachers' beliefs and assessment practices

Despite the abundance of publications on YL assessment, little is known about the ways in which teachers assess YLs' FL skills in their classrooms. Teachers' language assessment competence and literacy,

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

their knowledge, skills, and abilities they need in order to implement language assessment activities and to interpret YLs' FL skills and development (Fulcher, 2012) are, unfortunately, rarely studied. In one of the earlier studies, Edelenbos and Kubanek-German (2004, p. 260) identified teacher's 'diagnostic competence', that is, 'the ability to interpret students' foreign language growth, to skillfully deal with assessment material and to provide students with appropriate help in response to this diagnosis', as a key area and proposed descriptors of diagnostic competence. In particular, their research showed how teachers' language assessment literacy and diagnostic competences interact with their beliefs, attitudes, and knowledge about YLs' development.

Taking a closer look, Butler (2009) explored in her experimental case study how South Korean primary and secondary school teachers ($n = 26 + 23$) assessed 4 sixth-graders' English performance on two interactive oral tasks. First, teachers were asked to assess learners holistically, then to choose a few criteria from a list (e.g., fluency, accuracy, pronunciation, task completion, confidence in talking, motivation) and to assess them again relative to those criteria. In a follow-up activity, teachers discussed which criteria they chose, why, and how they arrived at their scores. Substantial variation was found across both groups of teachers in both types of assessments; primary school teachers were more concerned with YLs' motivation, fluency, and confidence in talking, while secondary teachers tended to focus on accuracy and less on affective traits. Even teachers in both groups who chose the same criteria varied in their responses as to why and how they applied them, reflecting their beliefs about language learning.

Feedback given by 41 Hungarian EFL teachers of YLs on 300 diagnostic tests was analyzed by Nikolov (2017). Teachers were invited to evaluate 20 tasks they volunteered to pilot with their students (age 6–13) to triangulate YLs' feedback and test scores. The comments revealed respondents' beliefs and practices. Many teachers found scoring time consuming and failed to see the value of giving YLs immediate feedback. Some disagreed with using self-assessment and did not ask their YLs to do so after each task. Many found paired oral tasks inappropriate because they could not listen to each pair and did not trust their pupils to score their own answers. Comments on 'unfamiliar' words 'not learned' (p. 260) were frequently found in teachers' feedback, indicating that they assumed children knew only what they had taught them, although some teachers were pleasantly surprised by how much children were able to figure out through context. Some, however, thought that if YLs could guess meaning it was not a true measure of their knowledge. Teachers were very critical about ambiguity in pictures, as they resulted in multiple correct answers.

In a follow-up single case study, Hild (2017) observed and interviewed a Hungarian EFL teacher with over 35 years of teaching experience whose students took 20 tasks, including oral tasks administered in pairs. The teacher noted that children were helping one another and scaffolded their partner's performance, but she found this unacceptable and asked them to move on to the next question and not to help their partner. The teacher strongly disagreed with self-assessment and explicitly asked YLs not to 'overcomplicate' (p. 708) what they meant to say, thus limiting their motivation and performance. She resisted all innovative ideas and explained why they would not work.

Similar traditional beliefs were reported by Tsagari (2016) who interviewed eight teachers of YLs in Greece and Cyprus about their classroom assessment practices. Although curricula did not require testing children, teachers used paper-and-pencil tests regularly. In both contexts, they tested YLs' vocabulary, grammar, and writing most frequently, whereas listening and speaking were hardly assessed at all. Sentence completion was the most popular test of vocabulary and grammar, and they tested classroom content. When they handed out the marked tests to YLs, they pointed out their mistakes. In their view, children looked forward to writing tests. Although some teachers were familiar with alternative assessment, they could not explain how they used it.

By contrast, other studies have found more positive teacher perceptions. Bacsa and Csíkos (2016) reported that teachers were more open to new ideas when they used listening comprehension tests in a follow-up project aimed at their students' development. After learning about diagnostic feedback, some teachers were motivated to design new diagnostic tests and one was pleasantly surprised that even some low proficiency YLs were able to complete the tests. Similarly, Brumen, Cagran, and

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Rixon (2009) surveyed 108 teachers of English and German in Croatia, the Czech Republic, and Slovenia to explore why and how they assess their YLs between the ages of 5 and 12. In all three countries, assessment was very much part of teachers' daily lives. YLs were regularly assessed to inform parents, children, and teachers themselves, most often by means of grades. Teachers claimed that they applied oral interviews, tests individually developed and borrowed from textbooks, and self-assessment most frequently. However, what exactly they subsumed under 'self-assessments' remained largely unclear.

4.2 Using alternative assessments with YLs

Alternative testing techniques such as self- and peer-assessments are expected to enhance learner autonomy and learning opportunities even at the early stages of language learning. However, few studies thus far have examined how self- and peer assessment works with YLs. With regard to using self-assessment, Butler (2016) discussed two approaches: (a) to target YLs' FL abilities in terms of assessment of learning, and (b) their learning potential, implementing assessment for learning. The challenges concern children's developing ability to reflect on their own learning and the difficulties they face in the learning process. Two empirical studies illustrate her points. In the first study, South Korean fourth and sixth graders ($n = 151$) were invited to self-assess their speaking ability in two modes: in general terms (off-task mode) and in a specific on-task mode (Butler & Lee, 2006). Children's predictions were closer to their teachers' assessments and the actual test scores in the second format, indicating, in our view, how children think in terms of concrete events. Also, older learners were better than younger students at estimating their speaking performance showing that self-assessment may become more precise over time. Additionally, Butler and Lee (2010) involved 254 South Korean sixth graders in an intervention study to examine the impact of YLs' ability to assess themselves on their attitudes, self-confidence, and learning of English, and how their teachers perceived the impact. The findings were mixed: improvement was minimal, probably due to the short time frame, and many contextual challenges such as the teachers' perceptions of the assessment emerged.

Peer assessment was used with fourth, fifth, and sixth graders ($n = 130$) in Taiwan. Hung (2018) compared YLs' assessments of their peers' English-speaking ability with their teacher's evaluations. Children's scores in the fifth and sixth grade were closer to the teacher's assessment than in the fourth grade, indicating development in the ability to assess one another. Although some children were dominant and some hurt their peers' feelings by offering harsh criticism, their teacher managed to follow up on the issues and teach them how to be constructive. Similar issues emerged in a study comparing the relationships between peer-, self-, and teacher assessments of 69 sixth graders in the same context (Hung, Samuelson, & Chen, 2016). Students self-assessed their English oral presentations and their peers and teacher also assessed them. Strong correlations were found between peers' and the teacher's scores and moderate relations between self- and teacher assessment. Most children were motivated by group discussions, were able to improve their presentations, but some were still concerned that their peers were not fair.

Twenty-four Chinese sixth graders performed two oral tasks in two modes: they interacted with their partner (of similarly high or low proficiency) or with their teachers (Butler & Zeng, 2011). After completing both tasks, children were asked to assess themselves, while their teachers also assessed them. YLs and their teachers tended to assess the oral performances mostly similarly and pair-work was found to be helpful for higher achiever pairs, as they produced more complex language. By contrast, lower proficiency learners benefited more from their teacher's support, as they 'stretched' their abilities more successfully than their higher proficiency peers.

Overall, alternative assessment for YLs is an area in need of further research. Although a few publications study their in-depth uses, survey data (e.g., Brumen et al., 2009) indicate that they may be more widely applied than researched. For instance, while portfolio assessment has been widely promoted in the literature and it may be part of classroom practice in some countries (e.g., Council of

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Europe, ELP https://www.coe.int/en/web/portfolio/elp-related-publications; Ioannou-Georgiou & Pavlou, 2003), no empirical study has been found on how teachers use it with their YL classroom. Peer- and self-assessment may not be in line with the assessment culture and classroom practices in many local contexts; therefore, teachers applying innovative techniques need to take into consideration contextual constrains when deploying alternative assessments in line with what is acceptable and desirable in their YL classrooms. In general, though, innovation should be seen in the larger context as part of the assessment culture (Davison, 2013).

4.3 Young learners' test-taking strategies

Test-taking strategies, strategies learners apply to answer test questions successfully, are widely assumed to be helpful, but little is known about YLs' behavior. Studies on how YLs use test-taking strategies may offer important validity evidence, providing insights into how YLs approach and interact with test items while revealing potential construct-irrelevant variance. Two studies focused on YLs' test-taking strategies: Nikolov (2006) collected data by using think aloud protocols from 52 YLs (age 12–13) as they were solving five reading comprehension and two writing tests at A1 level. Comparing high and low achievers revealed how more proficient test takers focused on what they knew not only in English but also about the world, whereas lower achievers were concerned with unfamiliar words. The analysis of the dataset offered important insights into patterns of strategy use: children combined cognitive and metacognitive strategies in unexpected ways, and many relied on translation for meaning making. Some children kept reflecting on their own experiences when working on a dialog, indicating self-centered reasoning.

Similarly, Gu and So (2017) examined what strategies children reported using on the TOEFL Primary tests. Sixteen Chinese test takers (age 6–11) were interviewed after taking four listening and three reading tests to find out why they chose and rejected certain options. Their strategies were categorized as construct-relevant learner strategies and test management strategies, and construct-irrelevant test-wiseness strategies. In line with the previous study, findings showed differences relative to ability level (low, medium, high scorers) and how often YLs applied different strategies—findings that provide important insights into how YLs approach listening and reading tasks.

4.4 Age-appropriate tasks, gamification, and technology-mediated assessment

What tasks are appropriate for developing and measuring YLs' abilities to use their FL are key validity issues. Tasks should be intrinsically motivating, cognitively challenging, as well as doable for YLs (Nikolov, 1999, 2016b). Not all widely used age-appropriate tasks and activities in YLs' classrooms can serve as valid measures of their progress and level of proficiency; however, all task types used in tests should be conducive to learning. Over the decades, many task types have been utilized in research with YLs, and authors have emphasized the complex relationship between cognitive validity and task types (e.g., Papp et al., 2018, pp. 128–269). Tasks are expected to be aligned with what YLs can do in their L1 (across oral/aural AND literacy skills) involving their background knowledge of the world (i.e., social AND academic uses of the FL) and cognitive abilities (i.e., working memory, inductive and deductive reasoning, metacognition). Much has been published about tasks that work with YLs (e.g., McKay, 2005, 2006; Nikolov, 2016b; Papp et al., 2018; Pinter, 2006/2017, 2011); however, not enough is known about the ways in which teachers use diagnostic information gained from deploying them as assessments. Moreover, in relation to tasks, a particular area of interest in YL assessment research has been task difficulty. It is important to explore the relationship between how challenging YLs find tasks and how difficult they are. Cho and So (2014) involved twelve South Korean EFL learners (age 9–12) with a wide range of intensive exposure at school and in private classes to find out what factors influenced perceived task difficulty of eight listening and four reading comprehension multiple-choice tests. After taking the tests, children were asked how clear the instructions were, which tests they found easy or difficult, and how they figured out the answers. Children identified

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

some construct-irrelevant factors causing difficulties, including the complexity of language in questions and answer options, the amount of information they had to remember in listening tasks, ambiguity in visuals, and simultaneous reading and listening, which was expected to be helpful.

In a large-scale diagnostic assessment project (Nikolov, 2017; for details, see Nikolov & Szabó, 2011, 2012; Szabó & Nikolov, 2013) 2,173 young EFL learners (age 6–13) and their 61 teachers piloted 300 diagnostic English tasks for the 4 language skills at 3 estimated levels of difficulty (A1-lower and middle range of A2). Children were invited to reflect on how difficult, familiar, and attractive the tasks were, and their teachers also gave feedback on each task. Moderate relationships were found between the ratings YLs gave on-task difficulty and their achieved scores, indicating YLs ability to use self-assessment. Similar relationships were found between task familiarity and achievements, whereas correlations were somewhat stronger between the extent to which they liked the tasks and how well they performed on them showing how task motivation impacts YLs' perceptions.

Additionally, task difficulty has been investigated in contexts of technology-mediated assessments. Uses of technology and gamification are recent foci in teaching and assessment (Bailey, 2017; Papp et al., 2018); how YLs' digital literacy skills interact with their FL abilities and beliefs is yet another recent avenue for explorations to examine the ways in which new genres and uses of technology (e.g., blogs, emails, text messages; oral presentations, computer- and app-based listening and speaking tasks) work together. For example, Kormos, Brunfaut, and Michel (2020) assessed 104 Hungarian learners of English using computer-administrated Listen-Speak and Listen-Write integrated tasks of the previously available TOEFL Junior™ Comprehensive test. Although YLs found the tasks motivating and performed well on them, they perceived the Listen-Speak tasks more difficult and more anxiety inducing due to time pressure. Studies using technology inform classroom pedagogy by highlighting what may make tasks more challenging.

Game-based assessment offers new insights into how intrinsically motivating gaming elements may serve YLs' needs. Courtney and Graham (2019) implemented an experimental study on YLs' perceptions of a digital game-based assessment in multiple languages. They involved 3,437 FL learners (mean age: 9.3) of English, Spanish, German, Italian, and French in four countries (England, Germany, Italy, Spain) in using digital game-like tests at two levels of difficulty. After taking the assessments, participants evaluated the tasks. The authors collected valuable data on children's reflections as to how motivating and challenging they found the tests. They tended to like the game regardless of their attainment, although they were aware of it being a low-stakes test.

5. Concluding remarks and future research

In this article, we reviewed the main trends and findings on the assessment of YLs' FL in studies published since 2000. Over the decades, YL FL assessment has become a rich bona fide research field in its own right, featuring research on a wide age range of YLs (3–14) in a variety of FL and content-based programs. The way in which the construct has been operationalized has become more varied in the domains of knowledge, skills, and abilities. This variation has resulted from focusing on YLs' communicative abilities, thus widening the initial narrow focus on form. While an increasingly growing body of research now exists, it has become obvious that, with few exceptions, most of the studies we identified were conducted in Western contexts (i.e., predominantly Europe and the United States). For example, we did not find any CLIL studies conducted in other geographical contexts that met the inclusion criteria. Overall, there are several important areas in need of further research that have emerged in this review. These areas would be best examined with YLs across multiple geographical contexts in order to increase and solidify the scope and credibility of the overall field of YL assessment.

The first major area concerns the operationalization of the construct. Approaches to defining constructs and designing frameworks either attempted to align constructs for older learners with YLs' cognitive, emotional, and social characteristics (e.g., Can do statements to CEFR levels) or focused bottom up on children's characteristics to define age-appropriate goals for FL education. Key characteristics of most frameworks include (a) the priority of listening, speaking, and interaction, (b) an

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

acknowledgement of younger learners' slower rate of FL learning, and (c) the realization that typically L2 learning routes are non-linear. Additionally, research highlights how YLs' oral/aural skills and literacy in their L1 or in multiple languages are still developing. However, assessment projects, unfortunately, often neglect some of these features.

Despite the explicit emphasis on listening comprehension, speaking, and interaction in age-appropriate teaching methodology and achievement targets, only a few studies thus far have assessed YLs' listening and speaking skills; instead research has focused on gauging L2 literacy skills, as they are easier to tap into. Insofar as they are missing from many large-scale national assessments (e.g., European Commission/EACEA/Eurydice, 2017), the most noticeable gap concerns attention to YLs' oral/aural L2 abilities. Additionally, when only a single aspect of YLs' FL is tested, it tends to be vocabulary. When asked what it means to learn an FL, children tend to refer to learning new words (Mihaljević Djigunović & Lopriore, 2011), sharing researchers' views, as they assess YLs' breadth of vocabulary as a key aspect of early FL learning. The challenge researchers face concerns whether to assess what children are taught by designing tests in line with the curricula or to apply external instruments for assessing YLs' receptive and productive vocabulary. Several studies used existing tests for L1 learners (e.g., PPVT-4,) instead of developing tests for YLs of the FL in line with the aims of the respective education programs. Researchers may want to focus on the development and validation of vocabulary tests based on curricula and consider assessing children's vocabulary through listening, speaking, and interactive tasks beyond the single word level to better reflect the larger construct of communicative ability.

Along similar lines, projects have tended to assess social uses of the FL, although content-based programs include academic language necessary in school subjects learned in an FL. Despite changes in the operationalization of the construct, studies on CLIL for YLs failed to tap into what YLs in CLIL programs can and cannot do in the FL and the content subject. Further research should examine in what domains content-based instruction is conducive to YLs' proficiency in the L2 and to content learning. If teaching is integrated, assessment should also integrate, or at least include, both domains. Classroom observations, analyses of teaching materials, and involvement of FL instructors, content teachers, and YLs are the next logical steps in developing, piloting, and validating assessments for CLIL programs to explore why results have been discouraging.

A recurring theme concerning results is that evidence underpinning two claims is missing: 'the earlier the better' and 'content-based programs are better than FL programs'. In our view, how constructs and frameworks are operationalized needs to be revisited to determine how assessment instruments relate to YLs' learning experiences, yet another argument for focusing on classroom-based learning and teaching processes to inform assessments for YLs. Hence, further research is needed to outline assessment constructs and domains in line with the goals of FL education programs, children's characteristics, and other contextual variables. Researchers and test developers should define their construct in terms of age-appropriate and domain-specific targets to make sure that what they assess is relevant to how children use language. Emerging construct-irrelevant features should also be analyzed and borne in mind when interpreting results and designing new assessments—to ultimately achieve a holistic, ecological approach (Larsen-Freeman, 2018) to defining and operationalizing the construct in early FL programs and acknowledge, explore, and explain YLs' differential success.

The second area concerns further investigation into the construct of YLs' FL learning and development. For instance, researchers may want to explore what proficiency levels are realistic at certain developmental stages for YLs who speak a certain L1 or multiple L1s. Studies are needed to show if, for example, B1 is a realistic CEFR level for 8-year-olds, or B2 for 12-year-olds; how YLs' performances compare to one another on the same task at various levels, and how children's cognitive, emotional and social skills, their L1(s) and their world knowledge contribute to their performances in the FL. It is also unclear how long-term memory and attrition work with YLs. It would be interesting to examine how YLs perform on the same tests after a few months or years and to investigate reasons for score gains or losses. Additionally, research should focus on YLs as individuals in their trajectories in

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

interaction with their peers and teachers in their specific contexts. Classroom-based case studies are needed on individuals and small groups together with their teachers to find out not only what children can and cannot do as they progress, learn, and forget, but also why, and to reveal how both YLs' and teachers' learning can be scaffolded by learning-oriented assessment. A focus on what YLs can do with support today is as important as what they can do without support tomorrow, and how learning-oriented assessment can scaffold and motivate YLs' slow development.

Third, how assessment is carried out in the classroom is an area in need of additional research. For instance, information regarding how teachers conduct (formative) assessment of YLs' FL abilities in the classroom is largely missing from the FL literature (e.g., Rea-Dickins & Gardner, 2000 for an ESL context). Although there is increased interest in teachers' language assessment literacy (e.g., Lan & Fan, 2019), little is known about teachers' daily assessment practices, how their feedback to YLs and grading of their performances motivate students to put more effort into doing similar or more difficult tasks or to shy away from further practice. Systematic observations of classroom practice coupled with interviews could provide valuable insights into daily assessment practices as well as potentially reveal areas in which teachers would benefit from additional support.

Overall, classroom observation as a useful technique for assessing YLs, also used in teacher education, is underutilized. Observation would be particularly appropriate for collecting evidence in preschools and content-based programs. Additionally, it can provide insights into what tasks teachers use, what they want children to be able to do, and how they determine to what extent they can do them. Understanding conflicting results of assessments is not possible without exploring the teaching and learning processes. In our view, more studies are needed of teachers' classroom assessment practices and their impact on YLs' learning, motivation, anxiety, willingness to communicate, etc. Unfortunately, no study was conducted by practicing teachers, although they are key players in YLs' lives. Hence future studies that involve YLs should draw more on observation as an assessment and data collection method.

Finally, more consideration and research is needed into test impact. Although socially responsible professionals should be aware of potential unintended consequences, few publications discuss ethical issues related to the impact of YL assessment. Most publications fail to share information on what happens to test results, or how stakeholders utilize them in children's interest. For example, for national examinations it would be important to document what decisions are made based on assessment results at the program levels and how they impact FL teaching and learning. Also, with regard to smaller-scale research projects such as assessments for learning, it would be valuable to examine how teachers use the diagnostic information and how the assessments impact what teachers and YLs do in the FL classroom. Accordingly, future studies should inquire into how test results are utilized, how they inform teaching, and how they impact children's and teachers' lives.

Furthermore, hardly any research is published on how the most vulnerable, less able, and anxious learners are impacted by testing techniques and results. For instance, very little is known about children with learning difficulties (Kormos, 2017) and from disadvantaged family backgrounds. Also, how diagnostic assessment is applied with children coping with SLDs, anxiety, or low self-esteem should be a priority. Moreover, in some cases, ethical issues concern why assessment, which may induce anxiety and take precious time away from learning activities, is necessary at all. For example, in experimental projects involving pre-school and lower-primary learners some children were sleepy or unwilling to participate in assessments far beyond their attention span and abilities. It is not a good idea to give children tasks they most likely cannot do successfully. We would argue against assessing pre-school children, based on the controversial findings of the empirical studies reviewed in this paper. In fact, we wonder why YLs, especially in lower-primary years, were assessed in many of the studies. Also, it is difficult to understand and justify how high-stakes examinations administered by strangers are conducive—especially toYLs' FL learning—in the long run. By contrast, classroom-based assessment projects, however, aiming to diagnose what YLs can and cannot do and why, are much needed because they can be highly informative for teachers, children, and parents. Hence, the field of YL assessment offers many possibilities for further developments—developments that are much needed in view of the

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

seemingly unstoppable fluctuations in the ebbs and flows of policy-related enthusiasm for early language learning.

Questions arising

1. What are the most age-appropriate data collection instruments and tasks for collecting evidence about YLs' aural, oral, and interaction abilities in an FL? How do they work with YLs of different ages in different cultural contexts?
2. What is the relationship between YLs' performances on tasks targeting only one skill vs. integrated tasks that target multiple skills?
3. What task types can develop YLs aptitude and academic L2? How do they work with YLs of different ages?
4. What do typical performances on the same tasks look like at different levels (e.g., A1-B1-B2) at different ages in different contexts?
5. What are the most age-appropriate data collection instruments and tasks for collecting evidence about YLs' knowledge, skills, and abilities in their FL AND content learned in the FL? How can integrated tasks tap into knowledge, skills, and abilities in both the FL and the content subject? How does assessment of content in L1 or FL impact results?
6. How do test results in national and international examinations impact YLs, teachers, parents, and decision-makers? How are test results used?
7. How do pre- and in-service teacher education programs prepare teachers for assessing young FL learners?
8. How are assessment practices (formative and alternative assessment, grading etc.) in FL classrooms and other school subjects related? How does the local assessment culture impact assessment in FL classrooms?
9. What types of assessment FOR learning do teachers use in the FL classroom? How do teachers and YLs benefit from diagnostic assessment practices? How does diagnostic feedback impact FL learning, as well as teachers' and learners' motivation, anxiety, and autonomy? How do test results support decision-making and FL learning in the classroom?
10. How can technology and gamification be applied in assessment of and for learning?

References

Agustín Llach, M. P. (2015). The effects of the CLIL approach in young foreign language learners' lexical profiles. *International Journal of Bilingual Education and Bilingualism*, *20*(5), 557–573.

Agustín Llach, M. P. (2016). Vocabulary growth in young CLIL and traditional EFL learners: Evidence from research and implications for education. *International Journal of Applied Linguistics*, *26*(2), 211–227.

Albaladejo Albaladejo, S., Coyle, Y., & de Larios, J. R. (2018). Songs, stories and vocabulary acquisition in preschool learners of English as a foreign language. *System*, *76*, 116–128.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London, UK: Continuum.

Alexiou, T. (2009). Young learners' cognitive skills and their role in foreign language vocabulary learning. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 46–61).Bristol, UK: Multilingual Matters.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Anthony, L., & Nation, I. S. P. (2017). *Picture Vocabulary Size Test (Version 1.2.0) [Computer software and measurement instrument]*. Tokyo, Japan: Waseda University.

Bacsa, É, & Csíkos, C. (2016). The role of individual differences in the development of listening comprehension in the early stages of language learning. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 263–289). Heidelberg, Germany: Springer.

Bailey, A. L. (2005). Cambridge young learners English (YLE) tests. *Language Testing*, *22*(2), 242–252.

Bailey, A. L. (2017). Theoretical and developmental issues to consider in the assessment of young learners' English language proficiency. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 25–40). New York, NY: Routledge.

Baron, P. A., & 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.

Benigno, V., & de Jong, J. (2016). A CEFR-based inventory of YL descriptors: Principles and challenges. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 43–64). Heidelberg, Germany: Springer.

Bishop, D. (2003). *Test for Reception of Grammar – Version 2 (TROG-2)*. London, UK: Pearson Assessment.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7–73.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Brumen, M., Cagran, B., & Rixon, S. (2009). Comparative assessment of young learners' foreign language competence in three Eastern European countries. *Educational Studies*, *35*(3), 269–295.
Butler, Y. G. (2009). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly*, *43*(3), 417–444.
Butler, Y. G. (2016). Self-assessment *of* and *for* young learners' foreign language learning. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 291–315). Heidelberg, Germany: Springer.
Butler, Y. G., & Le, V.-N. (2018). A longitudinal investigation of parental social-economic status (SES) and young students' learning of English as a foreign language. *System*, *73*, 4–15.
Butler, Y. G., & Lee, J. (2006). On-task vs. off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal*, *90*(4), 506–518.
Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31.
Butler, Y. G., Sayer, P., & Huang, B. (2018). Introduction: Social class/socioeconomic status and young learners of English as a global language. *System*, *73*, 1–3.
Butler, Y. G., & Zeng, W. (2011). The roles that teachers play in paired-assessments for young learners. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment* (pp. 77–92). Frankfurt am Main, Germany: Peter Lang.
Cameron, L. (2003). Challenges for ELT from the expansion in teaching children. *ELT Journal*, *57*(2), 105–112.
Carless, D., & Lam, R. (2014). The examined life: Perspectives of lower primary school students in Hong Kong. *Education*, *42* (3), 313–329.
Cenoz, J. (2003). The effect of age on foreign language acquisition in formal contexts. In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 77–93). Bristol, UK: Multilingual Matters.
Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical Analysis of CLIL: Taking Stock and Looking Forward. *Applied Linguistics*, *35*(3), 243–262.
Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London, UK: Routledge.
Chik, A., & Besser, S. (2011). International language test taking among young learners: A Hong Kong case study. *Language Assessment Quarterly*, *8*(1), 73–91.
Cho, Y., & Blood, I. (in progress). An analysis of TOEFL® Primary™ Repeaters: How much score change occurs? *Language Testing*.
Cho, Y., Ginsburgh, M., Morgan, R., Moulder, B., Xi, X., & Hauck, M. C. (2016). *Designing the TOEFL® Primary™ Tests*. (Research Memorandum No. RM-16-02). Princeton, NJ: Educational Testing Service.
Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young EFL learners' perceptions of test task difficulty* (TOEFL Research Memorandum ETS RM 14-04). Princeton, NJ: Educational Testing Service.
Courtney, L., & Graham, S. (2019). 'It's like having a test but in a fun way': Young learners' perceptions of a digital game-based assessment of early language learning. *Language Teaching for Young Learners*, *1*(2), 161–186.
Coyle, Y., & Gómez Gracia, G. (2014). Using songs to enhance L2 vocabulary acquisition in preschool children. *ELT Journal*, *68*(3), 276–285.
Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, *19*, 203–218.
Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Bristol, UK: Multilingual Matters.
Curtain, H. (2009). Assessment of early learning of foreign languages in the USA. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 60–82). Berlin, Germany/New York, NY: Mouton de Gruyter.
Davis, G. M., & Fan, W. (2016). English vocabulary acquisition through songs in Chinese kindergarten students. *Chinese Journal of Applied Linguistics*, *39*(1), 59–71.
Davison, C. (2013). Innovation in assessment: Common misconceptions. In K. Hyland & L. L. C. Wong (Eds.), *Innovation and change in English language education* (pp. 263–267). New York, NY: Routledge.
de Diezmas, E. N. M. (2016). The impact of CLIL on the acquisition of L2 competences and skills in primary education. *International Journal of English Studies*, *16*(2), 81–101.
de Wilde, V., & Eyckmans, J. (2017). Game on! Young learners' incidental language learning prior to instruction. *Studies in Second Language Learning and Teaching*, *7*(4), 673–694.
Edelenbos, P., Johnstone, R. & Kubanek, A. (2006). The main pedagogical principles underlying the teaching of languages to very young learners. Languages for the children of Europe: Published research, good practice and main principles. European Commission Report. Retrieved from http://ec.europa.eu/education/policies/lang/doc/youngsum_en.pdf.
Edelenbos, P., & Kubanek-German, A. (2004). Teacher assessment: The concept of 'diagnostic competence'. *Language Testing*, *21*(3), 259–283.
Edelenbos, P., & Vinjé, M. P. (2000). The assessment of a foreign language at the end of primary (elementary) education. *Language Testing*, *17*(2), 144–162.
European Commission/EACEA/Eurydice. (2017). *Key data on teaching languages at school in Europe – 2017 ed*. Eurydice Report. Luxembourg: Publications Office of the European Union.
Fenyvesi, K., Hansen, M., & Cadierno, T. (2018). The role of individual differences in younger vs. older primary school Danish learners of English. *International Review of Applied Linguistics in Language Teaching*. doi:10.1515/iral-2017-0053

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Fernandez-Sanjurjo, J., Arias Blanco, J. M., & Fernandez-Costales, A. (2018). Assessing the influence of socio-economic status on students' performance in Content and Language Integrated Learning. *System*, *73*(1), 16–26.
Fernández-Sanjurjo, J., Fernández-Costales, A., & Arias Blanco, J. M. (2017). Analysing students' content-learning in science in CLIL vs. non-CLIL programmes: Empirical evidence from Spain. *International Journal of Bilingual Education and Bilingualism*, *22*(6), 661–674.
Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113–132.
García Lecumberri, M. L., & Gallardo del Puerto, F. (2003). English FL sounds in school learners of different ages. In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 115–135). Bristol, UK: Multilingual Matters.
García Mayo, M. D. P., & García Lecumberri, M. L. (2003). *Age and the acquisition of English as a foreign language*. Bristol, UK: Multilingual Matters.
García Mayo, M. P. (2003). Age, length of exposure and grammaticality judgements in the acquisition of English as a foreign language. In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 94–114). Bristol, UK: Multilingual Matters.
Gattullo, F. (2000). Formative assessment in ELT primary (elementary) classrooms: An Italian case study. *Language Testing*, *17*(2), 278–288.
Getman, E., Cho, Y., & Luce C. (2016). *Effects of printed option sets on listening item performance among young English-as-a-foreign-language learners*. (Research Memorandum No. RM-16-16). Princeton, NJ: Educational Testing Service.
Griva, E., & Sivropoulou, R. (2009). Implementation and evaluation of an early foreign language learning project in kindergarten. *Early Childhood Education Journal*, *37*(1), 79–87.
Gu, L., & So, Y. (2017). Strategies used by young English learners in an assessment context. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 119–135). London, UK/New York, NY: Routledge.
Haenni Hoti, A., Heinzmann, S., & Müller, M. (2009). 'I can you help?' Assessing speaking skills and interaction strategies of young learners. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 119–140). Berlin, Germany: Mouton de Gruyter.
Haskins, R. (2018). Evidence-based policy: The movement, the goals, the issues, the promise. *The ANNALS of the American Academy of Political and Social Science*, *678*(1), 8–37.
Hasselgreen, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, *17*(2), 261–277.
Hasselgreen, A. (2003). *Bergen 'Can Do' project*. Strasbourg, France: Council of Europe. Retrieved from http://www.ecml.at/documents/pub221E2003_Hasselgreen.pdf
Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing*, *22*(3), 337–354.
Hasselgreen, A., & Caudwell, G. (2016). *Assessing the language of young learners*. Sheffield/Bristol, UK: Equinox.
Hild, G. (2017). A case study of a Hungarian EFL teacher's assessment practices with her young learners. *Studies in Second Language Teaching and Learning*, *7*(4), 695–714.
Hsieh, C. (2016). Examining content representativeness of a young learner language assessment: EFL teachers' perspectives. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 93–107). Heidelberg, Germany: Springer.
Hsieh, C., Ionescu, M., & Ho, T. (2017). Out of many, one: Challenges in teaching multilingual Kenyan primary students in English. *Language, Culture and Curriculum*, *31*(2), 199–213.
Hung, Y. (2018). Group peer assessment of oral English performance in a Taiwanese elementary school. *Studies in Educational Evaluation*, *59*, 19–28.
Hung, Y., Samuelson, B., & Chen, C. (2016). Relationship between peer- and self-assessment and teacher assessment of young EFL learners' oral presentations. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 317–338). Heidelberg, Germany: Springer.
Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 83–96). Berlin, Germany: Mouton de Gruyter.
Ioannou-Georgiou, S., & Pavlou, P. (2003). *Assessing young learners*. Oxford, UK: Oxford University Press.
Jaekel, N., Schurig, M., Florian, M., & Ritter, M. (2017). From early starters to late finishers? A longitudinal study of early foreign language learning in school. *Language Learning*, *67*(3), 631–664.
Johnstone, R. (2000). Context-sensitive assessment of modern languages in primary (elementary) and early secondary education: Scotland and the European experience. *Language Testing*, *17*(2), 123–143.
Johnstone, R. (2003). Evidence-based policy: Early modern language learning at primary. *The Language Learning Journal*, *28* (1), 14–21.
Johnstone, R. (2009). An early start: What are the key conditions for generalized success? In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: international perspectives* (pp. 31–42). Reading, UK: Garnet Education Publishing.
Johnstone, R. (2010). Introduction. In R. Johnstone (Ed.), *Learning through English: Policies, challenges and prospects* (pp. 7–23). London, UK: British Council.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Kane, M. (2011). Validating score interpretations and uses: Messick Lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing*, *29*(1), 3–17.
Kiss, C. (2009). The role of aptitude in young learners' foreign language learning. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 253–256). Berlin, Germany/New York, NY: Mouton de Gruyter.
Kiss, C., & Nikolov, M. (2005). Developing, piloting, and validating an instrument to measure young learners' aptitude. *Language Learning*, *55*(1), 99–150.
Klein-Braley, C. (1997). C-tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, *14*(1), 47–84.
Kondo-Brown, K. (2004). Investigating interviewer-candidate interactions during oral interviews for child L2 learners. *Foreign Language Annals*, *37*(4), 602–613.
Kormos, J. (2017). The effects of specific learning difficulties on processes of multilingual language development. *Annual Review of Applied Linguistics*, *37*, 30–44.
Kormos, J., Brunfaut, T., & Michel, M. (2020). Motivational factors in computer-administered integrated skills tasks: A study of young learners. *Language Assessment Quarterly*, *17*(1), 43–59.
Lan, C., & Fan, S. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. *Studies in Educational Evaluation*, *61*, 112–122.
Larsen-Freeman, D. (2018). Looking ahead: Future directions in, and future research into, second language acquisition. *Foreign Language Annals*, *51*, 55–72.
Lasagabaster, D., & Doiz, A. (2003). Maturational constraints on foreign-language written production. In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 136–159). Bristol, UK: Multilingual Matters.
Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, *16*(1), 33–51.
Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing*, *35*(2), 239–269.
Marshall, H., & Gutteridge, M. (2002). *Candidate performance in the young learner English tests in* 2000. In *Research Notes 7*. Cambridge, UK: University of Cambridge Local Examinations Syndicate.
McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics*, *25*, 243–263.
McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.
Merikivi, R., & Pietilä, P. (2014). Vocabulary in CLIL and in mainstream education. *Journal of Language Teaching and Research*, *5*(3), 487–497.
Mihaljević Djigunović, J. (2006). Role of affective factors in the development of productive skills. In M. Nikolov & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 9–24). Pécs, Hungary: Lingua Franca Csoport.
Mihaljević Djigunović, J. (2016). Individual differences and young learners' performance on L2 speaking tests. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 243–261). Heidelberg, Germany: Springer.
Mihaljević Djigunović, J. (2019). Affect and assessment in teaching L2 to young learners. In D. Prošić-Santovac & S. Rixon (Eds.), *Integrating assessment into early language learning and teaching practice* (pp. 19–33). Bristol, UK: Multilingual Matters.
Mihaljević Djigunović, J., & Lopriore, L. (2011). The learner: Do individual differences matter? In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 29–45). London, UK: The British Council.
Mihaljević Djigunović, J., & Nikolov, M. (2019). Motivation of young language learners. In M. Lamb, K. Csizér, A. Henry, & S. Ryan (Eds.), *Palgrave Macmillan handbook of motivation for language learning* (pp. 515–534). Basingstoke, UK: Palgrave Macmillan.
Mihaljević Djigunović, J., Nikolov, M., & Ottó, I. (2008). A comparative study of Croatian and Hungarian EFL students. *Language Teaching Research*, *12*(3), 433–452.
Mihaljević Djigunović, J., & Vilke, M. (2000). Eight years after: Wishful thinking vs facts of life. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 66–86). Pécs, Hungary: University Press Pécs.
Muñoz, C. (2003). Variation in oral skill development and age of onset. In M. P. García Mayo & M. L. García Lecumberri (Eds.), *Age and the acquisition of English as a foreign language* (pp. 161–181).Bristol, UK: Multilingual Matters.
Muñoz, C. (2006). *Age and the rate of foreign language learning*. Bristol, UK: Multilingual Matters.
Nikolov, M. (1999). 'Why do you learn English?' 'Because the teacher is short.' A study of Hungarian children's foreign language learning motivation. *Language Teaching Research*, *3*(1), 33–56.
Nikolov, M. (2006). Test-taking strategies of 12- and 13-year-old Hungarian Learners of EFL: Why whales have migraines. *Language Learning*, *56*(1), 1–51.
Nikolov, M. (2009). Early modern foreign language programmes and outcomes: Factors contributing to Hungarian learners' proficiency. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 90–107). Bristol, UK: Multilingual Matters.
Nikolov, M. (2016a). Trends, issues, and challenges in assessing young language learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 1–18). Heidelberg, Germany: Springer.
Nikolov, M. (2016b). A framework for young EFL learners' diagnostic assessment: 'Can do statements' and task types. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 65–92). New York, NY: Springer.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Nikolov, M. (2017). Students' and teachers' feedback on diagnostic tests for young EFL learners: Implications for classrooms. In M. P. García Mayo (Ed.), *Learning foreign languages in primary school: Research insights* (pp. 249–266). Bristol, UK: Multilingual Matters.

Nikolov, M., & Csapó, B. (2018). The relationships between 8th graders' L1 and L2 readings skills, inductive reasoning and socio-economic status in early English and German as a foreign language programs. *System*, *73*, 48–57.

Nikolov, M., & Curtain, H. A. (2000). *An early start: Young learners and modern languages in Europe and beyond*. Strasbourg, France: Council of Europe Pub.

Nikolov, M., & Mihaljević Djigunović, J. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review of Applied Linguistics*, *26*, 234–260.

Nikolov, M., & Mihaljević Djigunović, J. (2011). All shades of every color: An overview of early teaching and learning of foreign languages. *Annual Review of Applied Linguistics*, *31*, 95–119.

Nikolov, M., & Szabó, G. (2011). Establishing difficulty levels of diagnostic listening comprehension tests for young learners of English. In J. Horváth (Ed.), *UPRT 2011: Empirical studies in English applied linguistics* (pp. 73–82). Pécs, Hungary: Lingua Franca Csoport.

Nikolov, M. & Szabó G. (2012). Developing diagnostic tests for young learners of EFL in grades 1 to 6. In E. D. Galaczi & C. J. Weir (Eds.), Voices in language assessment: Exploring the impact of language frameworks on learning, teaching and assessment – Policies, procedures and challenges, Proceedings of the ALTE Krakow Conference, July 2011. Cambridge, UK: UCLES/Cambridge University Press, 347–363.

Nikolov, M., & Szabó, G. (2015). A study on Hungarian 6th and 8th graders' proficiency in English and German at dual-language schools. In D. Holló & K. Károly (Eds.), *Inspirations in foreign language teaching: Studies in applied linguistics, language pedagogy and language teaching* (pp. 184–206). Harlow, UK: Pearson Education.

Papageorgiou, S., & Bailey, A. (2019). Preface. In S. Papageorgiou & K. M. Bailey (Eds.), *Global perspectives on language assessment: Research, theory, and practice* (pp. xi–xv). New York, NY: Routledge.

Papageorgiou, S., & Baron, P. (2017). Using the Common European Framework of Reference to facilitate score interpretations for young learners' English language proficiency assessments. In M. K. Wolf & Y. G. Butler (Eds.), *English language proficiency assessments for young learners* (pp. 136–152). New York, NY: Routledge.

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly*, *12*(2), 153–157.

Papp, S., Khabbazbashi, N., & Miller S. (2012). *Computer-based speaking YLE test second trial: Movers and Flyers (China, May 2012)*, Report Number: VR 1388, Cambridge, UK: Cambridge ESOL internal report.

Papp, S., Rixon, S., with Field, J. (2018). *Examining young language learners: The Cambridge English approach to assessing children and teenagers in schools*. Cambridge, UK: Cambridge University Press.

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

Papp, S., Street, J., Galaczi, E., Khalifa, H., & French A. (2010). *YLE paired/group speaking modification trial*, Report Number: VR 1285, Cambridge, UK: Cambridge ESOL internal report.

Papp, S., & Walczak, A. (2016). The development and validation of a computer-based test of English for young learners: Cambridge English young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 139–190). Heidelberg, Germany: Springer.

Peng, J., & Zheng, S. (2016). A longitudinal study of a school's assessment project in Chongqing, China. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 213–241). Heidelberg, Germany: Springer.

Pfenninger, S. E., & Singleton, D. (2017). *Beyond age effects in instructional L2 learning: Revisiting the age factor*. Bristol, UK: Multilingual Matters.

Pfenninger, S. E., & Singleton, D. (2018). Starting age overshadowed: The primacy of differential environmental and family support effects on second language attainment in an instructional context. *Language Learning*, *69*(s1), 207–234.

Pinter, A. (2006/2017). *Teaching young language learners*. Oxford, UK: Oxford University Press.

Pinter, A. (2011). *Children learning second languages*. London, UK: Palgrave Macmillan.

Pižorn, K. (2009). Designing proficiency levels for English for primary and secondary school students and the impact of the CEFR. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives* (pp. 87–102). Arnhem, Netherlands: Cito, EALTA.

Pižorn, K., & Moe, E. (2012). A validation study of the national assessment instruments for young English language learners in Norway and Slovenia. *CEPS Journal*, *2*(3), 75–97.

Porsch, R., & Wilden, E. (2017). The development of a curriculum-based C-test for young EFL learners. In J. Enever & E. Lindgren (Eds.), *Early language learning: Complexity and mixed methods* (pp. 289–304). Bristol, UK: Multilingual Matters.

Puimège, E., & Peters, E. (2019). Learners' English vocabulary knowledge prior to formal instruction: The role of learner-related and word-related variables. *Language Learning*, *69*(4), 943–977.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing*, *17*(2), 115–122.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: Disentangling the construct of formative assessment. *Language Testing*, *17*(2), 215–243.

Rea-Dickins, P., & Rixon, S. (1997). The assessment of young learners of English as a foreign language. In C. Clapham & D. Corson (Eds.), *The encyclopaedia of language and education, Vol. 7: Language Testing* (pp. 151–161). Dordrecht, Netherlands: Kluwer.

Rixon, S. (2013). *British Council survey of policy and practice in primary English language teaching worldwide*. London, UK: 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 M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 19–41). Heidelberg, Germany: Springer.

Rixon, S., & Prošić-Santovac, D. (2019). Introduction: Assessment and early language learning. In D. Prošić-Santovac & S. Rixon (Eds.), *Integrating assessment into early language learning and teaching practice* (pp. 1–16). Bristol, UK: Multilingual Matters.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, J. (2015). *TOEFL® Junior Design Framework*. TOEFL Junior Research Report No. 02. ETS Research Report, No. RR-15-13. Princeton, NJ: Educational Testing Service.

Sun, H., Steinkrauss, R., Wieling, M., & de Bot, K. (2018). Individual differences in very young Chinese children's English vocabulary breadth and semantic depth: Internal and external factors. *International Journal of Bilingual Education and Bilingualism*, *21*(4), 405–425.

Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning: From theory and research to practice*. London, UK: Palgrave Macmillan.

Szabó, G., & Nikolov M. (2013). An analysis of young learners' feedback on diagnostic listening comprehension tests. In J. Mihaljević Djigunović & M. Medved Krajnović (Eds.), *UZRT 2012: Empirical studies in English applied linguistics*. Zagreb, Croatia: FF press, 7–21. Retrieved from http://books.google.hu/books?id=VnR3DZsHG6UC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

Szabó, T. (2018a). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Vol. 1: Ages 7–10: Collated representative samples of descriptors of language competences developed for young learners. Retrieved from https://rm.coe.int/collated-representative-samples-descriptors-young-learners-volume-1-ag/16808b1688

Szabó, T. (2018b). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Collated representative samples of descriptors of language competences developed for young learners aged 11–15 years. Retrieved from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680697fc9

Szpotowicz, M., & Lindgren, E. (2011). Language achievements: A longitudinal perspective. In J. Enever (Ed.), *ELLiE: Early Language Learning in Europe* (pp. 125–144). London, UK: British Council.

Timpe-Laughlin, V. (2018). *A good fit? Examining the alignment between the TOEFL Junior® Standard test and the English as a foreign language curriculum in Berlin, Germany*. (Research Memorandum No. RM-18-11). Princeton, NJ: Educational Testing Service.

Tragant, E., Marsol, A., Serrano, R., & Llanes, A. (2016). Vocabulary learning at primary school: A comparison of EFL and CLIL. *International Journal of Bilingual Education and Bilingualism*, *19*(5), 579–591.

Tsagari, D. (2012). FCE exam preparation discourses: Insights from an ethnographic study. *Research Notes*, *47*, 36–48.

Tsagari, D. (2016). Assessment orientations of primary state school EFL teachers in two Mediterranean countries. *In CEPS Journal*, *6*(1), 9–30. Retrieved from http://www.cepsj.si/doku.php?id=en:volumes:2016-vol6-no1.

Unesco. Institute for Statistics. (2012). *International standard classification of education: ISCED* 2011; [pdf] Retrieved from: http://uis.unesco.org/sites/default/files/documents/international-standardclassification-of-education-isced-2011-en.pdf.

Unsworth, S., Persson, L., Prins, T., & de Bot, K. (2014). An investigation of factors affecting early foreign language learning in the Netherlands. *Applied Linguistics*, *36*(5), 527–548.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Basingstoke, UK: Palgrave Macmillan.

Wilden, E., & Porsch, R. (2016). Learning EFL from year 1 or year 3? A comparative study on children's EFL listening and reading comprehension at the end of primary education. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 191–212). Heidelberg, Germany: Springer.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14.

Wolf, M. K., & Butler, Y. G. (2017). *English language proficiency assessments for young learners*. London, UK/New York, NY: Routledge.

Yeung, S. S., Ng, M. L., & King, R. B. (2016). English vocabulary instruction through storybook reading for Chinese EFL kindergarteners: Comparing rich, embedded, and incidental approaches. *Asian EFL Journal*, *18*, 81–104.

Yeung, S. S., Siegel, L. S., & Chan, C. K. (2013). Effects of a phonological awareness program on English reading and spelling among Hong Kong Chinese ESL children. *Reading and Writing*, *26*(5), 681–704.

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press

Zangl, R. (2000). Monitoring language skills in Austrian primary (elementary) schools: A case study. *Language Testing*, *17*(2), 250–260.

Marianne Nikolov is Professor Emerita of English Applied Linguistics at the University of Pécs, Hungary. Early in her career, she taught EFL to YLs for a decade. Her research interests include early learning and teaching of modern languages, assessment of processes and outcomes in language education, individual differences, teacher education, teachers' beliefs and practices, and language policy. Her work has been published in *Annual Review of Applied Linguistics, Language Learning, Language Teaching, Language Teaching Research, System* and by Mouton de Gruyter, Multilingual Matters, Peter Lang, and Springer. Her CV is at her website: http://ies.btk.pte.hu/content/nikolov_marianne

Veronika Timpe-Laughlin is a research scientist in the field of English Language Learning and Assessment at Educational Testing Service (ETS). Her research interests include pragmatics, young learners' language assessment, task-based language teaching, bilingual first language acquisition, and technology in L2 instruction and assessment. Veronika has recently published in *Language Assessment Quarterly* and *Applied Linguistics Review* and is the co-author of the 2017 book *Second language educational experiences for adult learners* (Routledge). Prior to joining ETS, Veronika worked and taught in the English Department at TU Dortmund University, Germany.

Cite this article: Nikolov, M., & Timpe-Laughlin, V. (2021). Assessing young learners' foreign language abilities. *Language Teaching*, *54*(1), 1–37. https://doi.org/10.1017/S0261444820000294

https://doi.org/10.1017/S0261444820000294 Published online by Cambridge University Press