

Educational Linguistics

Marianne Nikolov *Editor*

Assessing Young Learners of English: Global and Local Perspectives

 Springer

Educational Linguistics

Volume 25

Series Editor

Francis M. Hult, Lund University, Sweden

Editorial Board

Marilda C. Cavalcanti, Universidade Estadual de Campinas, Brazil

Jasone Cenoz, University of the Basque Country, Spain

Angela Creese, University of Birmingham, United Kingdom

Ingrid Gogolin, Universität Hamburg, Germany

Christine Hélot, Université de Strasbourg, France

Hilary Janks, University of Witwatersrand, South Africa

Claire Kramersch, University of California, Berkeley, U.S.A.

Constant Leung, King's College London, United Kingdom

Angel Lin, University of Hong Kong, Hong Kong

Alastair Pennycook, University of Technology, Sydney, Australia

Educational Linguistics is dedicated to innovative studies of language use and language learning. The series is based on the idea that there is a need for studies that break barriers. Accordingly, it provides a space for research that crosses traditional disciplinary, theoretical, and/or methodological boundaries in ways that advance knowledge about language (in) education. The series focuses on critical and contextualized work that offers alternatives to current approaches as well as practical, substantive ways forward. Contributions explore the dynamic and multi-layered nature of theory-practice relationships, creative applications of linguistic and symbolic resources, individual and societal considerations, and diverse social spaces related to language learning.

The series publishes in-depth studies of educational innovation in contexts throughout the world: issues of linguistic equity and diversity; educational language policy; revalorization of indigenous languages; socially responsible (additional) language teaching; language assessment; first- and additional language literacy; language teacher education; language development and socialization in non-traditional settings; the integration of language across academic subjects; language and technology; and other relevant topics.

The *Educational Linguistics* series invites authors to contact the general editor with suggestions and/or proposals for new monographs or edited volumes. For more information, please contact the publishing editor: Jolanda Voogd, Associate Publishing Editor, Springer, Van Godewijkstraat 30, 3300 AA Dordrecht, the Netherlands.

More information about this series at <http://www.springer.com/series/5894>

Marianne Nikolov
Editor

Assessing Young Learners of English: Global and Local Perspectives

 Springer

Editor

Marianne Nikolov
Department of English Applied Linguistics
University of Pécs
Pécs, Hungary

ISSN 1572-0292

Educational Linguistics

ISBN 978-3-319-22421-3

DOI 10.1007/978-3-319-22422-0

ISSN 2215-1656 (electronic)

ISBN 978-3-319-22422-0 (eBook)

Library of Congress Control Number: 2015951960

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

Acknowledgment

The editor and the authors of the book are grateful to two anonymous reviewers for their insights and comments on the first drafts. Their recommendations helped us tailor the text to our readers' needs. We would also like to thank Dr. Francis M. Hult, editor of Education Linguistics book series at Springer, for his helpful guidance and Mrs. Jolanda Voogd, senior editor, and Helen van der Stelt, her senior assistant, for their support on this project.

Contents

Trends, Issues, and Challenges in Assessing Young Language Learners	1
Marianne Nikolov	
Do Developments in Assessment Represent the ‘Coming of Age’ of Young Learners English Language Teaching Initiatives? The International Picture	19
Shelagh Rixon	
The “Global Scale of English Learning Objectives for Young Learners”: A CEFR-Based Inventory of Descriptors	43
Veronica Benigno and John de Jong	
A Framework for Young EFL Learners’ Diagnostic Assessment: ‘Can Do Statements’ and Task Types	65
Marianne Nikolov	
Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers’ Perspectives	93
Ching-Ni Hsieh	
Developing and Piloting Proficiency Tests for Polish Young Learners	109
Magdalena Szpotowicz and Dorota E. Campfield	
The Development and Validation of a Computer-Based Test of English for Young Learners: Cambridge English Young Learners	139
Szilvia Papp and Agnieszka Walczak	
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	191
Eva Wilden and Raphaela Porsch	

A Longitudinal Study of a School's Assessment Project in Chongqing, China	213
Jing Peng and Shicheng Zheng	
Individual Learner Differences and Young Learners' Performance on L2 Speaking Tests	243
Jelena Mihaljević Djigunović	
The Role of Individual Differences in the Development of Listening Comprehension in the Early Stages of Language Learning	263
Éva Bacsá and Csaba Csíkó	
Self-Assessment of and for Young Learners' Foreign Language Learning	291
Yuko Goto Butler	
Relationships between Peer- and Self-Assessment and Teacher Assessment of Young EFL Learners' Oral Presentations	317
Yu-ju Hung, Beth Lewis Samuelson, and Shu-cheng Chen	

Contributors

Éva Bacsa Kiss Bálint Reformed School, Szentes, Hungary

Veronica Benigno Pearson English, Pearson Education, London, UK

Yuko Goto Butler Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA

Dorota E. Campfield Educational Research Institute, Warsaw, Poland

Shu-cheng Chen Sianbei Elementary School, Tainan City, Taiwan

Csaba Csíkos Department of Educational Assessment and Planning, University of Szeged, Szeged, Hungary

John de Jong Pearson Assessment Centre, Pearson Education Inc., Iowa City, IA, USA

Amsterdam VU University, Amsterdam, Netherlands

Jelena Mihaljević Djigunović Department of English, University of Zagreb, Zagreb, Croatia

Ching-Ni Hsieh Center for English Language Learning and Assessment, Research and Development, Educational Testing Service, Princeton, NJ, USA

Yu-ju Hung Foreign Languages Division, R.O.C. Air Force Academy, Kaohsiung City, Taiwan

Marianne Nikolov Institute of English Studies, University of Pécs, Pécs, Hungary

Szilvia Papp Research and Validation Group, Cambridge English Language Assessment, Cambridge, UK

Jing Peng Research Centre of Language, Cognition and Language Application, Chongqing University, Chongqing, China

Raphaela Porsch Institute of Educational Science, University of Muenster, Muenster, Germany

Shelagh Rixon Centre for Applied Linguistics, University of Warwick, Coventry, UK

Beth Lewis Samuelson Department of Literacy, Culture and Language Education, Indiana University Bloomington, Bloomington, IN, USA

Magdalena Szpotowicz Educational Research Institute, Warsaw, Poland

Agnieszka Walczak Research and Validation Group, Cambridge English Language Assessment, Cambridge, UK

Eva Wilden English Department, University of Vechta, Vechta, Germany

Shicheng Zheng College of Foreign Languages and Cultures, Chongqing University, Chongqing, China

List of Acronyms

AO	age of onset
AoA	age of arrival for immigrants
ACTFL	American Council on the Teaching of Foreign Language
AfL	assessment <i>for</i> learning
CB	computer based
CEFR	<i>Common European Framework of Reference for Languages</i>
CLIL	content and language integrated learning
CPH	critical period hypothesis
CVIs	content validity indices
EAL	English as an additional language
EAP	English for additional purposes
EFL	English as a foreign language
ELL	early language learning
ELP	English language portfolio
ESL	English as a second language
ETS	educational testing services
EYL	English to young learners
GSE	Global Scale of English
IT	information technology
KET	Cambridge English: Key
KSAs	knowledge, skills, and abilities
L2	second language
LoE	length of exposure
LoR	length of residence
PA	peer assessment
PB	paper based
PET	Cambridge English: Preliminary
PPVT	Peabody Picture Vocabulary Test
SA	self-assessment
SAT	standards-based assessment

SBA	school-based assessment
SES	socioeconomic status
TAG	technical advisory group
TEFL	teaching English as a foreign language
TEYL	teaching English to young learners
TLU	target language use
WALT	we are learning to
YL	young learner
YLE	Cambridge English: Young Learners English

Trends, Issues, and Challenges in Assessing Young Language Learners

Marianne Nikolov

Abstract This introductory chapter aims to achieve multiple goals. The first part outlines the most important recent trends in early language learning, teaching and assessment and frames what the main issues are. The second part discusses the most frequent challenges policy makers, materials designers, test developers, researchers and teachers face. The third part introduces the chapters in the volume and explains how they are embedded in the trends. The last part suggests ideas for further research and points out some implications for educational and assessment practice.

Keywords Social dimensions • Framing assessment • Research on early language programs • Construct • Testing projects • Examinations • Uses of tests • Areas for further research

1 Introduction

The aim of this chapter is to offer insights into recent trends, emerging issues and challenges in the field of teaching and assessing young language learners and to outline which aspects the chapters in this volume highlight in various educational contexts. Recent developments are best viewed from a perspective of innovation (Davison, 2013; Davison & Leung, 2009; Kennedy, 2013). This approach to early language learning and assessment as a larger system (Markee, 2013) may allow us to understand how innovation works at various levels and how the classroom, institutional, educational, administrative, political and cultural level subsystems interact. A narrow focus on certain aspects of assessment practice is limited; innovation and change are necessary in the whole of assessment culture (Davison, 2013). The chapters in the book explore global issues and how they are embedded in local contexts. The findings may not directly translate into other situations, therefore, readers are expected to critically reflect on them and analyze how the lessons learnt can be relevant.

M. Nikolov (✉)

Institute of English Studies, University of Pécs, Pécs, Hungary

e-mail: nikolov.marianne@pte.hu

© Springer International Publishing Switzerland 2016

M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*, Educational Linguistics 25, DOI 10.1007/978-3-319-22422-0_1

Some of the studies included in the book fall into the narrow field of language testing and share information on frameworks and the time-consuming test design and validation processes of test development. Other chapters go beyond these domains and discuss results of large-scale national studies and smaller-scale classroom projects. The common denominator in these explorations reflect stakeholders' local needs. Alternative approaches to assessment, for example, peer and self-assessment, diagnostic testing, assessment *for learning*, and ways in which young learners' individual differences interact with test results are also discussed in depth. It is hoped that a wide readership will find food for thought in the book.

Specific uses of terms are clarified in the chapters and a list of acronyms is also included at the beginning of the volume. The ages covered by the term *young learners* in the chapters range from 6 to 12 or so; children in the projects learn a foreign language in the first 6 years of their studies. The use of key terms needs clarification. In this volume we follow the widely accepted tradition of using *assessment* and *testing* interchangeably, although we are aware that *assessment* is often used "as a superordinate term covering all forms of evaluation" (Clapham, 1997, xiv). The majority of sources on young learners tends to follow this tradition and this is what authors of this volume also do.

2 Main Trends in Early Language Learning and Assessment

2.1 The Social Dimension

These days, millions of children learn a foreign language (FL), most often English (EFL), in public and private schools around the Globe. The recent dynamic increase in the number of young language learners in early language programs is embedded in larger trends. Firstly, more and more people learn English as a lingua franca, aiming to achieve useful levels of proficiency in English, the means of international communication. Today, English is increasingly perceived as a basic competence and an asset for non-native speakers of English to succeed in life. Since access to English as a commodity is often limited, early language learning has a special social dimension. Proficiency in English can empower learners and early English may offer better access to empowerment over time.

These trends have important implications for curricula, assessment and equity. On the one hand, in many countries not all children have access to equal opportunities to start learning English at a young age. It has been widely observed that parents' socio-economic status plays an important role in access to English and choices of programs. In many places around the world parents empower their children by finding earlier, more intensive and better quality programs for their offspring. For example, an article in *The Economist* (December 20th 2014, p. 83) reported that 80 % of students at international schools around the world are locals because their parents want them to study later in an English speaking country and they believe

that earlier and better quality English learning opportunities allow them to do so. “When people make money, they want their children to learn English, when they make some more money, they want them to learn in English.” As a result of high investment in children’s learning of English, highly motivated parents make sure that their children learn English in the very best programs, as is documented by the recent interest in content and language integrated learning (CLIL). This new development poses new opportunities and challenges for assessment.

Parents would like to have evidence of their children’s proficiency in English at the earliest possible stage. This need has resulted in several internationally acknowledged external proficiency examinations offering young learners opportunities to take age-appropriate exams and document their level of proficiency. How these test results are used and why may vary (see e.g., Chik & Besser, 2011). Parents who want their children to get language certificates assume that the proficiency achieved at an early stage of language learning will be automatically maintained and built on over time.

Another line of test development is documented by national and international projects implemented in more and more countries as early language learning is becoming more the norm than the exception. Certain phases and steps of the arduous process of test development are discussed in five chapters in this book. Needs vary to a large extent, as the studies indicate and the uses of test results are also very different. Some projects are initiated by policy makers in order to establish a baseline or for gatekeeping purposes, others result from more bottom up initiatives based on local needs.

2.2 *An Inkblot Test or a Puzzle: ‘The Younger The Better’ vs. ‘The Slower’, or How and Why?*

The boom in early language learning is due to more and more parents’ and decision makers’ belief in *‘the younger the better’* slogan; young children are expected to outsmart older starters simply by starting at a younger age. The overwhelming optimism and overconfidence characterizing early language programs is well known in research in the social sciences and behavioral economics (Kahneman, 2011). Wishful thinking is supported by evidence in favor of one’s beliefs. The approaches to interpreting data on how young learners develop and what realistic expectations are after several years of exposure to L2 can be explained by two metaphors: an inkblot test and a puzzle (Nikolov, 2013). In the first approach, interpretations are projected into what there is in the data and they are biased by emotions, expectations, beliefs, etc. In the second approach, all data contribute to a better understanding of the whole as well as the small components of the larger picture. Although the puzzle metaphor is also limited, as it supposes a single correct outcome, it represents a more objective, scientific, and interactionist approach. The chapters in this volume hopefully add meaningful pieces to the picture of early language learning.

In recent years, concerns have been voiced about early learning of a foreign language both in national and local programs, as evidence on *'the younger the slower'* has emerged (e.g., deBot, 2014; García Mayo & García Lecumberri, 2003; Muñoz, 2006; Nikolov & Mihaljević Djigunović, 2006, 2011). Many experts have emphasized that focusing on starting age as the key variable is misleading in foreign language contexts. The age factor is not the main issue. There is a lot more to success over time. The quality and quantity of early provision, teachers, programs, and continuity are more important (Nikolov, 2000; Singleton 2014). Also, it is now widely acknowledged and documented that maintaining young learners' motivation over many years is an unexpected challenge emerging in most contexts: the earlier L2 learning is introduced, the sooner typical classroom activities and topics become boring for young learners. This is one of the reasons why there is a growing interest in integrating content areas and moving towards content-based curricula, which, in turn, pose further challenges in both teaching and assessment.

More and more stakeholders realize that offering early language learning opportunities is only the starting point. Issues related to curricula, teacher education, monitoring progress and outcomes over the years, and transition across different stages of education persist and pose new challenges (e.g., Nikolov 2009a, 2009b, 2009c; Rixon, 2013). In fact, the same old challenges are reemerging in a cyclic manner, as was implicitly predicted by Johnstone (2009).

An important shift can be observed from an emphasis on the *'fun and ease'* of early language learning to standards-based measurement of the outcomes in the target language (L2; e.g., Johnstone, 2009; Rixon, 2013, 2016 in this book). The shift towards standards is not limited to foreign language programs; it is an international trend in educational assessment for accountability in public educational policies in all subjects and competences.

2.3 Research on Early Language Learning and Teaching

Test results indicating how children progress and what levels they achieve in their L2 at the end of milestones in education are often used as one of several key variables interacting in the process of early foreign language learning and teaching. In other words, it has been realized that early language learning is not at all a simpler construct than language learning of older learner. Recent research projects apply all kinds of L2 tests as one of many data collection instruments in order to answer larger research questions, as they aim to build and test models of early foreign language learning. An important area of explorations concerns how young learners' individual differences, including attitudes, motivation, aptitude, anxiety, self-perceptions, self-confidence, strategies, etc., contribute to their development in their L2 (Bacsa & Csikos, 2016; Mihaljević Djigunović, 2016; Nikolov, 2016 all in this book). Another important avenue of explorations

gaining ground looks into how learners' first (L1) and other languages interact with one another over time (e.g., Nikolov & Csapó, 2010; Wilden & Porsch, 2016 in this volume).

Yet another important line of research examines how different types of curricula contribute to early language learning. Traditional FL programs are often supplemented or substituted by early content and language integrated learning curricula (CLIL). Overall, these research studies aim to find out not only what level of proficiency children achieve in their L2, but they also want to offer explanations as to how and why. The type of curriculum has important implications for the construct as well as for the way the curriculum is implemented in the classroom. On the one hand, some recent studies focus on the relationships between contextual factors and classroom processes. Highly age-appropriate innovative approaches, including *assessment for learning* (AfL, Black & Wiliam, 1998), diagnostic (Alderson, 2005; Nikolov, 2016), peer and self-assessment are examined in ELL contexts (Butler, 2016; Hung, Samuelson & Chen, 2016 in this volume). On the other hand, some research projects aim to find out how and to what extent different curricula contribute to L2 development.

In recent years, the field of early language learning research has grown remarkably. Many new studies have been published in refereed journals. (See for example Special Issues of *English Language Teaching Journal*, 2014 (3) edited by Copland and Garton; *International Journal of Bilingualism*, 2010 (3) edited by Nikolov; and *Studies in Second Language Learning and Teaching*, 2014 (3) edited by Singleton.) A range of books and research studies are available on the early teaching and learning of modern foreign languages offering food for thought for decision makers, teachers, teacher educators and researchers. (For critical overviews see e.g., Murphy, 2014; Nikolov & Mihaljević Djigunović, 2006, 2011.) Publications on large scale surveys give insights into the big picture (e.g., Edelenbos, Johnstone, & Kubanek, 2007; Emery, 2012; Garton, Copland & Burns, 2011; Rhodes & Pufahl, 2008; Rixon, 2013, 2016 in this volume). Excellent handbooks offer classroom teachers guidance on age-appropriate methodology and assessment (e.g., Cameron, 2001; Curtain & Dahlberg, 2010; Jang, 2014; McKay, 2006; Pinter, 2006, 2011).

The growing body of empirical studies (e.g., Enever, 2011; Enever, Moon, & Raman, 2009; García Mayo & García Lecumberri, 2003; Muñoz, 2006; Nikolov 2009a, 2009b) applies some kinds of tests, as they implement quantitative or mixed research methods (Nikolov, 2009c) and analyze test results in interaction with other variables. Testing young language learners' progress over time in their classrooms and their proficiency at the end of certain periods are often the aspects of studies. Thus, the assessment of young learners has become a central issue in early language learning research and daily practice (Butler, 2009; Inbar-Lourie & Shohamy, 2009; Johnstone, 2009; McKay, 2006; Nikolov & Mihaljević Djigunović, 2011; Rixon, 2013), as chapters in the present volume indicate. As Rixon (2016) put it in the title of her chapter, these developments in assessment represent the 'Coming of Age'.

3 Challenges in Early Language Learning, Teaching, and Assessment

3.1 *The Construct and Frameworks of Assessment*

The trends outlined above have important implications for the construct. Assessment of young language learners in early learning contexts was first brought to the attention of the testing community as a bona fide domain in a special issue of *Language Testing* edited by Pauline Rea-Dickins (2000). In her editorial she emphasized an array of issues: processes and procedures teachers used in their classrooms to monitor their learners' development and their own practice, the assessment of young learners' achievement at the end of their primary education, and teachers' professional development. At that time high hopes were typical in publications on early language programs and hardly any comparative studies were available on younger and older EFL learners. However, the field was characterized by variability and diversity, as Rea-Dickins pointed out (p. 119).

Over the past 15 years, the picture has become even more complex for several reasons:

- (1) The constructs (Inbar-Lourie & Shohamy, 2009; Johnstone, 2009) cover various types of curricula;
- (2) More evidence has been found on young learners' varied achievements and on how their individual differences and contextual variables, including teacher-related ones, contribute to outcomes over time (for an overview see Nikolov & Mihaljević Djigunović, 2011).
- (3) Accountability poses a recent challenge as standards-based assessment in early language programs has been introduced in many educational contexts.

The emergence of accountability in early language learning is not an unexpected phenomenon. As Johnstone (2009, p. 33) pointed out, the third phase of early learning became a "truly global phenomenon and possibly the world's biggest policy development in education. Thus, meeting 'the conditions for generalized success' becomes an awesome challenge." The task is to establish to what extent and in what conditions early language learning can be claimed to be successful in a range of very different situations where conditions vary a lot. Stakeholders are interested in seeing results. What can young learners actually do after many years of learning their new language? An important challenge for researchers concerns what curriculum is best and what realistic age-appropriate achievement targets are included in language policy documents. Once curricula are defined, and frameworks are in place, the construct and expected outcomes have to be in line with how young learners develop and how their motivation can be maintained over years.

Although early language learning is often seen as a simple proposition (start learning early), a lot of variation characterizes models according to when programs start, how much time they allocate, what type of curriculum and method they apply, who the teachers are, and how they implement the program. In the European contexts (Edelenbos, Kubanek, & Johnstone, 2007; Johnstone, 2009), three types of curricula are popular: (1) awareness raising to languages; (2) traditional FL programs

offering one to a few classes per week, and (3) content and language integrated learning (CLIL) curricula where up to 50% of the curriculum is taught in the L2. The first type does not aim to develop proficiency in an L2; the other two usually define L2 achievement targets. CLIL programs have become popular in Europe, Asia and South America. CLIL is typically taught by non-native teachers of English, and ‘could be interpreted as a foreign language enrichment measure packaged into content teaching’ (Dalton-Puffer, 2011, p. 184). In most schools ‘CLIL students nearly always continue with their regular foreign language program alongside their CLIL content lessons’ (p. 186). What the construct is in these two programs is one of the main challenges in early language learning research. As has been indicated, the increased interest in early CLIL programs is due to growing evidence that in traditional (type 2) programs children develop at a very slow rate and many of the motivating activities lose their appeal and soon become boring. Therefore, integrating not only topics from the main curriculum (as in type 2 programs), but also teaching subjects in the target language is supposed to result in killing two problems with one stone: a focus on intrinsically motivating content also offers opportunities to acquire L2 skills in all four skills. This means that both content and language have to be assessed.

As for the construct of early language learning, Inbar-Lourie and Shohamy (2009) suggest that different types of curricula should be seen along a continuum between programs focusing on language and content. Awareness raising is at one end, FL programs somewhere in the middle, and CLIL and immersion at the other end. They propose that in early language programs language should be “a tool for gaining knowledge and meaning making and for developing cognitive processing skills” (p. 91). In this framework, L2 is closely linked to the overall curriculum and learners’ L1, and the larger view of assessment culture where assessment is a means to improve. Their proposed framework integrates widely accepted principles of age-appropriate classroom methodology as well as assessment. The challenges concern how curricula define the aims set for language and content knowledge, and cognitive and other abilities and skills.

Achievement targets in L2 tend to be modest in early language programs. Young learners are not expected to achieve native level (e.g., Curtain, 2009; Haenni Hoti, Heintzmann, & Müller, 2009; Inbar-Lourie & Shohamy, 2009). Frameworks tend to build on developmental stages in early language programs and reflect how young learners move from chunks to analyzed language use (Johnstone, 2009). Most curricula include not only L2 achievement targets, but comprise further aims. Early learning is meant to contribute to young learners’ positive attitudes towards languages, language learning, speakers of other languages, and towards learners’ own culture and identity (e.g., Prabhu, 2009). In addition to linguistic and affective aims, they often include aims related to cognition, metacognition and learning strategies. There is a controversy in the multiplicity of aims. Testing in most contexts focuses on L2 achievements and the other aims are not assessed at all or they are discussed only in a few research projects. Testing in early language learning programs is most often concerned with: (1) how learners progress in their L2 over time and (2) what levels of proficiency they achieve in some or all of the four skills

by the end of certain periods. In addition to these areas, there is a need to explore how teachers assess YLs and how classroom practices interact with children's attitudes, motivation, willingness to communicate, anxiety, self-confidence and self-perception over time.

Early language learning assessment frameworks define the main principles of teaching and assessing young learners and aim to describe and quantify what children are expected to be able to do at certain stages of their L2 development (e.g., Curtain, 2009; Jang, 2014; McKay, 2006; Nikolov, 2016 in this volume). Frameworks developed in Europe tend to use the *Common European Framework of Reference for Languages (CEFR)* (Council of Europe, 2001) as a point of departure, despite the fact that it was not designed for young learners (e.g., Hasselgren, 2005; Pižorn, 2009; Papp & Salamoura, 2009; Papp & Walczak, 2016 in this volume). In contrast, research projects on early CLIL tend to follow a different tradition unrelated to testing children or standards-based testing. They frame CLIL as an add-on to FL instruction and analyze young learners' performances along three criteria (complexity, accuracy, and fluency) used in second language acquisition research (e.g., Hausen & Kuiken, 2009). Such a framework, however, is hardly suited to document very slow development (see e.g., Bret-Blasco, 2014).

Tests for young learners have been developed for various purposes. Standards-based tests are used in national and international projects and external examinations as well as in smaller-scale research studies. The majority of national and international projects tend to apply standards aligned to levels in *CEFR*. Test construction and validation is a long and complex process. Some important work has been published on the process of developing frameworks, *can do statements*, designing and validating tests for various purposes, for example, for large-scale proficiency tests, research projects and teacher-based assessments. These areas are discussed in five chapters.

3.2 National, International and Local Testing Projects

Early language learning is compulsory in many places. In Europe, it is more the norm than the exception. National curricula typically include achievement targets and in some countries national proficiency exams are implemented annually (e.g., in Germany, Wilden & Porsch, 2016 in this volume, in Poland, Szpotowicz & Campfield, 2016 in this volume; in Slovenia, Pižorn, 2009; in Switzerland, Haenni Hoti, Heinzmann & Müller, 2009; in Hungary, Nikolov & Szabó, [in press](#)). How these tests are administered, how the test results are used and how tests impact teaching and learning raises further questions. They have to be discussed in each particular situation bearing in mind the particulars of the assessment culture.

International research projects have also been implemented to collect test data for comparative purposes and to answer questions related to the rate and level of L2 development. For example, a longitudinal study, the Early Language Learning in Europe (ELLiE) project aimed to examine what level young learners achieved in a foreign language at public schools in England, Italy, the Netherlands, Poland, Spain,

Sweden and Croatia. In addition to L2, other factors were also included to find out how they contributed to processes and outcomes in the target languages as well as in the affective domain (Enever, 2011; Mihaljević Djigunović, 2012). Researchers faced challenges similar to those in previous longitudinal studies on early language learning (Enever, 2011; García Mayo & García Lecumberri, 2003; Muñoz, 2006). The same tests were used over the years to collect valid and reliable results on participants' L2 development and a single task was used for each skill.

Assessment projects are often narrowly limited and they aim to seek answers to research questions emerging from practice. For example, how achievement tests are applied by teachers (Peng & Zheng, 2016), and how innovative assessment techniques can change classroom processes (Butler, 2016; Hung, Samuelson & Chen, 2016, both in this volume). Other projects use tests in order to build new models or to test existing ones to find out to what extent they can reflect realities in early FL classrooms (Mihaljević Djigunović, 2016; Bacsa & Csikos, 2016; see chapters in this volume).

3.3 International Language Tests for Young Language Learners

In recent years, several international examinations have been developed and made available to young language learners whose parents want them and can afford them. Three widely known exams offer certificates on children's proficiency in English: (1) Cambridge Young Learners English Tests (www.cambridgeesol.org/exams/young-learners), (2) Pearson Test of English Young Learners (www.pearsonpte.com/PTEYoungLearners); and (3) TOEFL Primary (https://www.ets.org/toefl_primary). These examinations fall somewhere in the middle of the language–content continuum with a focus on some typically taught topics young language learners can be realistically expected to know. The levels cover A1 and A2 in the *CEFR* (Council of Europe, 2001). Besides aural/oral skills literacy skills are also included. How much work is devoted to developing and validating exams is discussed in three of the chapters (Benigno & de Jong, 2016; Hsieh, 2016; Papp & Walczak, 2016). Unfortunately, hardly any studies explore how these proficiency exams impact classroom processes or how children taking them benefit from their experiences in the long run. It would also be important to know how they maintain and further develop their proficiency after taking examinations.

3.4 Assessment for Learning

Recent research on early language learning assessment has focused on how teacher-based assessment can scaffold children's development in their L2 knowledge and skills so that they can apply their learning potential (Sternberg & Grigorenko, 2002).

In this developmental framework of *assessment for learning* children should benefit from ongoing classroom testing. Teachers consider assessment as an integral part of their teaching. They build on test results to inform their teaching (Black & Wiliam, 1998; Davison & Leung, 2009; McKay, 2006). This way the teaching process can be sensitive to readiness to develop (McNamara & Roever, 2006). These are key points in teacher-based assessment: learning oriented assessment is based on these principles (Nikolov, 2011, 2016 in this volume). Very little has been published on how assessment for learning works in early foreign language contexts and how teachers apply their diagnostic competence. 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” (Edelenbos & Kubanek-German, 2004, p. 48) is definitely an area where further classroom studies are necessary.

These approaches to assessment and uses of test results definitely require teachers to reflect on their practices in a new way. The visual and written samples in Rixon’s (2016) chapter clearly document a totally different assessment culture from what one would find in classrooms where the tradition is more focused on *assessment of learning*. Three other chapters in this book discuss further aspects of learning oriented assessment. Nikolov’s (2016) account shares outcomes of a diagnostic testing project: framework, main principles, *can do statements*, topics and task types designed for young learners in the first six grades of primary school. Butler’s (2016) overview offers multiple insights into how self-assessment can be used in various domains, whereas Hung, Samuelson and Chen report on how peer-, self-, and teacher-based assessments were implemented in the EFL classroom where traditions were not in line with assessment for learning principles.

3.5 *What Tests Are Used and How*

Researching and documenting how certain tests work with young learners is time-consuming and this is an area where there is a need and a lot of room for further work. Similarly to the most brilliant age-appropriate teaching materials and tasks, the most valid and reliable tests can also be misused or abused. The chapters in this volume offer insights into some actual tests and how researchers and teachers applied them. One interesting trend needs pointing out: most of the tests discussed in the early language learning assessment literature and these chapters are similar to language tests widely used and accepted in the L2 testing literature. However, some tests and criteria for assessment are borrowed from other traditions: for example, oral production was assessed along complexity, accuracy, and fluency in Bret Blasco’s (2014) study on CLIL.

As these are key issues in assessment, a detailed and critical analysis should focus on what tests are used in assessment projects involving young learners. Often a single task is used to tap into a skill and the same test is used over the years to document development (e.g., Bret Blasco, 2014; Enever, 2011). Recently elicited

repetition has been also used to assess speaking. It is important to approach these questions from the learners' and teachers' perspectives as well and to explore how tests can be linked to offer more reliable insights into young learners' development (e.g., Nikolov & Szabó, 2012; Szpotowicz & Campfield, 2016 in this volume). There is a lot of potential in learning about the traditions in the fields of second language acquisition and language testing, and most probably both areas would benefit from a comparative analysis.

4 How This Volume Contributes to a Better Understanding of the Challenges in Young Learners' Assessment and to Advancing the Field

Assessing young learners of a FL is a complex area requiring knowledge of age-appropriate classroom methodology, including teacher- and standards-based language assessment, second language acquisition, research methodology and the actual contexts. The issues and challenges should be approached, researched and interpreted as subcomponents of innovation requiring more than change in a single aspect. The complexity of teaching and assessment results from the fact that not only the constructs vary but also because young learners' individual differences, languages, and knowledge interact with specific contextual and teacher- and parent-related variables. In what follows, let us overview what this volume comprises.

The chapters focus on various aspects of assessment in early EFL programs around the world. The first two papers draw the larger picture; Marianne Nikolov and Shelagh Rixon outline the main trends, issues and challenges and the reasons why recent international developments represent the 'coming of age'. They provide an overview on how the main points are embedded in larger trends, and discuss the construct, various frameworks for test development, international and national projects and international examinations designed to tap into children's proficiency. These two chapters offer insights also into teacher-based alternative approaches: diagnostic and self-assessment.

Chapters "The "Global Scale of English Learning Objectives for Young Learners": A CEFR-Based Inventory of Descriptors, A Framework for Young EFL Learners' Diagnostic Assessment: 'Can Do Statements' and Task Types, Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers' Perspectives, Developing and Piloting Proficiency Tests for Polish Young Learners, and The Development and Validation of a Computer-Based Test of English for Young Learners: Cambridge English Young Learners" focus on how challenges are overcome in test development. Three papers present findings on the early stages and the fourth one on how a validated paper and pencil test can go online. In chapter "The "Global Scale of English Learning Objectives for Young Learners": A CEFR-Based Inventory of Descriptors", Veronica Benigno and John de Jong give an account of how Pearson developed their first batch of CEFR-based

inventory of young learners descriptors. Chapter “[A Framework for Young EFL Learners’ Diagnostic Assessment: ‘Can Do Statements’ and Task Types](#)”, by Marianne Nikolov, discusses how a framework was developed for young EFL learners for diagnostic assessment purposes and presents *can do statements* and task types found relevant in a national project in Hungary. In chapter “[Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers’ Perspectives](#)”, Ching-Ni Hsieh offers test validity evidence for TOEFL Primary: she discusses how content representativeness was ensured at ETS by integrating teachers’ views in the process. In chapter “[Developing and Piloting Proficiency Tests for Polish Young Learners](#)”, Magdalena Szpotowicz and Dorota E. Campfield reveal how they piloted proficiency tests and used children’s feedback in a national testing project in Poland. The very first examination for young learners of English was offered by Cambridge. In chapter “[The Development and Validation of a Computer-Based Test of English for Young Learners: Cambridge English Young Learners](#)”, Szilvia Papp and Agnieszka Walczak offer insights into how a computer-based test was developed and validated to make the tests more readily available.

Chapters “[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, A Longitudinal Study of a School’s Assessment Project in Chongqing, China, Individual Learner Differences and Young Learners’ Performance on L2 Speaking Tests](#)”, and [The Role of Individual Differences in the Development of Listening Comprehension in the Early Stages of Language Learning](#)” present five complex research projects where testing young learners’ L2 played a key part. In chapter “[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](#)”, Eva Wilden and Raphaela Porsch intended to find out if learning EFL from the first or the third year in German primary schools was a better model by examining young learners’ EFL listening and reading comprehension at the end of their primary education. Besides the modest advantage for earlier starters, their study revealed that children’s proficiency in other languages interacted with the outcomes in important and unexpected ways. In chapter “[A Longitudinal Study of a School’s Assessment Project in Chongqing, China](#)”, Jing Peng and Shicheng Zheng compare and contrast outcomes of a longitudinal teacher-based assessment study implemented at a school in China. They discuss how children performed on two achievement tests based on two course books and triangulate their findings by interviewing teachers. In chapter “[Individual Learner Differences and Young Learners’ Performance on L2 Speaking Tests](#)”, Jelena Mihaljević Djigunović discusses the dynamic changes in the ways how young Croatian language learners’ individual differences, motivation and self-concept, contributed to their performance on EFL speaking tests over a four-year period. The aim of chapter “[The Role of Individual Differences in the Development of Listening Comprehension in the Early Stages of Language Learning](#)”, by Éva Bacsa and Csaba Csíkos, was to model how aptitude, motivation anxiety, learners’ beliefs and their parental background interacted in the development of EFL in a semester-long study involving young learners in a small town in Hungary.

The last two chapters provide insights into how peer-, self-assessment and teacher assessment interact with one another. Yuko Goto Butler, in chapter “*Self-Assessment of and for Young Learners’ Foreign Language Learning*”, offers a critical overview of research into self-assessment *of* and *for* young learners’ foreign language learning and proposes five dimension for developing further research instruments, thus linking teaching, assessment and learning. The context of the final chapter is Taiwan. Yu-ju Hung, Beth Lewis Samuelson and Shu-cheng Chen explore the relationships between peer- and self-assessment and teacher assessment of young EFL learners’ oral presentations by applying both the teacher’s and her students’ reflections for triangulation purposes.

5 Areas for Further Research and Implications for Practice

This volume outlines some of the key areas where research has been conducted. Similar inquiries would allow us to find out how results would compare in other contexts. Researchers, including classroom teachers, should consider how replication studies could offer useful information on learners’ achievements in their countries and classrooms. Data collection instruments can be of invaluable help with instructions on how to apply them. Such data repositories, for example at <http://iris-database.org/iris/app/home/index>, are available. Test development is an extremely challenging and expensive process. Questionnaires, interviews, etc. also require special expertise to develop and validate. Sharing them would allow the early language learning field to advance more rapidly.

It is also important to note which key areas are not discussed in this book in full detail or at all, and where more research is needed.

- (1) In order to answer research questions related to the larger picture on early start programs, studies should aim to find out in what domains younger learners excel over time and why this is the case. This kind of research should work towards testing models of early language learning. Studies should include proficiency tests on learners’ aural/oral and literacy skills in their L1, L2, L3. Other instruments should tap into individual differences of young learners and their teachers, and contextual variables (including characteristics of programs, materials, methods, the quality of teaching) interacting in children’s development over several years. The main benefits of an early start are most probably not in higher L2 proficiency over time and this hypothesis may have important implications for language policy, curriculum design, teacher education and classroom practice.
- (2) Hardly any studies look into the relationships between access to early foreign language learning opportunities, assessment, and equity. Do all children have equal opportunities? Research is necessary to examine how parents’ motivation, learners’ socio economic status and achievements on tests interact and how test results are used.
- (3) A recurring theme in early language teaching programs concerns transition and continuity. Studies should go beyond the early years and focus on how teachers build on what learners can do in later years and what role assessment practices play in the process. In other words, research is necessary into how children are taught and assessed, and how teachers can apply diagnostic information in their teaching.

- (4) The impact of different kinds of assessment on young language learners, their teachers, and the teaching-learning process should be explored in depth. Teachers' and learners' emic perspectives are hardly ever integrated into studies. Exploring teachers' and their learners' beliefs and lived experiences could reveal why implementing innovation often poses a major challenge. Case studies could offer insights on what it means to a child to take an external examination, what challenges learners and their teachers face due to parental pressure to produce results, and why teachers may resist change in their teaching and testing practices.
- (5) It would be essential to learn more about the ways in which achievement targets defined in curricula are assessed by teachers on a daily basis. How they balance giving children feedback on their progress in test results with maintaining their motivation and keeping their debilitating anxiety low.
- (6) Yet another avenue for classroom research for practicing teachers should explore how teachers apply traditional (assessment *of* learning) and innovative assessment techniques (assessment *for* learning, peer and self-assessment). How do they use criteria for assessing speaking and writing and keys on closed items and students' responses to open items? How do they integrate other aspects of students' behavior into their assessments, for example, their willingness to communicate, attitudes, motivation, aptitude, anxiety?
- (7) Very little is known about testing learners' knowledge and skills in CLIL programs. Exploratory classroom studies are needed to find out how teachers tease out the two domains and how they can diagnose if learners' weaknesses are in their L2 or in the subject matter.

The studies in this volume discuss various aspects of test development, outcomes of large-scale surveys, national assessment projects, and innovative smaller-scale studies. The ideas shared and the frameworks and instruments used for data collection should be of interest to both novice and experienced teachers, materials and test developers, as well as for researchers. Readers should bear in mind which of the main points are worth further explorations. It is hoped that the volume offers exciting new ideas, and result in innovation and change.

References

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.
- 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*. New York: Springer.
- 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*. New York: Springer.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–71.
- Bret Blasco, A. (2014). *L2 English learners' oral production skills in CLIL and EFL settings: A longitudinal study*. Doctoral dissertation, Universitat Autònoma de Barcelona.
- Butler, Y. G. (2009). Issues in the assessment and evaluation of English language education at the elementary school level: Implications for policies in South Korea, Taiwan, and Japan. *The Journal of Asia TEFL*, 6, 1–31.

- 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*. New York: Springer.
- Cameron, L. (2001). *Teaching languages to young learners*. Cambridge: Cambridge University Press.
- Chik, A., & Besser, S. (2011). International language test taking among young learners: A Hong Kong case study. *Language Assessment Quarterly*, 8, 73–91.
- Clapham, C. (1997). Introduction. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education. Volume 7. Language testing and assessment* (pp. xiii–xix). Dordrecht: Kluwer Academic Publisher.
- Copland, F., & Garton, S. (Eds.). (2014). *English Language Teaching Journal. Special Issue*.
- 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. 59–82). Berlin, Germany: Mouton de Gruyter.
- Curtain, H. A., & Dahlberg, C. A. (2010). *Languages and children – Making the match: New languages for young learners* (4th ed.). Needham Heights, MA: Pearson Allyn & Bacon.
- Dalton-Puffer, C. (2011). Content and language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, 31, 182–204.
- Davison, C. (2013). Innovation in assessment: Common misconceptions. In K. Hyland & L. L. C. Wong (Eds.), *Innovation and change in English language education* (pp. 263–277). New York: Routledge.
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43, 393–415.
- deBot, K. (2014). The effectiveness of early foreign language learning in the Netherlands. *Studies in Second Language Learning and Teaching*. doi:10.14746/ssllt.2014.4.3.2
- Edelenbos, P., Johnstone, R., & Kubanek, A. (2007). *Languages for the children in Europe: Published research, good practice and main principles*. 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, 259–283.
- Emery, H. (2012). *A global study of primary English teachers' qualifications, training and career development*. London: British Council.
- Enever, J. (Ed.). (2011). *ElliE: Early language learning in Europe*. London: British Council.
- Enever, J., Moon, J., & Raman, U. (Eds.). (2009). *Young learner English language policy and implementation: International perspectives*. Reading, UK: Garnet Education Publishing.
- García Mayo, M. P., & García Lecumberri, M. L. (Eds.). (2003). *Age and the acquisition of English as a foreign language*. Clevedon: Avon/Multilingual Matters.
- Garton, S., Copland, F., & Burns, A. (2011). *Investigating global practices in teaching English to young learners*. London: British Council.
- 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.
- Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing*, 22, 337–354.
- Hausen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.
- 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*. New York: Springer.
- Hung, Y.-J., Samuelson, B. L., & Chen, S.-C. (2016). The relationships 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*. New York: 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.
- International schools: The new local. (2014, December 20). *The Economist*, pp. 83–84.
- Jang, E. E. (2014). *Focus on assessment*. Oxford: Oxford University Press.
- 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: Garnet Education Publishing Ltd.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Allen Lane/Penguin Books.
- Kennedy, C. (2013). Models of change and innovation. In K. Hyland & L. L. C. Wong (Eds.), *Innovation and change in English language education* (pp. 13–27). New York: Routledge.
- Markee, N. (2013). Contexts of change. In K. Hyland & L. L. C. Wong (Eds.), *Innovation and change in English language education* (pp. 28–43). New York: Routledge.
- McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell Publishing.
- Mihaljević Djigunović, J. (2012). *Early EFL learning in context – Evidence from a country case study*. London: The British Council.
- 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*. New York: Springer.
- Muñoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon: Avon/Multilingual Matters.
- Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford: Oxford University Press.
- Nikolov, M. (2000). Issues in research into early second language acquisition. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners: International perspectives* (pp. 21–48). Pécs: University Press Pécs.
- Nikolov, M. (Ed.). (2009a). *The age factor and early language learning*. Berlin/New York: Mouton de Gruyter.
- Nikolov, M. (Ed.). (2009b). *Early learning of modern foreign languages: Processes and outcomes*. Clevedon, UK: Multilingual Matters.
- Nikolov, M. (2009c). The age factor in context. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 1–38). Berlin, Germany/New York, NY: Mouton de Gruyter.
- Nikolov, M. (Ed.). (2010). *International Journal of Bilingualism. Special Issue*.
- Nikolov, M. (2011). Az angol nyelvtudás fejlesztésének és értékelésének keretei az általános iskola első hat évfolyamán [A framework for developing and assessing proficiency in English as a foreign language in the first six years of primary school]. *Modern Nyelvoktatás*, XVII(1), 9–31.
- Nikolov, M. (2013, August). *Early language learning: Is it child's play? Plenary talk*. EUROSLA Conference, Amsterdam. Retrieved from <http://webcolleges.uva.nl/Mediasite/Play/7883cb9b1cb34f98a21fb37534fc1ec61d>
- Nikolov, M. (2016). 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*. New York: Springer.
- Nikolov, M., & Csapó, B. (2010). The relationship between reading skills in early English as a foreign language and Hungarian as a first language. *International Journal of Bilingualism*, 14, 315–329.
- 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. (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* (pp. 347–363). Cambridge: UCLES/Cambridge University Press.
- Nikolov, M., & Szabó, G. (in press). 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 in honour of Peter Medgyes*. Harlow: Pearson Education.
- Papp, S., & Salamoura, A. (2009). *An exploratory study into linking young learners' examinations to the CEFR* (Research Notes, 37, pp. 15–22). Cambridge: Cambridge ESOL.
- 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*. New York: 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*. New York: Springer.
- Pinter, A. (2006). *Teaching young language learners*. Oxford: Oxford University Press.
- Pinter, A. (2011). *Children learning second languages*. Basingstoke: 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, The Netherlands: Cito/EALTA.
- Prabhu, N. S. (2009). Teaching English to young learners: The promise and the threat. In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: international perspectives* (pp. 43–44). Reading, UK: Garnet Education Publishing.
- Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing*, 17(2), 115–122.
- Rhodes, N. C., & Pufahl, I. (2008). *Foreign language teaching in U.S. Schools: Results of a national survey*. Retrieved from http://www.cal.org/projects/Exec%20Summary_111009.pdf
- Rixon, S. (2013). *British Council survey of policy and practice in primary English language teaching worldwide*. London: British Council.
- Rixon, S. (2016). Do developments in assessment represent the 'coming of age' of young learners English language teaching initiatives? The international picture. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.
- Singleton, D. (2014). Apt to change: The problematic of language awareness and language aptitude in age-related research. *Studies in Second Language Learning and Teaching*. doi:10.14746/ssllt.2014.4.3.9.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge: Cambridge University Press.
- Szpotowicz, M., & Campfield, D. E. (2016). Developing and piloting proficiency tests for Polish young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.
- 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*. New York: Springer.

Do Developments in Assessment Represent the ‘Coming of Age’ of Young Learners English Language Teaching Initiatives? The International Picture

Shelagh Rixon

Abstract This chapter draws upon two pieces of recent research undertaken for the British Council and in co-operation with Cambridge English concerning the state of the art of the teaching of English as a Foreign Language at primary school level, and of assessment of children’s English in particular. It is shown that, while some advances have been made in curricular planning over the past 15 years in different parts of the world and hence in target level-setting, the actual practices applied in assessment are not well-conceived in all places. In addition, the use of assessment data to improve continuity and coherence in English Language Teaching after transition from one level of schooling to another remains in most cases an opportunity missed.

Keywords Assessment • CEFR • English Language Teaching • Primary school • Target-setting • Transition

1 Introduction

The age range of learners discussed in this chapter is from 5 to 12 years old, corresponding with the ages between which children attend primary/elementary school in many countries. The focus is on the teaching of English to young learners (TEYL) in state rather than private schools.

The teaching of languages to primary school aged children has been described as one of the greatest areas of educational policy change world wide in the last 30 years.

S. Rixon (✉)

Centre for Applied Linguistics, University of Warwick, Coventry, UK

e-mail: shelaghr@hotmail.com

© Springer International Publishing Switzerland 2016

M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*, Educational Linguistics 25, DOI 10.1007/978-3-319-22422-0_2

Indeed EYL is often not just an educational project, but also a political and economic one. A remarkable number of governments talk not only about the need to learn a foreign language but of an ambition to make their country bilingual. (Graddol, 2006, pp. 88–91)

It is very well accepted, almost a truism, that attitudes to and practices within assessment are a strongly determinant factor in how teaching and learning takes place. Many authorities (e.g., Andrews, 2004; Black & Wiliam, 1998a, 1998b; Henry, Bettinger & Braun, 2006) have suggested that an indispensable way to promote and sustain an intended educational innovation or improvement, whether at curriculum or methodological level, is to adjust the assessment system so that it is coherent with the teaching and its content. Conversely, the best way to thwart change is to take no accommodating action with regard to assessment. In earlier times, this was often seen as applying principally to the formal, high-stakes, testing/examination system. See Rea-Dickins and Scott (2007) for a discussion with regard to language testing. However, attention to assessment at the classroom level, particularly “assessment for learning” or AfL (Black & Wiliam, 1998a) has more recently been shown to have an enormous influence on developing learners’ capacity for self-direction and more autonomous learning. Consideration of the range of assessment practices in the developing field of teaching English to primary school aged children is therefore surely of high relevance.

This chapter investigates the stated policies of regional and national educational authorities as well as the practices and perceptions of selected young learners’ practitioners with regard to the different roles that assessment currently plays in primary school level English Language Teaching. The focal areas concern its potential roles regarding quality of teaching and learning, in setting and checking targets and standards, for coherence between different levels of schooling and, in some contexts, for justice in allocating scarce educational opportunities. The argument is that a curricular/teaching innovation in a given context cannot be said to have ‘come of age’, until assessment is well understood and appropriately used at the classroom, local education authority and national education levels to support the intentions behind the innovation.

2 The History So Far

It might be hoped that, near the end of a 30-year or more ‘new wave’ of interest in the teaching of languages to young children, much would have fallen into place at the level of a range of recommended practices as well as generally agreed theory. This, however, cannot be taken for granted. The history of TEYL initiatives over the past 30 years has often been one of enthusiasm followed by some turbulence and often disappointment. There has often been more rhetoric on the part of educational authorities than willingness to put in place tangible support in terms of money and

time for training opportunities for teachers and for the supply or creation of suitable materials. Planning efforts have also frequently not been equal in energy to the content of ministerial decrees. See Enever and Moon (2009, pp. 5–20) for a fuller discussion of these issues. Surveys made near the beginning of the ‘boom’ and in the more recent past have shown that, time after time, compromises have been made with EYL initiatives, often, it seems, for the sake of speed of implementation for narrowly political motives. The main points of strain have frequently been found to be in the fundamental area of provision and preparation of teachers so that they are professionally well equipped to carry through the innovation. Rixon, summarising a survey of the decade from 1990 to 1999 found the following pattern in numerous state school systems. There was either:

... a relaxation of the official criteria or qualifications for eligibility as a teacher of English in the primary school system.

or

... an adequate supply of officially qualified teachers but considerable controversy about whether those teachers were adequately prepared in terms of language and methodology. (Rixon, 2000, p. 161)

This uncertainty over teacher supply and quality came in addition to considerable fluidity in, or, in some cases, absence of, specifications of syllabus content for primary-aged learners of English. Such fluidity was not in itself a bad thing, but was clearly inimical to any attempt to specify and promote assessment instruments which might, for example, support ongoing monitoring or lead to coherent and usable summative information on what had been learned at different stages of primary schooling.

3 Developments and Research in EYL Assessment Up Until the Early Twenty-First Century

As we have seen above, there was evidence even in 2000, nearly 20 years after the first stirrings of interest internationally in teaching English to younger children, that in many contexts EYL was still finding its feet in terms of decisions on curriculum and methodology and in recruiting or preparing teachers who were confident in the skills and knowledge they would need to function well in the classroom. Meanwhile, several strands of practice and thinking in the assessment area had been developing both in the English language teaching (ELT) world and the general mainstream educational world. These offered potentially useful approaches that could help tie together teaching and assessment in order to create more robust and coherent experiences for children learning English in school. However, these developments in themselves could also be seen as presenting yet more to be taken on board by Young Learners teachers still developing their new professional roles.

It was only in the late 1990s (e.g., Rea-Dickins & Rixon, 1997, 1999) that the assessment of the English language learning of primary school aged children started to be raised by researchers as an area of particular concern with the different purposes which assessment might serve in this area being spelled out and discussed. Among these the purposes of monitoring learning, allowing formative development and providing information to facilitate transition between one level of schooling and another were highlighted by writers who often had the improvement of pedagogy high amongst their priorities. For example, the models for assessment of children's language development that were deemed by Rea-Dickins and Rixon in their 1997 chapter to be the most interesting and likely to influence children's language learning for the better were mostly derived from work in mainstream UK schools with children with English as a second language (ESL, now known as EAL – English as an Additional Language). The techniques used in the main emphasised classroom assessment, continuous observation and record-keeping, with concern always for the development of the individual child and thus with a largely formative purpose.

It was recognised that this mainly classroom based tracking and record-keeping approach might not be familiar (and might hold little appeal) in contexts and educational systems which, for selection or other administrative purposes, required more speedily arrived-at summative results for large numbers of learners. The assessment events of this latter type might take place at the end of a term or school year or near the end of primary education. However, it was striking that this summative style of assessment was what also seemed to predominate in day-to-day classroom assessment in many of the EYL contexts that Rea-Dickins and Rixon were at that time researching. In an international survey involving 122 primary school teachers of English (Rea-Dickins & Rixon, 1999) 100 % of teachers' self-reports gave an account of classroom assessment practices which were exclusively based on 'paper and pencil' written tests and quizzes. This was in spite of the fact that they also claimed to be focusing mostly on developing speaking and listening skills.

From the late 1990s to the early 2000s, new editions of standard textbooks on language testing (e.g., Hughes, 2003) inserted new chapters on assessing children. However, the discussion tended to remain at the generic level of principle and the hunt for 'child-friendly' items largely within the familiar formats used with older learners. In the early 2000s, there came a welcome departure with the publication of an account of EYL assessment (Ioannou-Georgiou & Pavlou, 2003) which seemed to consider the area in a completely new way. Refreshingly, this book started with a persuasive discussion of portfolio-based evidence as a feasible norm for young learners' (YL) assessment and only then worked its way through to child-friendly versions of gradually more conventional and familiar assessment practices by the end of the book. This was a bold reversal of more timid accounts.

None of these works, however, included research into specific local understandings and practice in Young Learners assessment. A special issue of the journal *Language Testing* (Rea-Dickins, 2000) had addressed this area, albeit with a mainly European focus. Recently, research into specific contexts has increased. See, for example, Brumen, Cagran and Rixon (2009) on Slovenia, Croatia and the Czech Republic and other chapters in this present volume. This type of research serves to throw light on many of the issues covered in this chapter, in particular the understandings and actual practices of teachers regarding assessment compared with the ideals or the rhetoric to be found at an official level.

The growing interest in EYL assessment by academics and teacher educators such as those above roughly coincided with interest in younger learners from international providers of tests and exams aimed at a large-scale market (see Taylor & Saville, 2002). The aim of providers such as Cambridge English (then Cambridge ESOL), whose YLE tests were launched in May 1997, was to find ‘child-friendly’ yet practicable ways of assessing large numbers of youngsters and assigning them a summative grade that was reliable yet meaningful and informative.

4 Recent General Educational Assessment Movements and Their Influence on EYL Assessment

However, more influential still in some contexts have been movements in general educational assessment which affect the whole curriculum and may thereby also affect what takes place with regard to English. It is worth discussing three recent major movements in mainstream educational assessment at this point since overall educational reforms in some contexts may have been influenced by or directly adopted a version of one of these. In these cases it is likely that the assessment of English as one curricular subject amongst many will be affected by the general reform.

4.1 Standards-Based Assessment

The driving force of standards-based assessment is the attempt to ensure that schools and teachers strive to bring all learners to an acceptable minimum standard of learning (or beyond) and are held accountable for doing so. The No Child Left Behind movement in the USA is a striking early example of this as is the UK National Curriculum with its accompanying standard assessment tasks at the end of primary schooling. In educational systems using standards-based assessment, local or national tests aim to reveal the proportion of pupils attaining specified required

minimum standards either across the curriculum or in specified curricular subjects. To a great extent standards-based assessment removes competitive pressure from learners since it works with thresholds and broad bands of achievement rather than with ranking individuals in a minutely detailed way with respect to the performance of others. However, in the name of accountability, standards-based assessment shifts the pressure of competition on to schools and even on to individual teachers, since the success of institutions is judged by the proportions of their pupils reaching or exceeding the required standard.

Not all would agree that the standards-based assessment movement has been entirely positive. McKay (2005) critiques the support it has lent in countries such as the USA and Australia to managerialism, government control, competition amongst schools in education and the consequent disadvantaging and side-lining of minority groups such as learners for whom English is an additional language who might fare less well on the standard tests. However, standards-based assessment seems to be becoming increasingly influential internationally. In the discussion of the survey of present day EYL policy and practice later on in this chapter we shall see that a number of countries have adopted standards-based assessment across the curriculum and that the assessment of English learning is part of this greater system.

However, controversies may arise when movements for a change in the purpose and/or the format for assessment are introduced or imposed, especially when a significant paradigm shift is involved. For example, Davison (2007, p. 49) reports that transition from norm-referenced to standard-based, school-based assessment (SBA) at secondary school level in Hong Kong caused considerable unease amongst stakeholders. “As an outcome-oriented standards-referenced system, SBA is a significant cultural and attitudinal change, not only for teachers but for the whole school community, including students and parents.”

4.2 Assessment Using Specified Performance Criteria to Determine Levels

Although standards-based and performance-based assessment are often discussed under a single heading, it is important, especially in discussing language learning, to draw some distinctions. Both require detailed specification of what the learner should be able to do but performance-based assessment demands that learners perform in a way that closely reflects or can be directly linked with the use of real-life skills, so that, for example, a primary school numeracy assessment task concerning money might involve real coins and the challenge to children to check their change after a purchase in a shop keeping role-play. On the other hand, in many standards-based assessment systems attainment levels are actually indirectly extrapolated

from test scores and not necessarily directly approached by articulating what a learner ‘can do’ and setting up a challenge which gives them the opportunity to demonstrate it by performing using the required skills and functions. Links may be drawn from test scores to inferred skills and abilities, but this is a controversial area.

Assessment techniques within the performance-based tradition concerning language learning typically involve holistic tasks rather than responses to discrete test items. Role play, challenges involving information gaps and other requirements to simulate real language use as far as is possible are very common. Assessment judgements are made through observation, scrutiny of output such as written work in a required genre and are based on criteria derived from carefully-written performance descriptors. Self-assessment and reflection may be involved and collections of evidence of learning in portfolios may also play a part. *The European Language Portfolio* (ELP) in versions available for both older and younger learners (see <http://www.coe.int/t/dg4/education/elp/>) is a widely used device not only for collecting examples of work but for structuring self-assessment. It is directly linked with the performance descriptors set up by the *Common European Framework of Reference* (CEFR, Council of Europe, 2001).

4.3 Issues with the Common European Framework of Reference in Assessing EYLS

The CEFR (Council of Europe, 2001) is the most prominent example of a framework which can support a performance-based approach to assessment. It has been pointed out, however, (e.g., Jones, 2002) that the descriptors do not in themselves provide direct specifications for tasks which could form part of an assessment. An assessment-deviser would need to bring further detail to its “can do” statements and overall descriptions in order to set up appropriate assessment challenges to elicit a required performance that will demonstrate what the learner can do. There is also the issue that the judgement is not a stark ‘yes’ or ‘no’. There is also scope for judgements of a candidate’s performance concerning ‘how well’ and ‘how much’ they manage within a specified level.

Although the lower levels of the CEFR may seem to offer appropriate levels of language challenge for young children, there are some fundamental problems. As discussed by Hasselgren (2005), we do not currently have a CEFR designed for use with children involving domains that are appropriate for them and which includes skills and topics that are suited to their cognitive and social development and range of interests and experiences. Papp and Salmoura (2009) discuss attempts to calibrate the Cambridge YLE examinations against the CEFR. An additional issue is that, in cases where an A1 or A2 level is specified as the end-point for primary

school learning and the children in fact learn English for a number of years, there is probably a need to subdivide these already modest levels of attainment in order to be able to give sub-grades for levels of attainment arrived at before the final year of learning.

4.4 *Assessment for Learning*

The formative/summative assessment distinction is well known, particularly with regard to assessment within the classroom. Formative assessment is traditionally regarded as informing and shaping what the teacher should do next in class to support learning. However, in its evolved and refined *assessment for learning* (AfL) version (Black, Harrison, Lee, Marshall & Wiliam, 2003; Black & Wiliam, 1998a) the aim is to ensure that learners themselves are enabled to reflect on their current performance and state of understanding and can learn to decide on and plan what they need to do next. AfL provides a platform for the learners to develop an awareness of the goals of their own learning, how close they are to achieving them and the steps that they as individuals can best take to come closer to those goals. AfL, when successfully used, promotes autonomy and self-determination in learning.

Example 1 Emoticons as traffic lights



Common techniques to set up dialogue between teacher and pupils, as exemplified in the UK primary school system, are the use of overt statements of learning objectives – often referred to as WALT- “We Are Learning To” – and self-assessment support for pupils such as the requirement after each piece of work done to indicate the level of confidence they now have in the subject matter and/or skills involved. A ‘traffic light’ system is often used: Green for ‘OK, I understand’, Amber for ‘I’m nearly there’ and Red for when the child still has problems and would like further support. This may be used separately or in conjunction with ‘Emoticon/smiley’ faces as seen in Example 1 which is taken from a model for children shown on a classroom wall.

Example 2 The marking ladder

Friday 26th
October 2012

Marking Ladder

WALT: write a character profile based on a scene in a film

Pupil	Objective	Teacher
✓	I described my character's appearance using adjectives.	✓
✓ <i>but I could of used more</i>	I described how my character acted using powerful verbs.	✓
✓	I used speech to show how my character felt using powerful verbs.	✓
✓	I wrote clearly in complete sentences.	✓

What could I do to improve my writing next time?

I need to improve on my handwriting. ALSO I could of used more powerful verbs.

A super description of Tom in this scene! Really well done ~~the~~. You included all the features on the marking ladder and added your own personality to your writing. You also developed your use of powerful verbs since last time. Next time, I'd really like to see you checking your spellings with a dictionary a little more.

thank you Miss ~~Rebecca~~ WALT ✓
ps. I will start to use a dictionary for spellings.

Brilliant! You ☆

The ‘Marking Ladder’, seen as Example 2, sets up a detailed cooperative dialogue between pupil and teacher. This example is not for English as Foreign Language learning but for a writing challenge set for 9 year old native speakers of English. However, it serves well to demonstrate how the framework provided structures the interchange between learner and teacher:

1. first pupil, then teacher, indicates in their ‘own’ column if they feel that the WALT objectives have been met
2. pupil offers his/her own suggestion about the next step for improvement
3. teacher responds and pupil adds follow-up comments if they wish.

Assessment for learning is an area in which culturally-influenced views of childhood and children’s capacities, found in different YLs contexts, may play their part with regard to the reception of the approach by parents and professionals. Butler and Lee write, for example, of the challenges faced in convincing some colleagues that children are capable of reflection and self-assessment.

There are relatively few empirical investigations of self-assessment among young learners at the pre-elementary and elementary school levels. This may in part be due to the widespread notion that children are not capable of accurately self-evaluating their own performance or self-regulating their own learning. (Butler & Lee, 2010, p. 8)

So far we have considered recent developments in assessment both in language learning and in general education at primary level with a view to their potential influence on current actual practice in the teaching of YLs of English. Prominent amongst these developments have been Standards-based assessment, the interest in alternative methods of gathering evidence on attainment and growing interest in formative assessment, in particular *assessment for learning*. The rest of this chapter reports on evidence about current actual practice in YLs assessment drawn from two international surveys.

5 Analysis of Data on the Current State of the Art in YLs Assessment

The research questions concerning assessment to be answered by the surveys reported on here were similar in their interest in actual current practice although, because the informants differed, the scope and detail naturally also differ. The two surveys are as follows:

5.1 The Cambridge English Survey of Teachers’ Practices in Assessment

This survey (Papp, Chambers, Galaczi & Howden, 2011), which was questionnaire-based, covered the area of classroom teaching and assessment in great detail, with responses concerning their own practices from numerous individuals directly involved

professionally with Young Learners of English. The results of this survey are not publicly available in their entirety, although they will be drawn on in a future volume on assessing Young Learners in the Cambridge University Press *Studies in Language Testing* series (Papp and Rixon, forthcoming). Many thanks therefore go to Cambridge English for permission to publish summaries of key sections here. Because of the unavailability of the original document, page references will not be given.

The research interest was on individual perceptions as well as trends in the area of English language assessment. Much use was made of open response questions to which individual teachers gave detailed answers.

Respondents worked in private as well as state institutions, a number of them working in both. In all, 726 valid responses were returned from 55 different countries, the majority of respondents being from Greece, Italy, Mexico, Romania and Spain. Of the total sample, about 300 respondents worked mainly with learners in the 6–11 year old age range which is the subject of the present chapter. The rest worked with secondary school-aged learners. See Appendix A for the list of countries covered.

5.2 The British Council Survey of Policy and Practice in Primary ELT

The British Council survey (Rixon, 2013) was undertaken as a follow-up to an earlier survey on the same topic already quoted above (Rixon, 2000). The research scope of this survey was broader than that of the Cambridge survey in that it took in overall developments in policy and practice such as starting ages for English, availability of pre-school English, numbers of hours of English per year and over a whole primary school career, teaching materials and teacher qualifications and eligibility as well as relations between the public and private sectors. Because of the growing importance of assessment in Young Learners teaching, a special section of the survey questionnaire was devoted to policies regarding assessment.

Returns were mostly via an on-line questionnaire. The purpose of the survey was to collect data on policy and officially-supported practices in as many countries and regions as possible worldwide. In contrast with the Cambridge survey, this was a global ‘facts and figures’ exercise rather than an investigation into individual views and practices. It was thus felt appropriate not to make use of the questionnaire with a massive number of individuals but to identify one, or at most two, well-informed sources for each context. Authoritative informants on local policy and practice in a country or region were identified via the local British Council Offices. Responses were received from 64 separate countries or regions. See Appendix B for the list of contexts covered.

In many countries and regions, thanks to an increase in on-line information, much of the statistical data requested could be obtained and checked by reference to official websites. In cases where the answers were based not on official data but on an estimate or on the respondent’s personal experience, the respondents were asked

to state the degree of confidence with which they were answering. It is thus claimed that the data reported are of as good quality and as reliable as possible and, in cases where they are not independently verified, this fact is made transparent to the reader.

5.3 Findings from the Surveys

5.3.1 An Update on Teacher Preparation and Supply

There is evidence in the 2013 British Council survey that the tensions noted in 2000 between enthusiasm for innovation and less concern for practical provision have continued. Teacher supply and/or quality was judged adequate in only 17 (27 %) of contexts. In spite of the difficulties in teacher supply, the most frequently-reported recent policy change was the lowering of the age at which English was to be taught compulsorily in the primary school. Some verbatim comments from respondents illustrate issues encountered with keeping up or catching up with current demand for adequately trained teachers, with, for example, both the Taiwanese and the Israeli respondents complaining that teachers of English to primary school children often needed to be drawn from teachers specialising in other subjects.

However, the survey also revealed some cases in which, in spite of continued enthusiasm for lowering the age at which English could be begun, more realistic attitudes were evident.

There was a change introduced in the Regulations of the Cabinet of Ministers as to the age of starting the 1st foreign language – moving it to Year 1 (age 7), but it has been decided to wait with this change for a couple of years due to lack of funding (Latvia). (Rixon, 2013, p. 148)

In addition, there were cases where, in spite of problems reported at the time of response, planning was in place and attempts to improve teacher preparation for the future were evident: For example, in France, teaching personnel from numerous different backgrounds were still being used at the time of response. This had been an issue highlighted for France even as far back as the earlier, 2000, survey. However, the comment in the more recent survey showed that steps had been taken to ensure the supply of better qualified teachers in the future.

This is temporary as it is now compulsory for all new teachers to graduate from teacher training college (IUFM) with the required level of the foreign language. They will receive a certificate called CLES (Certificat de Langue de l'Enseignement Supérieur). This certificate certifies language competence only not methodology (France). (Rixon, 2013, p. 108)

It is notable that, here the emphasis is on the language levels of the graduating teachers rather than on the need also to cater for their preparation in appropriate language teaching methodology. However, when resources are stretched this seems a pragmatic if not ideal priority. It is one which remains widespread across other contexts. In a climate in which even language teaching methodology is rarely the subject of teacher preparation, one has then to ask how likely it is for new recruits

to receive specialist training in appropriate reasons for, and means of assessment of, children’s language learning.

One contention of this chapter is that the degree to which teachers are confident all-round ELT professionals, in ways which go beyond their own language proficiency, has huge implications for the nature and quality of language learning assessment. If elementary school teachers in many contexts are still learning to become fully skilled teachers of English, they might reasonably be expected still to be finding their way as implementers, informed critics or devisers of English language assessment approaches.

5.3.2 Teachers’ Growing Understanding of Assessment of Young Learners’ English

The Rea-Dickins and Rixon survey (1999) cited above showed teachers implementing class tests in a way that did not chime with their stated teaching priorities: 100 % of the sample of 122 teachers from nearly 20 countries stated that their main aims were to promote listening and speaking but none of them used class tests involving these skills. The Cambridge English survey of 12 years later involved more countries and teachers who came from private school as well as state school teaching backgrounds (although many had more than one job and some taught in both types of institution). Their self-reports concerning knowledge about and use of different assessment formats suggested that there was an awareness of a much broader variety of possibilities for assessment and of the different purposes it might serve.

Amongst the nearly 300 teachers of 6–11 year olds who responded to the Cambridge survey, the following types of assessment were selected as significant and actually used. These are listed in rank order according to the number of responses for each one:

1. Tests produced by the class teacher
2. Tests given in the textbook used in class
3. Collection of students’ work in a file or portfolio
4. Observation and written description of learner performance
5. Standardised tests and examinations
6. Self-assessment
7. Peer-assessment

The picture presented by the data from these teachers is of a good spread of actually-used assessment types per teacher. Out of just under 300 teachers, the numbers choosing these top seven assessment types was closely ranged between around 200 and 125. The top two choices of teacher-produced or textbook-supplied tests – similar perhaps to the written classroom tests used by the teachers in the Rea-Dickins and Rixon survey of 1999 – were made by approximately 200 respondents each, with the rest, apart from peer-assessment, at nearly the same level. Peer-assessment received the lowest number of selections, being chosen by approximately 125.

As might be expected, standardised tests were mostly reported as being used yearly for end of course summative assessment. Tests from textbooks and teacher-made tests, collection of student work and self-assessment tended to be used monthly, while written description of learner performance was usually provided once a year (as in school report-writing). No information is available on whether those once-a-year descriptions were in fact based on regular record keeping based on observation.

As is often the case, it is the individual comments that are the most suggestive of teachers' understanding of the topics covered in the questionnaire. A number claimed that the function of tests that they found most valuable was that of informing them of the success or even the quality of their own teaching.

There was little indication in responses to this survey that self-assessment had yet moved into a fully-developed *assessment for learning* mode such as is described in the first section of this chapter.

Portfolios were sometimes used as part of peer-assessment or as a stimulus for self-assessment. They were not usually graded. In many cases they were used as 'self-evident' proof of achievement to stakeholders such as parents for them to look at and judge for themselves.

It is important to remember that the sample for the Cambridge English survey was drawn from teachers who already had had contacts with one major international provider of tests and examinations. Many of the respondents also self-reported as having prepared children for international tests and examinations from other providers. They might therefore be expected to be amongst the best informed in the profession.

5.3.3 National Policies and Assessment

The British Council survey (Rixon, 2013) supports the discussion of assessment from the more top-down perspective of national or regional policy. Officially-endorsed assessment principles and skills may or may not already have percolated down to the classroom level in a given context but first signs of *coming of age* at a national or regional level may also be traced when officialdom puts in place an assessment policy that is likely to add to clarity about the standards expected or is presented as having the intention of bringing about a positive impact on classroom teaching.

The following themes regarding assessment explored by the survey will be discussed:

1. Standards-setting and the growing role of the *CEFR*
2. Assessment as an official requirement in EYL teaching in primary school
3. Means of assessing if standards are reached
4. Consequences and lack of consequences of assessment
5. The role of assessment in facilitating transitions between school levels.

5.3.4 Standards-Setting and the Growing Role of the *CEFR*

In a substantial number of the contexts (33 %) there were reports of innovations regarding standards-setting as a major recent policy change. Other places had already had standards in place for a longer time. Where standards for English and other languages taught had been recently set, those standards were described as based upon the *CEFR*. From 20 contexts mention is made of A1 or A2 at *CEFR* as the required level of attainment at the end of primary school. From Croatia comes a detailed report of foreign language standards being set as a part of a general educational reform and using the *CEFR* as a reference point.

The Croatian Education System has undergone many changes in the last decade in the Government’s attempt to align it to European trends and policies. In the primary sector, the most significant changes have been the introduction of the competence-based, student-centred Croatian National Education Standards which were introduced in all schools in 2006/07, the Primary Education Syllabus in 2006, the new Act on Education in Primary and Secondary school in 2008 and the new National Framework Curriculum for Pre-School, Primary and Secondary Education in 2010. ...The Primary Education Syllabus defines the levels of English proficiency according to *CEFR* levels depending on the number of years the language is taught in Croatia. (Croatia) (Rixon, 2013, p. 88)

It should be noted that the *Common European Framework of Reference* (2001) is now also widely used in contexts that are outside Europe, for example, Colombia, Mexico, Georgia and Kazakhstan.

The widespread setting of standards seems to be the most significant change since the last global survey of EYL policy for the British Council took place at the end of the twentieth century. However, although standards may be set, they are not necessarily always checked through formal assessment. For example, the Czech Republic, Greece and Lithuania which set *CEFR* A1 as their target standard at the end of primary school do not require formal assessment. There were numerous contexts, including Indonesia, Pakistan and Saudi Arabia in which no target standard is yet in place and formal assessment is not required.

5.3.5 Assessment as an Official Requirement in YL Teaching in Primary School

At the beginning of an innovation involving the teaching of a foreign language to young children, it is common for there to be a ‘honeymoon’ period in which no assessment is built into the project. There are often good reasons for not insisting on assessment, or on formal summative assessment at least. Firstly, the project will need time to ‘settle in’ before any valid and reliable results may be expected. Secondly, as discussed above, teachers may have enough to cope with in the early stages, just concerning the teaching, without the additional burden of working with assessment in an area in which they are not yet confident. Thirdly, there may be professional or ideological views held by the initiators of a project, according to which children, particularly very young children, should not be disturbed in their

learning by assessment. Kubanek-German writes eloquently of the situation with regard to Germany when primary school teaching of foreign languages was still in its early stages:

When the new primary programmes started in the early 90s, there was a marked unwillingness among teachers and curriculum planners to administer tests or describe progress in a systematic fashion. The principle of child-orientation (holistic approach, integrative approach, use of stories, avoidance of anxiety, fostering motivation and intercultural openness) seemed to exclude formal testing. (Kubanek-German, 2000, p. 65)

However, in many contexts, once the first cohorts to start English in the early years of schooling began to reach the end of primary school, attitudes and policies often changed. Assessment of attainment at the end of primary school became required in Germany in the later 1990s in the same way as it had in the early 1990s in France, another context in which in the early years of the innovation no assessment had been required. A similar change took place in Italy in 1997 when a section was added to the school report form concerning the child's attainments and progress in learning a foreign language (in which English was the most popular choice). The present author remembers attending a number of in-service courses in France, Germany and Italy designed to support teachers in their new assessment responsibilities.

In the British Council survey, there were reports from 11 out of 64 (17 %) of the locations surveyed that there had been recent policy changes concerning the introduction of assessment. In addition to these 11 cases, we should not forget that in a number of other countries, such as those mentioned immediately above, assessment of English had been already well established some years previously. A later question in the survey allowed for respondents to make comments and explain more about how assessment was carried out.

5.3.6 Means of End of Primary School Assessment

Before discussing means by which end of primary school assessment is carried out, it should be remembered, as noted above, that in a large number of contexts (28; 44 % of the sample) it was stated that there was no requirement for formal assessment of English language learning at the end of primary school. This involves a number of contexts in which standards have been set but there are no formal means by which their attainment is ascertained.

Where assessment at the end of primary school takes place, this may be by formal testing but it may also be by a means devised within the school or following a framework supplied from outside but implemented by teachers. France provides an example of a recently introduced highly systematic application of this latter practice:

In France, there is continuous assessment from Year 3 to Year 5. At the end of year 5, teachers complete an evaluation (*Palier 2 CM2 La pratique d'une langue vivante étrangère*) which covers five skills areas: oral interaction, understanding, individual speaking with no interaction, e.g. reproducing a model, a song, a rhyme, a phrase, reading aloud, giving a short presentation e.g. saying who you are and what you like. (France). (Rixon, 2013, p. 107)

Other contexts favour the more conservative means of formal testing which may also be linked with official evaluation of school success. This is usually as a result of English, as one curricular subject among many, being included in a wider educational policy. Russia and Bahrein, for example, were reported as having instituted new systems of formal assessment across the curriculum at the end of primary schooling. The stated purpose for this in both cases was in order to monitor and evaluate school performance.

In Taiwan, assessment specific to English is being implemented at a local level with, it seems, a diagnostic purpose as well as a school evaluation purpose.

Cities and counties in Taiwan are now developing and administering their own English proficiency tests at the primary level. The purpose is to assess the effectiveness of English instruction and to identify those in need of remedial teaching. Assessment is mid-term and final, starting from the third grade. (Taiwan) (Rixon, 2013, p. 224)

Even when formal assessment by an official test is not required by regulation, there seem in some contexts to be strong social pressures to have objective test-based measures in place. A frank comment on the situation in Finland suggests how different stakeholders may influence what actually happens on the ground, resulting in the widespread use of additional unofficial tests in English amongst other school subjects. An unofficial league-table of schools, familiar in the UK with regard to the press reporting of examination results, seems to be developing in Finland:

Many (we don't know how many) primary schools use a voluntary 'national' test of English (or some other school subject) designed by the English teachers' association of Finland (or another such association depending on the subject) at the end of primary school, to guide their final grading of the students, to get some information for themselves about how they are doing against the average of the other schools that have opted to take the same test, and so on. This is quite unofficial and varied as to how the teachers and schools use the information from those tests that are not really standardised in the strict meaning of the word. Recent information indicates that some school rectors and municipal education authorities insist on the teachers/schools using these tests so that they would know how well their school(s) are doing against the other schools. This violates the stated purpose of these tests but seems to be happening anyway, at least in some municipalities and schools. (Finland) (Rixon, 2013, p. 106)

By contrast, the recently introduced innovations regarding primary English teaching in Cyprus have included assessment through the use of portfolios.

A New National Curriculum, part of the education reform happening in Cyprus, has been implemented in September 2011. This introduces English from pre-primary, emphasises the role of portfolio assessment and introduces content and language integrated learning (CLIL). (Cyprus) (Rixon, 2013, p. 91)

5.3.7 Consequences and Lack of Consequences from End of Primary School Assessment

In contexts in which assessment takes place we have seen above that there are different purposes, including the wish to monitor and evaluate school performance. It was not clear in many of these cases how draconian the consequences of failure to reach adequate standards would be.

Clearly, political and economic conditions in different contexts make a difference to what is at stake and therefore to the type of assessment that is in place. In a few contexts, results in English might affect the category of secondary school to which pupils might go. However, there was a group of countries included in the survey in which assessment in English at the end of primary school was very high stakes. These countries were ones in which English had a status as an official rather than a foreign language and had been established for many years as the medium of education as well as a school subject. In these cases more traditional end-of-primary-schooling examinations were used and had been in force for a long period of time. They could form a very important part of the decision-making process concerning a child's educational future. In countries where educational opportunities are hard to come by they could even help to determine whether a pupil might go to secondary school at all. This was reported in Bangladesh, Namibia and Zambia, for example.

It was intriguing that, in some contexts, formal end-of-primary-school assessment in English was reported as taking place but that it was also reported that this assessment would have no consequences with regard to secondary school entry (or for any other purpose).

5.3.8 Transitions Between School Levels

Ensuring continuity and coherence in learning between levels of schooling is both a well-known and a long-standing problem area in the field of curriculum planning involving an early start for language learning. It was most famously signalled as long ago as the 1970s (Burstall, Jamieson, Cohen & Hargreaves, 1974) with regard to learners of French leaving primary school in England and Wales and moving to secondary schools. A common experience was that their achievements in learning French tended to be belittled and if there were differences in levels of French among the children the whole class was often made to start again from zero. Since then, this phenomenon has been observed in many educational contexts. It has often been identified as one source of the failure of early start programmes in foreign language learning to yield the hoped-for results by the end of secondary schooling. See, for example, Hunt, Barnes, Powell and Martin (2008).

Often the disjunction between school levels is partly a consequence of teaching cultures in which primary and secondary teachers have little in common and few chances to be in contact with one another. This distancing may be exacerbated in societies in which there is a stark difference in professional and/or social status between primary and secondary school teachers. Assessment alone is therefore not to be seen as the solution but if used strategically it can go some way towards improving matters. For example, giving secondary colleagues a realistic and in-depth view of what children have actually achieved could help to break down the prejudice that little is achieved at primary school. The European Language Portfolio is sometimes used for this purpose, as was reported in the Pri-Sec-Co project (Education, Audiovisual & Culture Executive Agency, 2008), a piece of work funded by a European Union Comenius grant specifically to investigate transition in school systems in some European countries as regards language learning.

Table 1 Responses in the British Council Survey concerning transition from primary school to the next level of education

	Always	Often	Quite often	Sometimes	Rarely	Never	I don't know/no info
Teachers from the two levels of schooling meet to discuss the transition	2	1	2	8	14	24	13
Information on children's levels from externally provided formal testing at the end of Primary School is passed to the new school	7	0	3	4	8	25	17
Information on children's levels from school-based assessment is passed to the new school	14	4	3	4	5	21	13

The British Council survey (Rixon, 2013, pp. 39–40) aimed to investigate ways in which assessment data is used or fails to be used in order to promote coordination between primary/elementary school and secondary school level language learning. Table 1 shows the numbers of responses of each type to the three questions below regarding assessment and transition.

1. Do primary and secondary school English teachers meet to discuss pupils moving to secondary school?
2. Is school-based assessment information passed to the next school?
3. Is information from externally provided formal testing passed to the next school?

The three questions covered three levels of possible formality with which information might be passed from one school to the next: It seems from the results of this part of the survey that the opportunity for making good use of information on children's attainments in English whether through assessment results or informal data was often missed (yet again).

6 Limitations

The data in this chapter come mostly from surveys in which summaries of prevailing practices are given by experts and experienced teachers and there has been no opportunity for analysis or discussion of materials used in assessment or of the experiences and understandings of ordinary teachers and their pupils. Although some practices may be shared or imitated across national boundaries and instruments such as the CEFR may be influential, it does not make sense to seek for trends on an international scale.

7 Implications for Practice

As pointed out above, a chapter based mainly on survey data cannot make detailed recommendations for assessment practice in a given context. However, from the discussion it may be seen that the signs that EYL initiatives are on their way to *coming of age* with regard to assessment are rather few. As with much in the field of the teaching of languages to young learners, statements of the ideal in good practice in the learning/assessment bond often outstrip the reality. It was to be expected that, given the global nature of the two main surveys quoted, there would be a wide range of practice found, much of which would be affected by the beliefs and traditions of local teaching and assessment cultures. However, in some contexts, local authorities and experts are introducing new approaches which may require a considerable revision of mind-set on the part of teachers and public alike. The research reported on in this chapter also suggests a wide range of technical assessment expertise, from contexts in which assessment practices may be haphazard or occasionally diametrically at odds with the stated pedagogic aims of the teaching programme to those in which assessment seems to be well understood at both an official and a classroom practitioner level.

The following key points seem to have emerged:

1. Teachers who in many contexts are still not yet fully bedded in as language teachers may be expected to lag a little in classroom language assessment practices. More and higher quality pre-and in-service teacher education on the topic is needed.
2. There is a notable increase in the setting of target levels but there is not always provision of means to ascertain whether those levels are in fact obtained. There is an urgent need for assessment instruments to be developed that are a good match with the targets.
3. Assessment instruments provided by specialists for regional/national use have increased since 1999/2000 in terms of quantity. This is a positive development provided that these instruments in fact match with stated aims.
4. Sharing of assessment information at school transition remains patchy. This is an area in which all but a few countries need to take serious stock and devise means to improve continuity and coherence.

8 Need for Future Research

There seems to be much that could be learned now and in the near future from detailed qualitative accounts of the development in assessment of children's English language learning in some of the contexts from which the information in this chapter was collected. It is to be hoped that publication of close-up, localised, studies of assessment practices with young learners will be on the increase.

Appendices

Appendix A: Countries from Which a Minimum of Five Responses to the Cambridge Survey Were Obtained

Argentina	Hong Kong	Russia
Brazil	Italy	South Korea
Bulgaria	Japan	Spain
Chile	Macedonia	Sri Lanka
China	Malaysia	Switzerland
Croatia	Mexico	Turkey
Cyprus	Peru	Uruguay
France	Poland	Vietnam
Germany	Portugal	
Greece	Romania	

Appendix B: Countries and Regions from Which Responses to the British Council Survey Were Obtained

Algeria	India: Goa	Qatar
Argentina	India: South India	Romania
Armenia	India: Tamil Nadu	Russia
Azerbaijan	Indonesia	Saudi Arabia
Bahrain	Israel	Senegal
Bangladesh	Italy	Serbia
Brazil	Japan	Sierra Leone
Cameroon	Jordan	South Africa
China	Kosovo	South Korea
China: Hong Kong	Latvia	Spain
Colombia	Lithuania	Sri Lanka
Croatia	Mexico	Sweden
Cyprus	Montenegro	Taiwan
Czech Republic	Morocco	Turkey
Denmark	Namibia	Uganda
Egypt	North Cyprus	United Arab Emirates
Finland	Pakistan	Uzbekistan
France	Palestine	Venezuela
Georgia	Peru	Yemen
Germany	Poland	Zambia
Greece	Portugal	Zimbabwe

References

- Andrews, S. (2004). Washback and curriculum innovation. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research* (pp. 37–52). Mahwah, NJ: Laurence Erlbaum and Associates.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Black, P., & Wiliam, D. (1998b). *Inside the black box: Raising standards through classroom assessment*. London: King's College.
- 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.
- Burstall, C., Jamieson, M., Cohen, S., & Hargreaves, M. (1974). *Primary French in the balance*. Slough: NFER Publishing Company.
- Butler, G., & Lee, Y. (2010). The Effects of self-assessment among Young Learners of English. *Language Testing*, 27(1), 5–31.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Davison, C. (2007). Views from the chalkface: English language school-based assessment in Hong Kong. *Language Assessment Quarterly*, 4(1), 37–68.
- Education, Audiovisual & Culture Executive Agency. (2008). Pri-sec-co. Primary and secondary continuity in foreign language teaching. Project no. 134029-LLP-1-2007-1-DE-COMENIUS-CMP. Retrieved from http://eacea.ec.europa.eu/lfp/project_reports/documents/comenius/all_com_mp_134029_prisecco.pdf
- Enever, J., & Moon, J. (2009). New global contexts for teaching primary ELT: Change and challenge. In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 5–20). Reading: Garnet.
- European Languages Portfolio. Retrieved from <http://www.coe.int/t/dg4/education/elp/>
- Graddol, D. (2006). *English next*. London, UK: British Council.
- Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing*, 22(3), 337–354.
- Henry, A. K., Bettinger, E., & Braun, M. K. (2006). *Improving education through assessment, innovation, and evaluation*. Cambridge, MA: American Academy of Arts and Sciences.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.
- Hunt, M. J., Barnes, A., Powell, B., & Martin, C. (2008). Moving on: The challenges for foreign language learning on transition from primary to secondary school. *Teaching and Teacher Education*, 24(4), 915–926.
- Ioannou-Georgiou, S., & Pavlou, P. (2003). *Assessing young learners*. Oxford: Oxford University Press.
- Jones, N. (2002). Relating the ALTE framework to the common European framework of reference. In J. C. Alderson (Ed.), *Case studies on the use of the common European framework of reference* (pp. 167–183). Strasbourg: Council of Europe Publishing.
- Kubaneck-German, A. (2000). Early language programmes in Germany. In M. Nikolov & H. Curtain (Eds.), *An early start: Young learners and modern languages in Europe and beyond* (pp. 59–70). Strasbourg: Council of Europe Publishing.
- McKay, P. (2005). Research into the assessment of school-age language learners. *Annual Review of Applied Linguistics*, 25, 243–263.
- Papp, S., Chambers, L., Galaczi, E., & Howden, D. (2011). *Results of Cambridge ESOL 2010 survey on YL assessment*. University of Cambridge ESOL Examinations: Cambridge ESOL internal document VR1310.

- Papp, S., & Salmoura, A. (2009). An exploratory study into linking young learners' examinations to the CEFR. *Research Notes*, 37, 15–22.
- Rea-Dickins, P. (Ed.). (2000). Assessing young language learners [special issue]. *Language Testing*, 17(2).
- Rea-Dickins, P., & Rixon, S. (1997). The assessment of young learners of English as a foreign language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Language testing and assessment, Vol. 7, pp. 151–161). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Rea-Dickins, P., & Rixon, S. (1999). Assessment of young learners: Reasons and means. In S. Rixon (Ed.), *Young learners of English: Some research perspectives* (pp. 89–101). Harlow: Pearson Education.
- Rea-Dickins, P., & Scott, C. (2007). Washback from language tests on teaching, learning and policy: Evidence from diverse settings. *Investigating washback in language testing and assessment [Special Issue]. Assessment in Education: Principles, Policy and Practice*, 14(1), 1–7.
- Rixon, S. (2000). Collecting eagle's eye data on the teaching of English to young learners: The British Council overview. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 153–167). Pécs: University of Pécs Press.
- Rixon, S. (2013). *British Council survey of policy and practice in primary English Language Teaching worldwide*. Retrieved from <http://www.teachingenglish.org.uk/article/british-council-survey-policy-practice-primary-english-language-teaching-worldwide>
- Taylor, L., & Saville, N. (2002). Developing English language tests for young learners. *Research Notes*, 7, 2–5.

The “Global Scale of English Learning Objectives for Young Learners”: A CEFR-Based Inventory of Descriptors

Veronica Benigno and John de Jong

Abstract This chapter presents an ongoing project to create the “Global Scale of English Learning Objectives for Young Learners” – CEFR-based functional descriptors ranging from below A1 to high B1 which are tailored to the linguistic and communicative needs of young learners aged 6–14. Building on the CEFR principles, a first set of 120 learning objectives was developed by drawing on a number of ELT sources such as ministry curricula and textbooks. The learning objectives were then assigned a level of difficulty in relation to the CEFR and the Global Scale of English and calibrated by a team of psychometricians using the Rasch model. The objectives were created and validated with the help of thousands of teachers, ELT authors, and language experts worldwide – with the aim to provide a framework to guide learning, teaching, and assessment practice at primary and lower-secondary levels.

Keywords Young learners • Descriptors • Assessment • Teaching • Learning objectives • Can Do Statements • Rating • Scaling • CEFR (Common European Framework of Reference for Languages) • GSE (Global Scale of English) Learning Objectives for Young Learners

1 Introduction

The *Common European Framework of References for Languages (CEFR; Council of Europe, 2001)* was compiled with an adult and young adult audience in mind. Consequently, the majority of descriptors refer to communicative acts performed by learners who are likely to use the foreign language in the real world. The CEFR is

V. Benigno (✉)
Pearson English, Pearson Education, London, UK
e-mail: veronica.benigno@pearson.com

J. de Jong
Pearson Assessment Centre, Pearson Education Inc., Iowa City, IA, USA
Amsterdam VU University, Amsterdam, Netherlands
e-mail: john.dejong@pearson.com

therefore less appropriate for describing proficiency of young learners (YL, primary, and lower secondary learners), and particularly of the youngest ones whose life experience is substantially different from that of adults.

In this chapter we discuss an ongoing project at Pearson English which aims to develop a set of functional descriptors for young learners: the “Global Scale of English Learning Objectives for Young Learners” (Pearson, 2015b; here also referred to as “descriptors” or “learning objectives”). These CEFR-based “Can Do” statements cover the levels from below A1 to high B1 and are tailored to motivations and needs of young learners aged 6–14, a period during which they are still developing linguistic and cognitive skills in their own mother tongue. Level B2 and higher are not taken into account because they assume more adult skills. The CEFR was used as a reference guide to identify valid theoretical and methodological principles for the development and the scaling of the new descriptors.

We believe this work represents a contribution to the ongoing debate on what young learners can do and what instruments can be used to assess their performance. Setting standards requires us to define what learners should be able to do with the language at a certain level of proficiency and how to observe proficiency gains in relation to a defined scale. Standard setting does not imply a prescriptive pedagogy but allows for comparability between curricula based on a definition of extraneous, i.e., non-school, functional learning goals. If standards refer to a common framework they will allow the implementation of a transparent link between content development, teaching, and assessment.

Teaching English to Young Learners (TEYL) has recently received much attention. Under the impact of globalization, the last few decades have seen an increasing tendency to introduce English in primary school curricula around the world, particularly in Europe (Nikolov & Mihaljević Djigunović, 2006; Nikolov, 2016). Nowadays, millions of primary age children learn English in response to parents’ expectations and supported by educational policy makers. There has been an increase not only in the number of young learners and their teachers, but also in the volume of documents about and for young learners: language policy documents, teachers’ handbooks, teaching materials, empirical studies, conference reports and proceedings, and academic publications (Nikolov & Mihaljević Djigunović, 2011). Early language learning policies have been promoted by European institutions since the 1990s (Speitz, 2012). According to the European Commission, early language learning yields a positive impact in terms of education and cross-cultural communication:

Starting to learn a second/foreign language early can help shape children’s overall progress while they are in a highly dynamic and receptive developmental stage in their lives. Starting early also means that learning can take place over a longer period, which may support the achievement of more permanent skills. When the young brain learns languages, it tends to develop an enhanced capacity to learn languages throughout life. (European Commission, 2011, p. 7)

Support of intercultural education is claimed to be among the benefits of early language learning: “raising awareness of language diversity supports intercultural

awareness and helps to convey societal values such as openness to diversity and respect” (European Commission, 2011, p. 7).

It is generally believed that early foreign language (FL) introduction provides substantial benefit to both individuals (in terms of linguistic development, social status, and opportunities) and governments (as a symbol of prestige and economic drive). However, some concerns have been raised about the dangers of inadequate preparation and limited knowledge about who young learners are, how they develop, and what they need. This has led some researchers to argue against the validity of “the earlier the better” hypothesis. Among the most common arguments against this principle are: (a) learning is not exclusively determined by age but also by many other factors, e.g., the effectiveness of teaching; and (b) younger learners have an imprecise mastery of their L1 and poorer cognitive skills in comparison to older learners. Studies on the age factor (e.g., Lightbown & Spada, 2008) have shown that, at least in the early stages of second language development, older learners progress faster than younger ones, questioning the benefit of the early introduction of an FL in the curriculum. Other studies (e.g., Singleton, 1989), however, have argued that early language learning involves implicit learning and leads to higher proficiency in the long run. There is indeed some evidence to support the hypothesis that those who begin learning a second language in childhood in the long run generally achieve higher levels of proficiency than those who begin in later life (Singleton, 1989, p. 137), whereas there is no actual counter evidence to disprove the hypothesis.

It is worth highlighting that “the earlier the better” principle is mainly questioned in FL contexts, whereas several studies on bilingual acquisition show great benefits for children who learn two linguistic systems simultaneously (Cummins, 2001).

Another major concern among TEYL educators and stakeholders is the lack of globally (or widely) accepted guidelines to serve as a reference for standard setting. Although there is some consensus on who young learners are and how their proficiency develops at different cognitive stages, there seems to be a lack of consistency in practices around the world. According to Inbar-Lourie and Shohamy (2009, pp. 93–94, cited in Nikolov & Szabó, 2012, p. 348), early programmes range from awareness raising to language focus programmes and from content-based curricula to immersion. It appears to be particularly problematic to develop a global assessment which fits the richness of content aimed at young learners of different ages and with different learning needs worldwide. While the CEFR has become accepted as the reference for teaching and assessment of adults in Europe, different language institutions have produced different, and sometimes conflicting, interpretations of what the different levels mean. Moreover, there is no single document establishing a common standard for younger learners, but rather several stand-alone projects that try to align content to the CEFR or to national guidelines (e.g., Hasselgren, Kaledaité, Maldonado-Martin, & Pizorn, 2011). Pearson’s decision to develop a CEFR-based inventory of age-appropriate functional descriptors was motivated by the awareness of (1) the lack of consensus on standards for young learners and (2) the consequent need for a more transparent link between instructional and assessment materials, on the one hand, and teaching practices, on the other.

Although it is not the purpose of the present study to provide a detailed picture of all aspects of TEYL, we will briefly touch upon some of the main issues related to its implementation (see Nikolov & Curtain, 2000 for further details). In the first section of this chapter we present the heterogeneous and multifaceted reality of TEYL and discuss the need for standardisation. We outline the linguistic, affective and cognitive needs which characterize young learners. This brief overview is intended to provide the reader with some background on the current situation of TEYL and to support our arguments for the need of a set of descriptors for young learners. In the second section we discuss the limitations of the CEFR as a tool to assess young learners. We also describe the reporting scale used at Pearson, the Global Scale of English -henceforth GSE- (Pearson, 2015a), which is aligned to the CEFR. Then, we move to the main focus of our paper and explain how we developed the learning objectives by extending the CEFR functional descriptors and how we adapted them to the specific needs of a younger audience. Our descriptor set is intended to guide content development at primary and lower-secondary levels and to serve as a framework for assessment for EFL learners aged 6–14 and on the CEFR levels below A1 to high B1. The last section discusses the contribution of our paper to the research on young learners and briefly mentions some issues related to assessment.

2 The Heterogeneous Reality of TEYL and the Characteristics of Young Learners

2.1 The Need for Standardisation in TEYL

One of the major concerns related to TEYL is the absence of globally agreed and applied standards for measuring and comparing the quality of teaching and assessment programmes. Nikolov and Szabó (2012) mention a few initiatives aimed at adapting the CEFR to young learners' needs and examinations, along with their many challenges. According to Hasselgren (2005), the wide diffusion of the European Language Portfolio checklists developed by the Council of Europe (2014) for young learners has shown the impact of the CEFR on primary education. However, a glimpse into the different local realities around the world reveals a chaotic picture. Consider the obvious variety of foreign language programmes across Europe in terms of starting age, hours of instruction, teachers' proficiency in the foreign language, teachers' knowledge of TEYL, and support available to them (McKay, 2006; Nikolov & Curtain, 2000). Although there may be arguments for using different methods, approaches, and practices, a problem arises when no or little effort is made to work toward a common goal. Because of the absence of agreed standards, even within national education systems, existing learning, teaching and assessment resources are extremely diverse, leading to a lack of connectedness and resulting inefficacy. The implementation of a standard is therefore needed

to describe what learners are expected to know at different levels of schooling. At the national level, common learning goals should be clearly defined and students’ gains at each transition should be accounted for in order to guarantee continuity between different school grades. At the international level, standardisation should be promoted so as to increase the efficacy of teaching programmes in order to meet the requirements from increasing international mobility of learners and to allow for the comparison of educational systems.

2.2 Who Are Young Language Learners?

According to McKay (2006), young language learners are those who are learning a foreign or second language and who are doing so during the first 6 or 7 years of formal schooling. In our work we extend the definition to cover the age range from 6 to 14, the age at which learners are expected to have attained cognitive maturity. In our current definition, the pre-primary segment is excluded and age ranges are not differentiated. In the future, however, we may find it appropriate to split learners into three groups:

1. Entry years age, usually 5- or 6-year-olds: teaching often emphasizes oral skills and sometimes also focuses on literacy skills in the children’s first and foreign language
2. Lower primary age, 7–9: approach to teaching tends to be communicative with little focus on form
3. Upper primary/lower secondary age, 10–14: teaching becomes more formal and analytical.

In order to develop a set of learning objectives for young learners, a number of considerations have been taken into account.

- Young learners are expected to learn a new linguistic and conceptual system before they have a firm grasp of their own mother tongue. McKay (2006) points out that, in contrast to their native peers who learn literacy with well-developed oral skills, non-native speaker children may bring their L1 literacy background but with little or no oral knowledge of the foreign language. Knowledge of L1 literacy can facilitate or hinder learning the foreign language: whilst it helps learners handle writing and reading in the new language, a different script may indeed represent a disadvantage. In order to favour the activation of the same mechanisms that occur when learning one’s mother tongue, EFL programmes generally focus on the development of listening and speaking first and then on reading and writing. The initial focus is on helping children familiarize themselves with the L2’s alphabet and speech sounds, which will require more or less effort depending on the learners’ L1 skills and on the similarity between the target language and their mother tongue. The approach is communicative and tends to minimize attention to form. Children’s ability to use English will be

affected by factors such as the consistency and quality of the teaching approach, the number of hours of instruction, the amount of exposure to L2, and the opportunity to use the new language. EFL young learners mainly use the target language in the school context and have a minimal amount of exposure to the foreign language. Their linguistic needs are usually biased towards one specific experiential domain, i.e. interaction in the classroom. In contrast, adolescents and adult learners are likely to encounter language use in domains outside the classroom.

- The essentials for children’s daily communication are not the same as for adults. Young children often use the FL in a playful and exploratory way (Cazden, 1974 cited in Philp, Oliver & Mackey, 2008, p. 8). What constitutes general English for adults might be irrelevant for children (particularly the youngest learners) who talk more about topics related to the here and now, to games, to imagination (as in fairy tales) or to their particular daily activities. The CEFR (2001, p. 55) states that children use language not only to get things done but also to play and cites examples of playful language use in social games and word puzzles.
- The extent to which personal and extra-linguistic features influence the way children approach the new language and the impact of these factors are often underestimated (to this regard, see Mihaljević Djigunović, 2016 in this volume): learning and teaching materials rarely make an explicit link between linguistic and cognitive, emotional, social and physical skills.

Children experience continuous growth and have different skills and needs at different developmental stages. The affective and cognitive dimensions, in particular, play a more important role for young learners than for adults, implying a greater degree of responsibility on the part of parents, educators, schools, and ministries of education. One should keep in mind that because of their limited life experience each young learner is more unique in their interests and preferences than older learners are. Familiar and enjoyable contexts and topics associated with children’s daily experience foster confidence in the new language and help prevent them from feeling bored or tired; activities which are not contextualised and not motivating inhibit young learners’ attention and interest. From a cognitive point of view, teachers should not expect young learners to be able to do a task beyond their level. Tasks requiring metalanguage or manipulation of abstract ideas should not come until a child reaches a more mature cognitive stage. Young learners may easily understand words related to concrete objects but have difficulties when dealing with abstract ideas (Cameron, 2001, p. 81). Scaffolding can support children during their growth to improve their cognition-in-language and to function independently. In fact children are dependent upon the support of a teacher or other adult, not only to reformulate the language used, but also to guide them through a task in the most effective way. Vygotsky’s (1978) notion of the teacher or “more knowledgeable other” as a guide to help children go beyond their current understanding to a new level of understanding has become a foundational principle of child education: “what a child can do with some assistance today she will be able to do by herself tomorrow” (p. 87). The implication of this for assessing what young learners can do in a new language has been well expressed by Cameron (2001, p. 119):

Vygotsky turned ideas of assessment around by insisting that we do not get a true assessment of a child’s ability by measuring what he can do alone and without help; instead he suggested that what a child can do with helpful others both predicts the next stage in learning and gives a better assessment of learning.

3 Project Background: The CEFR and the Global Scale of English

The above brief overview of the main characteristics of young learners shows the need for learning objectives that are specifically appropriate for young learners. Following the principles laid out in the CEFR, we created such a new, age-appropriate set of functional descriptors. Although adult and young learners share a common learning core, only a few of the original CEFR descriptors are suitable for young learners.

Below we first discuss the limitations of the CEFR as a tool to describe young learners’ proficiency and present our arguments for the need to complement it with more descriptors across the different skills and levels. Then, we present the Global Scale of English, a scale of English proficiency developed at Pearson (Pearson, 2015a). This scale, which is linearly aligned to the CEFR scale, is the descriptive reporting scale for all Pearson English learning, teaching, and assessment products.

3.1 The CEFR: A Starting Point

The CEFR (Council of Europe, 2001) has acquired the status of the standard reference document for learning, teaching, and assessment practices in Europe (Little, 2006) and many other parts of the world. It is based on a model of communicative language use and offers reference levels of language proficiency on a six-level scale distinguishing two “Basic” levels (A1 and A2), two “Independent” levels (B1 and B2), and two “Proficient” levels (C1 and C2). The original Swiss project (North, 2000) produced a scale of nine levels, adding the “plus” levels: A2+, B1+ and B2+. The reference levels should be viewed as a non-prescriptive portrayal of a learner’s language proficiency development. A section of the original document published in 2001 explains how to implement the framework in different educational contexts and introduces the European Language Portfolio, the personal document of a learner, used as a self-assessment instrument, the content of which changes according to the target groups’ language and age (Council of Europe, 2001).

The CEFR has been widely adopted in language education (Little, 2007) acting as a driving force for rigorous validation of learning, teaching, and assessment practices in Europe and beyond (e.g., CEFR-J, Negishi, Takada & Tono, 2012). It has been successful in stimulating a fruitful debate about how to define what learners

can do. However, since the framework was developed to provide a common basis to describe language proficiency in general, it exhibits a number of limitations when implemented to develop syllabuses for learning in specific contexts. The CEFR provides guidelines only. We have used it as a starting point to create learning objectives for young learners, in line with the recommendations made in the original CEFR publication:

In accordance with the basic principles of pluralist democracy, the Framework aims to be not only comprehensive, transparent and coherent, but also open, dynamic and non-dogmatic. For that reason it cannot take up a position on one side or another of current theoretical disputes on the nature of language acquisition and its relation to language learning, nor should it embody any one particular approach to language teaching to the exclusion of all others. Its proper role is to encourage all those involved as partners to the language learning/teaching process to state as explicitly and transparently as possible their own theoretical basis and their practical procedures. (Council of Europe, 2001, p. 18)

The CEFR, however, has some limitations. Its levels are intended as a general, language-independent system to describe proficiency in terms of communicative language tasks. As such, the CEFR is not a prescriptive document but a framework for developing specifications, for example the *Profile Deutsch* (Glabionat, Müller, Rusch, Schmitz & Wertenschlag, 2005). The CEFR has received some criticism for its generic character (Fulcher, 2004) and some have warned that a non-unanimous interpretation has led to its misuse and to the proliferation of too many different practical applications of its intentions (De Jong, 2009). According to Weir (2005, p. 297), for example, “the CEFR is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing”. In this respect, we acknowledge the invaluable contribution of the CEFR as a reference document to develop specific syllabuses and make use of the CEFR guidelines as the basis on which to develop a set of descriptors for young learners.

A second limitation in the context of YL is that the framework is adult-centric and does not really take into account learners in primary and lower-secondary education. For example, many of the communicative acts performed by children at the lower primary level lie at or below A1, but the CEFR contains no descriptors below A1 and only a few at A1. Whilst the CEFR is widely accepted as the standard for adults, its usefulness to teach and assess young learners is limited and presents more challenges. We therefore regard the framework as not entirely suitable for describing young learners’ skills and the aim of this project is to develop a set of age-appropriate descriptors.

Thirdly the CEFR levels provide the means to describe achievement in general terms, but are too wide to track progress over limited periods of time within any learning context. Furthermore, the number of descriptors in the original CEFR framework is rather limited in three of the four modes or language use (listening, reading, and writing), particularly outside of the range from A2 to B2. In order to describe proficiency at the level of precision required to observe progress realistically achievable within, for example, a semester, a larger set of descriptors, covering all language modes, is needed.

Finally, the CEFR describes language skills from a language-neutral perspective and therefore it does not provide information on the linguistic components (grammar and vocabulary) needed to carry out the communicative functions in a particular language. We are currently working on developing English grammar and vocabulary graded inventories for different learning contexts (General Adult, Professional, Academic, and Young Learner) in order to complement the functional guidance offered in the CEFR. The YL learning objectives will also have an additional section dedicated to enabling skills, including phonemic skills.

3.2 A Numerical Scale of English Proficiency

Pearson’s inventory of learning objectives differs from the CEFR in a number of aspects, most importantly, in the use of a granular scale of English proficiency, the GSE. This scale was first used as the reporting scale of Pearson Test of English Academic -PTE Academic- (Pearson, 2010) and will be applied progressively to all Pearson’s English products, regardless of whether they target young or adult learners. The GSE is a numerical scale ranging from 10 to 90 covering the CEFR levels from below A1 to the lower part of C2. The scale is a linear transformation of the logit scale underlying the descriptors on which the CEFR level definitions are based (North, 2000). It was validated by aligning it to the CEFR and by correlating it to a number of other international proficiency scales such as IELTS and TOEFL (De Jong & Zheng, *forthcoming*; Pearson, 2010; Zheng & De Jong, 2011).

The GSE is a continuous scale which allows us to describe progress as a series of small gains. The learning objectives for young learners do not go beyond the B1+ level because communicative skills required at B2 level and beyond are generally outside of the cognitive reach of learners under 15 (Hasselgren & Moe, 2006). Below 10 on the GSE any communicative ability is essentially non-linguistic. Learners may know a few isolated words, but are unable to use the language for communication. Above 90 proficiency is defined as being likely to be able to realize any communication about anything successfully and therefore irrelevant on a language measurement scale.

The GSE breaks the wide CEFR levels into smaller numeric values along its 10–90 scale; it specifies 81 points as opposed to the six levels of the CEFR (see Fig. 1). For young children especially, who progress at a slower pace than adults, this is particularly crucial. The scale offers a consistent, granular, and actionable measurement of English proficiency. It provides an instrument for a detailed account of

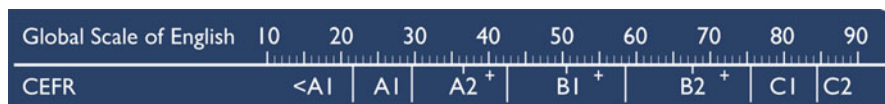


Fig. 1 The relation between the GSE and the CEFR

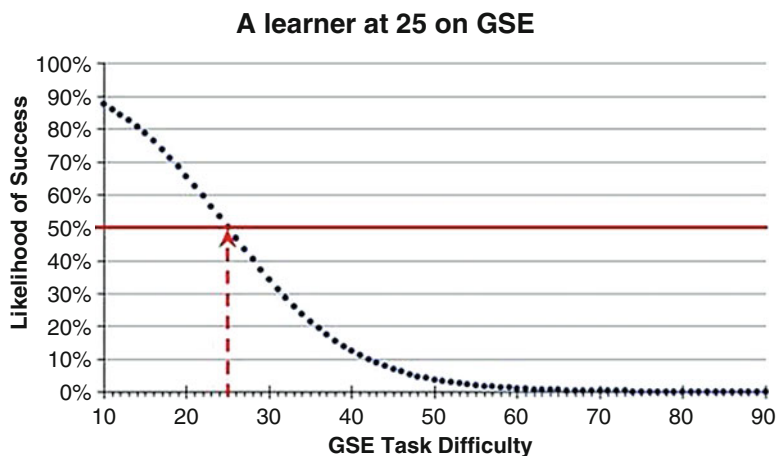


Fig. 2 Probability of success along the GSE of a learner with a score of 25

learners' levels and it offers the potential of more precise measurement of progress than is possible with the CEFR itself. The CEFR consists of six main levels to describe increasing proficiency and defines clear cut-offs between levels.

We should point out that learning a language is not a sequential process since learners might be strong in one area and weak in another. But what does it mean then to be, say, 25 on the GSE? It does not mean that learners have mastered every single learning objective for every skill up to that point. Neither does it mean that they have mastered no objectives at a higher GSE value. The definition of what it means to be at a given point of proficiency is based on probabilities. If learners are considered to be 25 on the GSE, they have a 50 % likelihood of being capable of performing all learning objectives of equal difficulty (25), a greater probability of being able to perform learning objectives at a lower GSE point, such as 10 or 15, and a lower probability of being able to cope with more complex learning objectives. The graphs below show the probability of success along the GSE of a learner at 25 and another learner at 61 (Figs. 2 and 3).

4 The Development of Learning Objectives

Pearson's learning objectives for young learners were created with the intention of describing what language tasks learners who are aged 6–14 can perform. Our inventory describes what learners can do at each level of proficiency in the same way as a framework, i.e. expressing communicative skills in terms of descriptors. In the next section we explain how we created YL descriptors sourcing them from different inputs. Then, we describe the rating exercise and the psychometric analysis carried out to validate and scale the descriptor set. Our work is overseen by a Technical Advisory Group (TAG) including academics, researchers, and practitioners

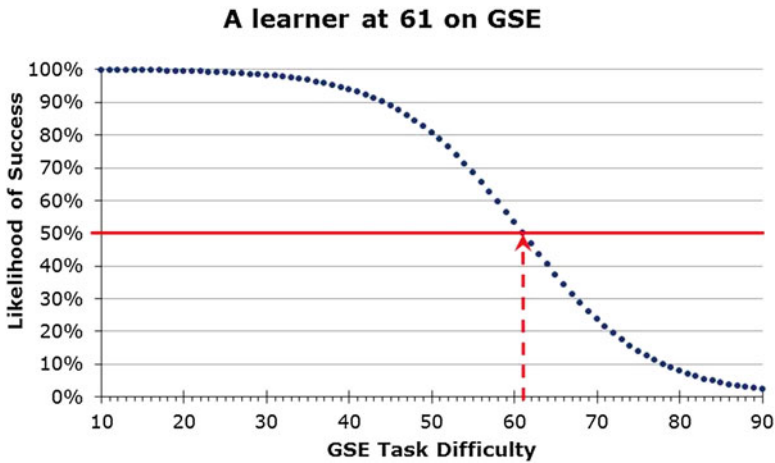


Fig. 3 Probability of success along the GSE of a learner with a score of 61

working with young learners who provide critical feedback on our methodology and evaluate the quality and appropriateness of our descriptor set and our rating and scaling exercises.

4.1 The Pool of Learning Objectives

The learning objectives were developed with the aim of describing early stages of developing ELT competencies. Accordingly, descriptors are intended to cover areas connected with personal identity such as the child’s family, home, animals, possessions, and free-time activities like computer games, sports and hobbies. Social interaction descriptors refer to the ‘here and now’ of interaction face to face with others. Descriptors also acknowledge that children are apprentice learners of social interaction; activities are in effect role-plays preparing for later real world interaction, such as ordering food from a menu at a restaurant. The present document is a report on the creation of the first batch: 120 learning objectives were created in two phases as described below: drawing learning objectives from various sources and editing them. In the next descriptor batches we are planning to refer to contexts of language use applicable particularly to the 6- to 9-year-old age range, including ludic language in songs, rhymes, fairy tales, and games.

Phase 1 started in September 2013 and lasted until February 2014. A number of materials were consulted to identify learning objectives for young learners: European Language Portfolio (ELP) documents, curriculum documents and exams (e.g., Pearson Test of English Young Learners, Cambridge Young Learners, Trinity exams, national exams), and Primary, Upper Primary and Lower Secondary course books. This database of learning objectives was our starting point to identify linguistic and communicative needs of young learners.

Phase 2 started in February 2014 and is still in progress: we are currently (summer 2014) working on two new batches of learning objectives (batch 2 and batch 3). With regard to batch 1, 120 new descriptors were created by qualified and experienced authors on the basis of the learning objectives previously identified. Authors followed specific guidelines and worked independently on developing their own learning objectives. Once a pool of learning objectives was finalised, they were validated for conformity to the guidelines and for how easy it was to evaluate their difficulty and to assign a proficiency level to them. We held in-house workshops to validate descriptors with editorial teams. Authors assessed one another's work. If learning objectives appeared to be unfit for purpose or no consensus was reached among the authors, they were amended or eliminated.

The set of 120 learning objectives included 30 for each of the four skills. Additionally, twelve learning objectives were used as anchor items with known values on the GSE, bringing the total number of learning objectives to 132. Among the anchors, eight learning objectives were descriptors taken verbatim from the CEFR (North, 2000) and four were adapted from the CEFR: they had been rewritten, rated and calibrated in a previous rating exercise for general English learning objectives. In our rating exercises for the GSE, the same anchors are used in different sets of learning objectives in order to link the data. The level of the anchors brackets the target CEFR level of the set of learning objectives to be rated: for example, if a set of learning objectives contains learning objectives targeted at the A1 to B2 levels, anchors are required from below A1 up to C1. A selection of the most YL-appropriate learning objectives from the CEFR was used as anchors.

A number of basic principles are applied in editing learning objectives. Learning objectives need to be relatively generic, describing performance in general, yet referring to a specific skill. In order to reflect the CEFR model, all learning objectives need to refer to the quantity dimension, i.e., what are the language actions a learner can perform, and to the quality dimension, i.e., how well (in terms of efficacy and efficiency) a learner is expected to perform these at a particular level. Each descriptor refers to one language action. The quantity dimension refers to the type and context of communicative activity (e.g., listening as a member of an audience), while the quality dimension typically refers to the linguistic competences determining efficiency and effectiveness in language use, and is frequently expressed as a condition or constraint (e.g., if the speech is slow and clear). Take, for example, one of our learning objectives for writing below:

- Can copy short familiar words presented in standard printed form (below A1 – GSE value 11).

The language operation itself is copying, the intrinsic quality of the performance is that words are short and familiar, and the extrinsic condition is that they are presented in standard printed form.

The same communicative act often occurs at different proficiency levels with a different level of quality.

See, for example, the progression in these two **listening** learning objectives developed by Pearson:

- Can recognise familiar words in short, clearly articulated utterances, with visual support. (below A1; GSE value 19)
- Can recognise familiar key words and phrases in short, basic descriptions (e.g., of objects, animals or people), when spoken slowly and clearly. (A1; GSE value 24)

The first descriptor outlines short inputs embedded in a visual context, provided that words are familiar to the listener and clearly articulated by the speaker. The listener needs to recognize only specific vocabulary items to get the meaning. The second descriptor shows that as children progress in their proficiency, they are gradually able to cope with descriptions that require more linguistic resources than isolated word recognition and the ability to hold a sequence in memory.

Similarly, for **speaking**, the earliest level of development is mastery of some vocabulary items and fixed expressions such as greetings. Social exchanges develop in predictable situations until the point where children can produce unscripted utterances. See, for example, the difference between a learner at below A1 and another learner at A1:

- Can use basic informal expressions for greeting and leave-taking, e.g., Hello, Hi, Bye. (below A1; GSE value 11).
- Can say how they feel at the moment, using a limited range of common adjectives, e.g., happy, cold. (A1; GSE value 22).

For **writing**, the following learning objectives show a progression from very simple (below A1) to elaborate writing involving personal opinions (B1):

- Can copy the letters of the alphabet in lower case (below A1; GSE value 10).
- Can write a few basic sentences introducing themselves and giving basic personal information, with support (A1; GSE value 26).
- Can link two simple sentences using “but” to express basic contrast, with support. (A2; GSE value 33).
- Can write short, simple personal emails describing future plans, with support. (B1; GSE value 43).

The third example above shows that ‘support’ (from interlocutor, e.g., the teacher) is recognized in the learning objectives as a facilitating condition. Support can be realized in the form of a speaker’s gestures or facial expressions or from pictures, as well as through the use of adapted language (by the teacher or an adult interlocutor).

Similarly, the following **reading** descriptors show the progression from basic written receptive skills to the ability to read simple texts with support:

- Can recognise the letters of the Latin alphabet in upper and lower case. (below A1; GSE value 10).
- Can recognise some very familiar words by sight-reading. (A1; GSE value 21)
- Can understand some details in short, simple formulaic dialogues on familiar everyday topics, with visual support. (A2; GSE value 29)

A number of other secondary criteria were applied. North (2000, pp. 386–389) lists five criteria learning objectives should meet in order to be scalable.

- Positiveness: Learning objectives should be positive, referring to abilities rather than inabilities.
- Definiteness: Learning objectives should describe concrete features of performance, concrete tasks and/or concrete degrees of skill in performing tasks. North (2000, p. 387) points out that this means that learning objectives should avoid vagueness (“a range of”, “some degree of”) and in addition should not be dependent for their scaling on replacement of words (“a few” by “many”; “moderate” by “good”).
- Clarity: Learning objectives should be transparent, not dense, verbose or jargon-ridden.
- Brevity: North (2000, p. 389) reports that teachers in his rating workshops tended to reject or split learning objectives longer than about 20 words and refers to Oppenheim (1966/1992, p. 128) who recommended up to approximately 20 words for opinion polling and market research. We have used the criterion of approximately 10–20 words.
- Independence: Learning objectives should be interpretable without reference to other learning objectives on the scale.

Based on our experience in previous rating projects (Pearson, 2015b), we added the following requirements.

- Each descriptor needs to have a functional focus, i.e., be action-oriented, refer to the real-world language skills (not to grammar or vocabulary), refer to classes of real life tasks (not to discrete assessment tasks), and be applicable to a variety of everyday situations. E.g. “Can describe their daily routines in a basic way” (A1, GSE 29).
- Learning objectives need to refer to gradable “families” of tasks, i.e., allow for qualitative or level differentiations of similar tasks (basic/simple, adequate/standard, etc.), e.g., “Can follow short, basic classroom instructions, if supported by gestures” (Listening, below A1, GSE 14).

To ensure that this does not conflict with North’s (2000) ‘Definiteness’ requirement, we have added two further stipulations:

- Learning objectives should use qualifiers such as “short”, “simple”, etc. in a sparing and consistent way as defined in an accompanying glossary.
- Learning objectives must have a single focus so as to avoid multiple tasks which might each require different performance levels.

In order to reduce interdependency between learning objectives we produced a glossary defining commonly used terms such as “identify” (i.e., pick out specific information or relevant details even when never seen or heard before), “recognize” (i.e., pick out specific information or relevant details when previously seen or heard), “follow” (i.e., understand sufficiently to carry out instructions or directions, or to keep up with a conversation, etc. without getting lost). The glossary also provides definitions of qualifiers such as “short”, “basic”, and “simple”.

4.2 *The Rating of Learning Objectives*

Once the pool of new learning objectives was signed off internally, they were validated and scaled through rating exercises similar to the methodology used in the original CEFR work by North (2000). The ratings had three goals: (1) to establish whether the learning objectives were sufficiently clear and unambiguous to be interpretable by teachers and language experts worldwide; (2) to determine their position on the CEFR and the GSE scales; and (3) to determine the degree of agreement reached by teachers and experts in assigning a position on the GSE to learning objectives.

The Council of Europe (2009) states that to align materials (tests, items, and learning objectives) to the CEFR, people are required to have knowledge of (be familiar with) policy definitions, learning objectives, and test scores. As it is difficult to find people with knowledge of all three, multiple sources are required (Figueras & Noijons, 2009, p. 14). The setting of the rating exercise for each group was a workshop, a survey or a combination of both workshop and online survey for teachers. Training sessions for Batch 1 were held between March and April 2014 for two groups accounting for a total of 1,460 raters: (1) A group of 58 expert raters who were knowledgeable about the CEFR, curricula, writing materials, etc. This group included Pearson English editorial staff and ELT teachers. (2) A group of 1,402 YL teachers worldwide who claimed to have some familiarity with the CEFR. The first group took part in a face-to-face webinar where they were given information about the CEFR, the GSE, and the YL project and then trained to rate individual learning objectives. They were asked to rate the learning objectives, first according to CEFR levels, and then, to decide if they thought the descriptor would be taught at the top, middle or bottom of the level. Based on this estimate, they were asked to select a GSE value corresponding to a sub-section of the CEFR level. The second group participated in online surveys, in which teachers were asked to rate the learning objectives according to CEFR levels only (without being trained on the GSE).

All raters were asked to provide information about their knowledge of the CEFR, the number of years of teaching experience and the age groups of learners they taught (choosing from a range of options between lower primary and young adult/adult – academic English). We did not ask the teachers to provide information on their own level of English, as the survey was self-selecting; if they were familiar with the CEFR and able to complete the familiarisation training, we assumed their level of English was high enough to be able to perform the rating task. They answered the following questions:

- How familiar are you with the CEFR levels and descriptors?
- How long have you been teaching?
- Which of the following students do you mostly teach? If you teach more than one group, please select the one you have most experience with – and think about this group of students when completing the ratings exercise.
- What is your first language?
- In which country do you currently teach?

Appendixes A and B comprise the summary statistics of survey answers by selected teachers and by selected expert raters respectively. They report data of only 274 raters ($n=37$ expert raters and $n=237$ teachers out of the total of 1,460 raters) who passed the filtering criteria after psychometric analysis.

The total of 120 new learning objectives was then subjected to rating together with twelve anchors (a total of 132 learning objectives) by the two groups. For the online ratings by the 1,402 teachers, the learning objectives were divided into six batches each containing 20 new learning objectives and four anchors. Each new descriptor occurred in two batches and each anchor occurred in four batches. The teachers were divided into six groups of about 230 teachers. Each group of teachers were given two batches to rate in an overlapping design: Group 1 rated Batches 1 and 2, Group 2 Batches 2 and 3, etc., so each new descriptor was presented to a total of about 460 teachers, whereas the anchors occurred in twice as many batches and were rated by close to a thousand teachers, producing a total of over 61,000 data points. The descriptor set was structured according to specific criteria. Similar learning objectives were kept in separate batches to make sure each descriptor was seen as completely independent in meaning. Moreover, each batch was balanced proportionally, so that each contained approximately the same proportion of learning objectives across the skills and levels in relation to the overall set. The 58 experts each were given all 120 learning objectives and the twelve anchors to rate, resulting in a total data set of more than 6,500 data points.

4.3 *The Psychometric Analysis*

After all ratings were gathered, they were analysed and were assigned a CEFR/GSE value. The data consisted of ratings assigned to a total of 132 learning objectives by 58 language experts and 1,402 teachers. Below we describe the steps we followed to assign a GSE value to each descriptor and to measure certainty values of the individuals' ratings.

As the GSE is a linear transformation of North's (2000) original logit-based reporting scale, the GSE values obtained for the anchor learning objectives can be used as evidence for the alignment of the new set of learning objectives with the original CEFR scale. Three anchor learning objectives were removed from the data set. One anchor descriptor had accidentally been used as an example (with a GSE value assigned to it) in the expert training. Independence of the expert ratings could therefore not be ascertained. Another anchor did not obtain a GSE value in alignment with the North (2000) reported logit value. For the third descriptor no original logit value was available in North (2000), although it was used as an illustrative descriptor in the CEFR (Council of Europe, 2001). Therefore, the number of valid anchors was nine and the total number of rated learning objectives was 129.

The values of the anchors found in the current project were compared to those obtained for the same anchors used in preceding research rating adult oriented learning objectives (Pearson, 2015b). The correlation (Pearson's r) between ratings

assigned to anchors in the two research projects was 0.95. The anchors had a correlation of 0.94 (Pearson’s *r*) with the logit values reported by North (2000), indicating a remarkable stability of these original estimates, especially when taking into account that the North data were gathered from teachers in Switzerland more than 15 years ago.

The rating quality of each rater was evaluated according to a number of criteria. As previously explained, the original number of 1,460 raters (recruited at the start of the project) reduced to only 274 raters after running psychometric analysis of all data. Raters were removed if (1) the standard deviation of their ratings was close to zero as this was an index of lack of variety in their ratings; (2) they rated less than 50 % of the learning objectives; (3) the correlation between their ratings on the set of learning objectives and the average rating from all raters was lower than 0.7; and (4) if they showed a deviant mean rating (*z* mean beyond $p < 0.05$). As a result, from the total of 1,460 (37 of 58 expert raters and 237 of 1,402 teachers) only 274 raters passed these filtering criteria. The selected teachers came from 42 different countries worldwide.

Table 1 shows the distribution of the learning objectives along CEFR levels according to the combined ratings of the two groups. It was found to peak at the A2 and B1 levels, indicating the need to focus more on low level learning objectives in the following batches.

Table 2 shows the certainty index distribution based on the two groups’ ratings. Certainty is computed as the proportion of ratings within two adjacent most often

Table 1 Learning objectives’ distribution across CEFR levels

GSE	CEFR	n	%
<22	<A1	4	3
22–29	A1	20	16
30–42	A2	66	51
43–58	B1	37	29
59–75	B2	2	2
76–84	C1	0	0
≥85	C2	0	0
Total		129	100

Table 2 Certainty index distribution of ratings

Certainty	n	%
>.90	29	22
.80–.90	66	51
.75–.79	25	19
.70–.74	4	3
.60–.69	5	4
<.60	0	0
Total	129	100

selected levels of the CEFR. Let us take, for example, a descriptor which is rated as A1 by a proportion of .26 of the raters, as A2, by .65 of the raters, and by .09 as B1. In this case the degree of certainty in rating this descriptor is the sum of the proportions observed with the two largest adjacent categories, i.e., A1 and A2 with .26 and .65 respectively. The sum of these yields a value of .91. This is taken as the degree of certainty in rating this descriptor. Only 4 % of the data set showed certainty values below .70, while only 7 % of the learning objectives showed certainty below .75. At this stage we take the low certainty as an indication of possible issues with the descriptor, but will not reject any descriptor. At a later stage, we will combine the set reported on here with all other available descriptor sets and evaluate the resulting total set using the one-parameter Rasch model (Rasch, 1960/1980) to estimate final GSE values. This will increase the precision of the GSE estimates and reduce the dependency on the raters involved in the individual projects. At that time the certainty of ratings will be re-evaluated and learning objectives with certainty below a certain threshold will be removed.

5 Final Considerations

In this paper we described work in progress to develop a CEFR-based descriptor set targeting young learners. In Sect. 3 we discussed limitations of the CEFR, with a special focus on its restricted suitability to describe what young learners can do in their new language. The system of levels provided by the CEFR has widely spread among practitioners and the framework has been the theme of international conferences such as EALTA and LTRC 2014. The CEFR has been validated by numerous follow-up initiatives since its publication in 2001. Since the principle of a qualitative and a quantitative dimension of language development of the CEFR is applicable to learners of all age groups, we believe the framework provides firm guidance and is suitable to be adapted to young learners. Although the present paper reports on the initial stage of the project, the analysis of the first batch of 120 learning objectives has allowed us to review our methodology to inform the next phases of the project. The current batch has shown high reliability and methodological rigour.

Next steps will include the calibration of more sets of learning objectives and the inclusion of these sets in a larger set including data on general academic and professional English learning objectives to be analysed using the Rasch (1960/1980) model for final scaling. In the near future, we hope to be able to report on the development of these additional batches of learning objectives as well as the standardisation of Pearson teaching and testing materials based on the same learning objectives and the same proficiency scale.

Appendices

Appendix A: Summary Statistics of Survey Answers by Selected Teachers (Tables 3, 4, and 5)

Table 3 Familiarity with CEFR descriptors

How familiar are you with the CEFR levels and descriptors?	n	%
I have a detailed knowledge of them	37	16
I have a general understanding of them	200	84
Total	237	100

Table 4 Age groups taught

Which of the following students do you mostly teach? ^a		
Age group	n	Percentage
6–9	55	23
9–12	71	30
10–14	90	38
12–15	92	39
15–18	82	35

^aMost teachers responded they were teaching in more than one age group

Table 5 First language

What is your first language?	n	%
Spanish	51	22
Russian	38	16
English	28	12
Italian	15	6
Greek	12	5
Polish	13	5
Romanian	11	5
Serbian	10	4
Portuguese	10	4
21 Other languages	39	16
No answer	10	4
Total	237	100

Appendix B: Summary Statistics of Survey Answers by Selected Expert Raters (Tables 6, 7, 8, 9, and 10)

Table 6 Teaching experience

How long have you been teaching?	n	%
Less than 2 years	5	2
2–5 years	5	12
More than 5 years	31	86
Total	37	100

Table 7 Country of teaching

In which country do you currently teach?	n	%
Russia	28	12
Argentina	22	9
Greece	19	8
Italy	20	8
Poland	13	5
Spain	12	5
42 other countries	109	46
No answer	14	6
Total	237	100

Table 8 Age groups taught

Which of the following students do you <u>mostly</u> teach?	n	%
<i>Lower Primary (age 6–9)</i>	10	27
Upper Primary (age 9–12)	12	32
Upper Primary/Lower Secondary (age 10–14)	3	8
Lower Secondary (age 12–15)	8	22
Upper Secondary (age 15/16–18)	2	5
Young adult/adult students – general English	2	5
Total	37	100

Table 9 First language

What is your first language?	n	%
English	10	27
Cantonese	9	24
Polish	7	19
Portuguese	3	8
Catalan	2	5
Hungarian	2	5
Spanish	3	8
Italian	1	3
Total	37	100

Table 10 Teaching experience

How long have you been teaching?	n	%
2–5 years	5	14
More than 5 years	32	86
Total	37	100

References

- Cameron, L. (2001). *Teaching languages to young learners*. Cambridge: Cambridge University Press.
- Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe. (2009). *Relating examinations to the Common European framework of reference for languages: Learning, teaching, assessment (CEFR): A Manual*. Strasbourg: Council of Europe.
- Council of Europe. (2014). *ELP checklists for young learners: Some principles and proposals. European Language Portfolio templates and resources language biography*. Retrieved June 20, 2014, from http://www.coe.int/t/dg4/education/elp/elp-reg/Source/Templates/ELP_Language_Biography_Checklists_for_young_learners_EN.pdf.
- Cummins, J. (2001). *Negotiating identities: Education for empowerment in a diverse society* (2nd ed.). Los Angeles: California Association for Bilingual Education.
- De Jong, J. H. A. L. (2009). *Unwarranted claims about CEF alignment of some international English language tests*. Paper presented at EALTA Conference, June 2009. Retrieved May 25, 2015, from http://www.ealta.eu.org/conference/2009/docs/friday/John_deJong.pdf.
- De Jong, J. H. A. L., & Zheng, Y. (forthcoming). Linking to the CEFR: Validation using a priori and a posteriori evidence. In J. Banerjee & D. Tsagari (Eds.), *Contemporary second language assessment*. London/New York: Continuum.
- European Commission. (2011). *Commission staff working paper. European strategic framework for education and training (ET 2020). Language learning at pre-primary school level: making it efficient and sustainable a policy handbook*. Retrieved July 10, 2014, from http://ec.europa.eu/languages/policy/language-policy/documents/early-language-learning-handbook_en.pdf.
- Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. Arnhem: Cito/EALTA.
- Fulcher, G. (2004). Deluded by artifices? The common European framework and harmonization. *Language Assessment Quarterly*, 1(4), 253–266.
- Glabionat, M., Müller, M., Rusch, P., Schmitz, H., & Wertenschlag, L. (2005). *Profile deutsch A1-C2 (Lernzielbestimmungen, Kannbeschreibungen, Kommunikative Mittel)*. München: Langenscheidt.
- Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing*, 22(3), 337–354.
- Hasselgren, A., Kaledaitė, V., Maldonado-Martin, N., & Pizorn, K. (2011). *Assessment of young learner literacy linked to the common European framework for languages*. European Centre of Modern languages/Council of Europe publishing. Retrieved July 15, 2014, from http://srvcs-npbs.xtec.cat/cirel/cirel/docs/pdf/2011_08_09_Ayllit_web.pdf.
- Hasselgren, A., & Moe, E. (2006). *Young learners' writing and the CEFR: Where practice tests theory*. Paper presented at the Third Annual Conference of EALTA, Kraków. Retrieved July 15, 2014, from http://www.ealta.eu.org/conference/2006/docs/Hasselgren&Moe_ealta2006.pdf.
- 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.

- Lightbown, P. M., & Spada, N. (2008). *How languages are learned*. New York: Oxford University Press.
- Little, D. (2006). The common European framework of reference for languages: Content, purpose, origin, reception and impact. *Language Teaching*, 39(3), 167–190.
- Little, D. (2007). The common European framework of references for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal*, 91(4), 645–655.
- McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.
- Mihaljević Djigunović, J. M. (2016). Individual differences and young learners' performance on L2 speaking tests. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.
- Negishi, M., Takada, T., & Tono, Y. (2012). *A progress report on the development of the CEFR-J* (Studies of Language Testing, 36, pp. 137–165). Cambridge: Cambridge University Press.
- Nikolov, M. (2016). 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*. New York: Springer.
- Nikolov, M., & Curtain, H. (Eds.). (2000). *An early start: Young learners and modern languages in Europe and beyond*. Strasbourg: Council of Europe.
- Nikolov, M., & Mihaljević Djigunović, J. (2006). Recent research on age, second language acquisition, and early foreign language learning. *Annual Review 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. (2012). Developing diagnostic tests for young learners of EFL in grades 1 to 6. In Galaczi E. D. & Weir C. J. (Eds.), *Voices in language assessment: Exploring the impact of language frameworks on learning, teaching and assessment: Policies, procedures and challenges* (pp. 347–363). Proceedings of the ALTE Krakow Conference, July 2011. Cambridge: UCLES/Cambridge University Press.
- North, B. (2000). *The development of a common framework scale of language proficiency*. New York: Peter Lang.
- Oppenheim, A., N. (1966/1992) (2nd ed.) *Questionnaire design, interviewing and attitude measurement*. London: Pinter Publishers.
- Pearson. (2010). *Aligning PTE Academic Test Scores to the common European framework of reference for languages*. Retrieved June 2, 2014, from http://pearsonpte.com/research/Documents/Aligning_PTEA_Scores_CEF.pdf.
- Pearson. (2015a). *The Global Scale of English*. Retrieved May 25, 2015, from <http://www.english.com/gse>.
- Pearson. (2015b). *The Global Scale of English Learning Objectives for Adults*. Retrieved May 25, 2015, from <http://www.english.com/blog/gse-learning-objectives-for-adults>.
- Philp, J., Oliver, R., & Mackey, A. (Eds.). (2008). *Second language acquisition and the young learner: Child's play?* Amsterdam: John Benjamins.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago: The University of Chicago Press.
- Singleton, D. (1989). *Language acquisition. The age factor*. Clevedon: Multilingual Matters.
- Speitz, H. (2012). Experiences with an earlier start to modern foreign languages other than English in Norway. In A. Hasselgren, I. Drew, & S. Bjørn (Eds.), *The young language learner: Research-based insights into teaching and learning* (pp. 11–22). Bergen: Fagbokforlaget.
- Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Oxford, UK: Palgrave.
- Zheng, Y., & De Jong, J. (2011). *Establishing construct and concurrent validity of Pearson Test of English Academic* (1–47). Retrieved May 20, 2014, from <http://pearsonpte.com/research/Pages/ResearchSummaries.aspx>.

A Framework for Young EFL Learners' Diagnostic Assessment: 'Can Do Statements' and Task Types

Marianne Nikolov

Abstract The aim of this chapter is to present a framework for assessing young learners of foreign languages for diagnostic purposes. The first section outlines the most important trends in language assessment and describes the educational context where the project was implemented. Then, the chapter discusses how children between the ages of 6 and 12 develop in a foreign language and outlines the most important principles of assessing young language learners. The actual framework was designed for the four skills; it aimed to cover the first 6 years of primary education in Hungarian public schools. The document used the *Common European Framework of Reference (CEFR, 2001)* as a point of departure and includes age-specific 'can do statements' and task types corresponding to them. Readers are encouraged to critically reflect on how the findings could be adopted in their own contexts.

Keywords Principles of early language teaching • Framework • Diagnostic assessment • Task types • CEFR

1 Introduction

The chapter presents some of the results of a national project conducted in Hungary in the hope that readers may find them useful in their own contexts. The first part of the chapter embeds the project in recent trends in educational and language assessment and the educational context where the project was implemented. In order to develop age-appropriate diagnostic tests for learners of English as a foreign language (EFL) in the first 6 years of primary school (ages 6–12) in the four skills, a framework was designed in line with the *Common European Framework of Reference (CEFR, 2001)*, including 'can do statements' and task types corresponding to them (Nikolov, 2011). As a next step, diagnostic tests were developed and validated (Nikolov & Szabó, 2012a, 2012b; Szabó & Nikolov, 2013). These

M. Nikolov (✉)

Institute of English Studies, University of Pécs, Pécs, Hungary

e-mail: nikolov.marianne@pte.hu

calibrated tasks are meant to be available to teachers for their classroom use. This chapter focuses on the main features of the framework, what *can do statements* and various task types were specified and what lessons were learned from various phases of the project.

2 Contextualizing the Project in Recent Educational Assessment Trends

Recent trends in educational research are highly relevant to early language learning, since they have opened new avenues on how different approaches to assessment, diagnostic and dynamic testing as well as peer and self-assessment, can boost learners' learning potential (Alderson, 2005; Rixon, 2016, Hung, Samuelson, & Chen, 2016 in this volume; Sternberg & Grigorenko, 2002) and also offer teachers feedback on their own work and where students are in their development. Besides traditional ways of *assessment of learning*, the need to focus on *assessment for learning* has been widely emphasized not only in language learning but in other domains as well (Assessment Reform Group, 2002; Black & Wiliam, 1998; Davison & Leung, 2009; Leung & Scott, 2009; McKay, 2006; Teasdale & Leung, 2000; also see Rixon, 2016 in this volume). These shifts in emphasis on how children can benefit from classroom testing, and how teachers can scaffold their development have resulted in new studies. Assessment should be sensitive to the issue of readiness to develop (McNamara & Roever, 2006, pp. 251–252); this is an area where more research is needed to find out how learners can benefit from different kinds of interaction (Nikolov & Mihaljević Djigunović, 2011, p. 111) and how their teachers can use diagnostic information. These points are crucial for young learners, as their progress in their new language depends on their classroom experiences and feedback from their teachers and peers. Techniques applied in diagnostic assessment may also open new avenues for developing learner autonomy by involving students' in their own development.

Before moving on we need to discuss how diagnostic assessment is defined, what the key characteristics are, and how the concept fits the picture outlined so far. Definitions of diagnostic assessment share the following features:

- (1) “diagnostic tests seek to identify those areas in which a student needs further help” (Alderson, Clapham & Wall, 1995, p. 12);
- (2) records on diagnostic assessments indicate “specific areas of strengths and weaknesses in language ability” (Bachman & Palmer, 2010, p. 196);
- (3) diagnostic tests can be theory or syllabus-based (Bachman, 1990, p. 60);
- (4) tests developed for other purposes, for example, for progress, proficiency or placement, can be and are often used diagnostically (Alderson, 2005; Bachman & Palmer, 2010);
- (5) information on learners' strengths and weakness can lead to action: teachers can use results to tune their teaching to learners' needs and learners may seek out more opportunities to practice in the problem areas;
- (6) diagnostic tests are hard to develop and are rarely investigated (Alderson, 2005, p. 6).

In sum, diagnostic assessment is an area where learning, instruction, and assessment overlap and interact. In the case of young learners, diagnostic tests have to be driven by both theory and curriculum, since they have to reflect how children learn in general and how they develop in their FL in particular, and also, what they have had access to in their contexts.

The focus on how learning and assessment interact with young learners' individual differences may shed light on a new challenge in early language learning: why after the first enthusiastic period of learning a new language, young children's motivation, similarly to experiences with older learners, declines over time (Mihaljević Djigunović, 2009a; Nikolov, 1999) and how young children's self-perceptions are shaped by what happens in the classroom over the months and years. Most studies exploring these issues have found that varied levels of performances on tests play a key role in how motivated and anxious learners are, how they perceive themselves and what goals they set for themselves in the long run. There is an important relationship between young language learners' ID and assessment.

The recent international trends in assessment mentioned above motivated a large-scale innovative project for students in the first 6 years of primary education in Hungary. A longitudinal project implemented at the Center for Research for Learning and Instruction at the University of Szeged aimed to develop an online system for diagnostic assessment for the first six grades in public schools in reading in Hungarian as a first language (L1), mathematics and science (Csapó & Csépe, 2012; Csapó & Szabó, 2012; Csapó & Szendrei, 2011) as well as in English as a foreign language (EFL). The framework of the project discussed in this chapter (Nikolov, 2011; Nikolov & Szabó, 2011a, 2011b) is part of this larger one (Csapó & Zsolnai, 2011) and was funded by The EU Social Renewal Operational Program (TÁMOP-3.1.9-08/01-2009-0001).

3 The Context of the Project

It is particularly important for Hungarians to be able to use foreign languages (FLs), since the official language, Hungarian, is hardly used in other countries. For many decades, Russian was the mandatory FL and few people had access to other languages. Despite the high expectations after the change of regime in 1989 when access to more desirable foreign languages, most importantly to English and German, became easier in schools, the expected dynamic increase was not observed in the ratio of people developing proficiency in these languages (Medgyes & Nikolov, 2014; Nikolov, 2009). Although the Eurobarometer surveys (2006, 2012) show an increase in the ratio of FL speakers (in 2006, 29 % of respondents said they could use at least one FL; in 2012 the ratio was 35 %), still only every third citizen claimed, based on their self-assessment, to be able to use a FL (compared to the average of 54 % of Europeans). Over the last two decades German and English have enjoyed a special status in Hungarian education: both have been perceived as valuable assets for students' future careers. Therefore, parents and students tend to

consider the quality of foreign language instruction at schools when they choose an institution. Since the 1990s, some important changes have emerged in foreign language education: (1) the demand for English as a lingua franca has dynamically increased and, in contrast, German has lost some of its appeal (Medgyes & Nikolov, 2014); (2) due to parental pressure, the age when children start learning a FL has decreased, despite the fact that language policy documents have maintained grade 4 as the mandatory start of FL learning (Nikolov, 2009). As a result of this controversial regulation, parents who are very keen on their children's early learning of a foreign language press schools to lower the time of starting a FL. Schools are supported per capita by the ministry, therefore, it is their interest to satisfy needs by launching early language programs to attract students.

This situation is further complicated by the increasingly higher value attached to English than to German, and the fact that teachers are tenured in their jobs and German classes also have to be filled. As English is a lot more popular, schools stream students in different language groups. More able and socially more privileged students tend to start learning a FL earlier and the ratio of English learners is higher than that of learners in German classes. Also, students with higher socio-economic background and better achievements in other school subjects attend more intensive programs, whereas less able students, often coming from poorer and less educated families, tend to start later, they are taught in fewer weekly classes and are often placed in German classes, although they would prefer to learn English.

Due to these interrelated reasons, in various large-scale testing projects involving representative samples of students in years 6, 8, 10 and 12, significant differences have been found in students' proficiency levels studying English and German: results tend to be higher in English (Csapó & Nikolov, 2009; Nikolov, 2011; Nikolov & Józsa, 2006). Another important outcome is that a very wide range of achievements is typical across all levels of education and the differences increase as students make progress in their studies, thus, many children are left behind. Learners of English tend to achieve higher scores and their attitudes and motivation are consistently more favorable than those of their peers learning German (Dörnyei, Csizér, & Németh, 2006; Nikolov, 2003). Classroom practice, however, is typically characterized by similar practices often focusing on form and applying grammar-translation type of drills rather than focusing on meaning even in the younger age groups (Nikolov, 2003, 2008).

As for how much it matters when children start learning a foreign language, minimal contributions were found of an early start in a national project involving representative samples of English and German learners in their 6th and 10th grades (age 12 and 16). As the results of regression analyses indicate in Table 1, the number of years students studied English and German explains 3 and 4 % of variance in their scores, whereas the number of weekly classes between 10 and 14; however, students' socio-economic status explains 25–24 % of variance in English and 18–17 % in German achievements. In other words, whether students started early or late, made hardly any difference in their levels of proficiency in any of the 2 years or languages.

Table 1 Variables contributing to Hungarian learners' performances in English and German ($r\beta\%$) (Nikolov & Józsa, 2006, p. 211)

Independent variables	Year 6		Year 10	
	English	German	English	German
Parents' education (SES)	25	18	24	17
Weekly classes	13	10	14	13
Years of language study	3	3	4	4
Private tuition	ns	ns	2	2
Variance explained (%)	41	31	44	36

4 The Diagnostic Assessment Project

The aims and achievement targets of the Hungarian diagnostic assessment project had to be in line with theories on how children learn a FL, curricular requirements, and realities in schools. For FLs, various versions of National Core Curricula (NCC; for a critical overview see Medgyes & Nikolov, 2010) preceded the version published in 2007. This was the version the diagnostic project had to be in line with. Despite the fact that in 2006 every fourth school started teaching a FL before the mandatory grade 4 (Nikolov, 2009), the official curriculum maintained that all schools had to offer students at least one FL from fourth grade (age 10) and it allowed them to start earlier upon parents' requests. However, no official curriculum was available for the first three grades (ages 6–9), and no goals and achievement targets were set for the first years (Nikolov, 2011). Therefore, one of the aims was to outline a framework for EFL for the first six grades of public schools.

The NCC (2007) prescribed dual levels of achievement targets for the 9 years of compulsory FL learning between grades 4 and 12 (age 10 and 18), depending on long term goals: whether students aimed to take an intermediate (B1 level) or advanced (B2) level school-leaving examination at the end of their education at age 18. The NCC explicitly stated that the construct was communicative competence (useful language ability) in the four skills (listening, speaking, reading, and writing) and the required levels were in line with the *CEFR* (2001); the levels students had to achieve were independent of when they started learning a FL and how intensive their courses were. By the end of year 6, students were expected to be at the A1- or A1 level, whereas at the end of year 8, at the A1+ or A2- level in the four skills. The NCC specified provision in loose terms: in grades 1 to 4, 2–6 % of the overall classes (1 or 2 per week) could be devoted to teaching a FL, whereas in grades 5 to 8, 12–20 % (2–6 classes). However, some schools could also launch content and language integrated learning type of dual-language classes, teaching some subjects in the target language, but achievement targets were not specified until a new version of NCC (2012) was published.

Besides achievement targets in the foreign language, the NCC (2007) specified some further aims: they included the development of learners' positive attitudes towards language learning and towards other cultures, their motivation to improve their proficiency and to learn about the target culture as well as other cultures, and

their language learning strategies. Therefore, these were also included in the framework.

The language testing background to our study is based on the conceptualization of communicative competence and language ability (Council of Europe, 2001): learners' performances are assessed in their four language skills. In the choice of task and text types, piloting and validating tests, and evaluating results, we followed the principles of communicative language testing in general (Alderson, 2005; Alderson et al., 1995; Bachman & Palmer, 2010), and assessing young learners in particular (McKay, 2006; Nikolov, 2011; Nikolov & Szabó, 2011a, 2011b).

4.1 Aims and Phases

The aims for the first phase of the diagnostic assessment project were (1) to design a framework based on research into how young learners of a FL develop and the main principles of teaching and assessing them; (2) to draw up a list of can do statements for young learners in the first six grades of public schools for the levels required in the curriculum; (3) to identify topics, text types and task types that would allow valid, reliable and age-appropriate diagnostic assessments of the target age group in the four skills in EFL in line with curricular requirements. In the following sections these three points are discussed.

As a first step, a detailed analysis of the literature was conducted with the following focal points: (1) how young learners of various first languages, including Hungarian as L1, develop in English as a foreign language, (2) the main principles of teaching and assessing children in their new language in the first six grades, and (3) what is known about classroom practice in the first 6 years of EFL in Hungarian public schools. In addition to these, a small-scale focused project was implemented to explore (4) what teaching materials and tests are used in EFL classes and how teachers apply them for assessment.

4.2 The Framework

In this section we summarize the main points related to how children between the ages of 6 and 12 develop in a FL and outline the most important principles for assessing their development. This short overview is based on a range of handbooks and empirical studies on early language learning and teaching (e.g., Nikolov & Mihaljevic-Djigunovic, 2006, 2011).

It has been widely accepted that the younger the learners are, the more similar the process of their FL development tends to be to the acquisition of their first language(s) and the less able they are to learn and apply language rules consciously. Language learning is based on two distinct processes (MacWhinney, 2005, Paradis, 2004, 2009, Skehan, 1998, Ullman, 2001). Implicit learning is based on memoriz-

ing unanalyzed wholes, chunks, formulaic expressions in context, as well as rules, whereas explicit learning is rule-based, and it allows learners to formulate new utterances and express their ideas in new ways. The ability to rely on explicit learning emerges in all learners over time and it gains a major role around puberty. The two processes interact with one another dynamically; however, the younger children are the more decisive implicit learning is.

Young learners find pleasure in age-appropriate activities including telling rhymes, singing songs, playing games, listening to and telling picture stories, acting out roles, etc., and they tend to pick up unanalyzed chunks from these and from classroom language in contexts where they can understand meanings. At later stages guessing games, stories, role-plays and a range of meaning-focused tasks recycle familiar language and offer opportunities to learn new meanings in intrinsically motivating and cognitively challenging activities (Nikolov, 1999, 2002), two key qualities necessary for maintaining young learners' interest in tasks. Most young learners tend to enjoy telling rhymes and singing songs at the early stage of language learning, and body language and other visual support can scaffold comprehension and their FL development. As a lot of revision and recycling is necessary, activities need to be varied and build on one another to avoid boredom and scaffold development (Curtain & Dahlberg, 2010; Nikolov, 2002). Activities have to offer opportunities to recognize and use familiar chunks and expressions, including, for example, greetings, instructions, feedback and other types of language related to classroom management. Many young learners prefer acting and speaking in groups at the early stage and they become more willing to perform tasks individually and in pairs at later stages. Some children may be anxious, contrary to common belief assuming that all young learners have low anxiety and are motivated (Mihaljević Djigunović, 2009b).

Children are able to comprehend a lot more than they can produce; if tasks are tuned to their abilities and background knowledge of the world, they are able to figure out new meanings they are not familiar with. Their inductive reasoning skills allow them to guess meaning in context and if they are encouraged to do so, they will be able to apply this extremely useful strategy over time. Guessing often happens in the children's first language allowing them to make sense of one another's comments (Nikolov, 2002). This process can offer teachers important insights into children's thought processes, strengths and weaknesses. Therefore, it is a must for them to be able to comprehend their young learners' L1 in order to build on their meaning making and recycle what they say in the target language, thus building on what they know and are familiar with and what they are not.

A key principle concerns comprehensible input (Krashen, 1985): children need to make sense of what they hear and read by relying on their background knowledge of the world and of contexts, what others' intentions may be. However, they also need opportunities to apply what they are ready to use (Swain, 2000) and interact and experiment in order to get feedback allowing them to develop further. In fact, focus on knowing equivalents in the two languages, translation of word meaning is not necessary for children to be able to comprehend and use the target language.

Errors are typical and they indicate where the children are in their process of learning the target language; similarly to L1 development, errors emerge and then tend to disappear with time if enough learning opportunities are offered. Certain features of interlanguage indicate the developmental stages children are at. Many young learners progress from a silent period in their foreign language class and they may be willing to respond by movements or body language or in their FL, indicating their level of listening comprehension. Typical developmental stages are marked, for example, by the use one-word or two-word utterances, or omission of certain words (e.g., copula) or the use of external *no* in negation (*no dog*) in speaking. They often transfer their L1 pronunciation in the case of cognates (e.g., elephant, television, computer) or intonation patterns in questions, for example. These developmental errors indicate the learning process and they tend to disappear over time or with the help of tasks helping children notice gaps (Schmidt, 1990) at a further developmental stage when they are ready.

The distinction between basic interpersonal communication skills and cognitive academic skills (Cummins, 2000) can highlight yet another important principle: most children develop along similar lines in their oral and aural skills, but more visible individual differences tend to emerge in their literacy development. These differences are related to children's aptitude, literacy skills in their first and other languages and these interact with their socio-economic status. Several empirical studies revealed important relationships between young learners' level of aptitude, their L1 skills and their socio-economic status in the Hungarian educational context (Bacsá & Csíkos, 2016 in this volume; Csapó & Nikolov, 2009; Kiss & Nikolov, 2005; Nikolov & Csapó, 2010; Nikolov & Józsa, 2006) and in other countries as well (e.g., Alexiu, 2009; Mihaljević Djigunović, 2012; see also findings by Wilden & Porsch, 2016 in the present volume on multilingual young learners' receptive skills in English and German).

The interaction between young learners' languages is further underpinned by findings in classroom research. In a language, like Hungarian, with a highly transparent sound – letter correspondence all children who can read words in their L1 will apply their L1 phonetic rule in English and read out words phonetically. This strategy may support memorizing the spelling of words phonetically in L1 but may negatively impact reading (Nikolov, 2002). Hungarian learners of all ages who can spell and write well tend to apply this strategy.

The younger the learners are the slower their development is in their new language compared to older learners (Krashen, 1985, Nikolov & Mihaljević-Djigunović, 2006). Findings of two longitudinal studies provide evidence in European EFL contexts (García Mayo & García Lecumberri, 2003; Muñoz, 2006), whereas studies in English as a second language (ESL) contexts have found that 5–7 years are necessary for children to achieve native-like proficiency in immersion programs (Wong Fillmore, 1998) where the teachers and many of the peers are native speakers and the language of instruction is English. This slow speed of progress has important implications for teaching and assessment.

The main argument for an early start is often the critical period hypothesis; the assumption that language acquisition has to start before a certain time in one's life,

otherwise an accent is unavoidable and proficiency will be limited. A range of publications have pointed out why the argument is hardly relevant in foreign language contexts (e.g., DeKeyser & Larson-Hall, 2005, Nikolov & Mihaljevic-Djigunovic; 2006, Scovel, 2000; Singleton & Ryan, 2004) where young learners have limited access to authentic language use, especially because their teachers use English as a lingua franca. Achievement targets have to be in line with what is available in and outside the classroom.

As children's attention span is short, tasks have to be in line with how long young learners can focus on a certain activity. Also, tasks have to be intrinsically motivating so that they are worth doing and repeating. Extrinsic motives in the form of feedback on achievements are also important, so it is essential that all tasks have clear outcomes. Repetition of the same task may lead to boredom, so activities should be varied and children should be familiar with a range of task types so that recycling is possible in a motivating way. Tasks also have to challenge learners and offer them opportunities to make efforts and develop a growth mindset (Dweck, 2006). It is crucial that young learners believe that learning a new language is possible and they can do it. These affective aims concerning positive beliefs, self-esteem and self-confidence may turn out to be more important over time than the actual language skills young learners develop in the first few years.

Tasks have to be realistic and tuned to learners' abilities and needs. They should be neither too easy nor too difficult, as both can discourage young learners from extending themselves and showing their true abilities (McKay, 2006). Young learners need feedback on how they perform on tasks, what they are good at and what they need to practice more to perform better. Lack of success may demotivate young learners in the long run (Nikolov, 2001).

Although young learners are often assumed to be all highly motivated and lacking anxiety, important differences characterize even children as language learners (Mihaljević Djigunović, 2009b). Instrumental motives concerning how useful proficiency in a FL will be in adulthood will not guarantee focused attention on classroom tasks and motivated behavior despite the fact that children are aware of such long-term goals (Nikolov, 2002). The teacher plays a special role in young learners' motivation: the younger they are the more their attitudes and motivation are influenced by their teacher and this relationship gradually weakens over the years (Nikolov, 1999).

Language learning strategies emerge at the earliest stages and their number and conscious uses gradually increase over the years. Young learners can notice similarities and differences between the pronunciation and vocabulary use of their teacher's and cartoon characters, guess meaning, repeat and memorize words and chunks (Mihaljević Djigunović, 2001), encourage and help themselves and others, reflect on and self-evaluate their performances (Nikolov, 2002; Pinter, 2006, 2007a, 2007b).

Some further points concern tasks and how assessment should be implemented. First of all, all tasks used for assessment should be familiar to learners. They should be appropriate not only for assessment but also for learning. The setting where the assessment is conducted should also be familiar to the children and they should be

able to understand what they can and cannot do well. An emphasis on positive outcomes and encouragement are crucial when assessing young learners; as they need to feel successful, tasks should be doable to avoid frustration. It is also important to bear in mind how performing in front of others may induce anxiety in children, so working in pairs or small groups should be alternatives (Nikolov & Mihaljević Djigunović, 2011).

Tasks should focus on meaning (not form) and allow young learners to communicate with their peers and their teacher (Nikolov, 2011). As at the early stages of language learning children are not proficient in their literacy skills in their mother tongue, both teaching and assessment should focus on their listening comprehension and speaking skills; and reading comprehension and writing should be introduced gradually when they are ready for them.

Tasks used in course books often integrate more than one language skill; however, during assessment it is important to try to focus on skills separately so that the skill and subskill is specified where children's strengths and weaknesses are identified (Alderson, 2005; McKay, 2006).

Feedback and evaluation must always come right after students' performance, it should be individualized and also motivating for further learning (Nikolov, 2011).

Diagnostic assessment should be regular, it should tap into the small developmental steps and should provide clear feedback so that young learners can feel that they are making progress and achieving what they are expected to (Nikolov, 2011).

Both self- and peer-assessment can be effectively used in diagnostic assessment, as they may contribute to learner autonomy, encourage the use of learning strategies and children can scaffold one another's FL learning (for detailed discussions see Rixon, 2016 and Hung et al., 2016 in this volume and McKay, 2006).

As for the content of assessment tasks, themes and topics listed in curricula and discussed in typically used teaching materials should be drawn on bearing in mind both the children's local and the target language cultures.

The first draft of the above framework for English as a foreign language was one of the documents used in the project and then, after piloting diagnostic tests, integrated into the final framework published in Hungarian (Nikolov, 2011).

4.3 Findings on EFL Teachers' Assessment Practices in Early Language Programs in Hungary

Prior to the project a lot of data were available on how teachers develop but less on how they assess their young EFL learners in primary schools. Observations and interviews were conducted (Bors, Lugossy, & Nikolov, 2001; Nikolov, 2008) and questionnaire data were also collected from students in large-scale national testing projects (Csapó & Nikolov, 2009; Nikolov, 2003; Nikolov & Józsa, 2006). The main findings indicated that the most typical classroom activities were rarely in line with age-appropriate teaching methodology; teachers tended to focus on grammar,

and translation and reading out loud were the most often applied techniques of meaning making and testing. These activities were the most disliked ones among students in addition to other written tests, whereas the most favored classroom activities included watching videos, acting out role plays, and other oral tasks; these were the least often applied. Overall, these classroom-based studies shed some light on why the efficiency of early start programs was low and the need for further research.

As these surveys did not directly focus on teachers' assessment practices, a small-scale project was designed to explore what specific tests highly experienced teachers of young learners used and how they assessed their learners with the help of these test tasks in their classrooms in the first six grades (Hild & Nikolov, 2011). A convenience sample of twelve Hungarian teachers of English volunteered (for payment) to choose and characterize tests they often used in their lessons for diagnostic assessment purposes. The respondents were asked to describe and attach the actual tasks and to fill in a short questionnaire on them to reveal how they actually diagnosed their students' strengths and weaknesses, how they gave them feedback, and what level the tasks were in their views. Teachers analyzed 119 tasks; most of them integrated various skills or comprised a sequence of tasks building on one another. The largest category of tasks integrated reading comprehension and writing skills; tasks in the second main category integrated listening comprehension and speaking skills, whereas the third group integrated three skills. Five tests were meant to develop listening, speaking and writing; four reading, writing and speaking; two listening, reading and speaking; and one listening, reading and writing. Two of these tasks assessed surprising domains: reading comprehension, practice of punctuation and negative forms; listening, lip reading, and speaking. Twelve tasks assessed speaking exclusively. The fifth category comprised eleven tasks that integrated reading and speaking, whereas eleven tasks assessed writing and nine listening skills. The last three categories comprised seven speaking and writing tasks, seven reading and two other tasks (listening and reading; reading and vocabulary) (Hild & Nikolov, 2011).

In sum, the tests teachers used varied to a great extent and the main findings were that (1) teachers found it hard to apply the categories we clarified in the data collection instrument and they were supposed to be familiar with; (2) they applied fuzzy terms for assessing learners' performances and not criteria; (3) the tests either tapped into two or more skills in an integrated manner, thus it was not possible to find out which skill they measured, or they comprised sequences of tasks where the outcomes of the first part determined how well students could perform on the next ones; (4) they tended to focus on errors, accuracy and what students cannot do rather than fluency, vocabulary and what students can do; (4) the feedback teachers gave learners typically meant rewards for best performances, but no reward for less good performances, thus only top achievers got feedback. These techniques could demotivate less able learners and rewards did not give information on what areas needed improvement (Hild & Nikolov, 2011).

As a result of the above small-scale survey and an extensive analysis of the task and text types used in teaching materials, an exhaustive list of test and text types was

compiled. Then, these were compared and contrasted with can do statements in *CEFR* at A1 and A2 levels in a two-day workshop in Pécs in June 2010. Participants included highly experienced primary-school teachers of English, and a team of Hungarian and international experts on researching and testing young learners (see Acknowledgements). The themes and topics in the teaching materials were also overviewed and matched with the ones listed in the NCC (2007) before the final list was drawn up.

In the final list of task types, the following criteria were used for inclusion: (1) task was age-appropriate; (2) task was in line with how children develop in a L2; (3) it reflected good practice; (4) children's performance on the task could be measured (quantified); (5) task was appropriate both for developing and testing one or more clearly specified skills or subskills in 'can do statements' listed in the framework; (6) task was within the attention span of the target group; (7) task was expected to be intrinsically motivating for young learners. In the next sections the results are presented: first the 'can do statements', then the topics, text and task types are discussed.

5 'Can Do Statements', Topics, Types of Texts and Tasks for Diagnostic Tests for Children Between the Ages of 6 and 12

5.1 Can Do Statements

One of the many challenges in drawing up what children can do concerns their slow progress in the first few years of their learning of a new language. Some previous work was available on how the *CEFR* had adapted to accommodate young learners' needs (e.g., Hasselgren, 2005; Pižorn, 2009; Papp & Salamoura, 2009); these were consulted before the actual list of can do statements were drawn up.

As the teachers we intended to involve in the project needed reference points to guide them in estimating the level of their students in an educational context where children may start learning a foreign language in any of the grades, we tried to establish three levels within the continuum specified in the curriculum for grades 1 to 6. The following criteria were used to define these levels and we labelled them as (1) beginner, (2) beginner plus, and (3) elementary levels, corresponding to the A1-, A1, and A2- levels in the *CEFR* (2001).

An important point concerned how teachers who joined the piloting phase of the diagnostic assessment project could decide which level their classes should target. The list of can do statements were meant for them, too, to help them estimate the level of difficulty of the tasks. The following criteria were drawn up to help teachers decide in terms of how much instruction was most probably in line with the levels.

A1- Beginner: This level describes what children can realistically be expected to do by the end of 4th grade (age 10), after studying EFL for 1–4 years, in 1–3 h per

week. Included in this level are absolute beginners (with no previous exposure to English at all) as well as false beginners (who may have been exposed to some English by hearing it from their parents, in kindergarten, in private lessons, on television, in computer games or while staying abroad).

A1 Beginner Plus: This level describes what learners can realistically be expected to do by the end of 5th grade (age 11), after studying EFL for 2–5 years in 1–3 h per week.

A2- Elementary: This level is assumed to be what learners can realistically achieve by the end of 6th grade (age 12), after studying EFL for 3–6 years in 1–3 h per week.

In addition to these points, it was clarified that as children starting to learn English at age 6 are at a very low level in their literacy skills in their L1, the can do statements in reading and writing are not relevant in their case, only the listening comprehension, speaking and interaction ones are. In other words, the levels in the various skills can vary. Thus, young learners are not expected to achieve the same level in the four skills, as curricula may vary a lot.

As Table 2 shows, the can do statements are arranged in three skill areas. In the first one listening comprehension, speaking and interaction are listed together, whereas reading and writing are put in two groups. There are many more statements in the first group, as this is where at this very low level (A1-) young learners are expected to be able to do more in listening comprehension, speaking and interaction than in their literacy skills.

As Table 3 shows, the list of can do statements is longer, and in some only a single word or expression is different from the wording in Table 2. The statements are listed in the same order as in Table 2 in order to allow users to notice the differences. Some of the can do statements are specific to the teaching traditions of Hungarian learners, for example, spelling is included under reading. This is level A1 in *CEFR*.

As Table 4 shows, can do statements for the elementary (A2-) level expand the ones in the previous two tables. In some of them references to classroom contexts are included, for example, “Can ask a question or help peers when they are stuck.” An additional feature refers to accuracy: at this level students are expected to be able to do what they could not do very well without mistakes. It was felt that this was a necessary addition in order to avoid fossilization and the typical complaint on the part of teachers in later years that there is hardly anything to rely on when young learners enter secondary schools.

5.2 Language Learning Strategies

In addition to the can do statements listed in Tables 2, 3, and 4, a list of various learner strategies were also collected. These are the ones children need to be able to apply to develop and in order to demonstrate what they can do. Some language

Table 2 What can young learners do at the beginner level?

Listening comprehension, speaking and interaction	Can follow simple instructions in English in familiar contexts and can respond through total physical response, using body language, facial expression, or one-word answers in English.
	Can participate in activities and tasks by following classroom instructions in English.
	Can participate in classroom activities and tasks individually, in pairs and in groups.
	Can comprehend the meaning of frequently used words, expressions, requests and questions in English.
	Can guess the meaning of familiar or new English words from short, simple definitions/explanations in Hungarian or by pointing.
	Can follow the gist of short stories and tales in English with the help of illustrations.
	Can follow picture descriptions in English.
	Can participate in 4–5 round games with peers.
	Can respond to questions in English by using body language, speaking in Hungarian or giving a one-word answer in English.
	Can join discourse in English by using body language or Hungarian or single words in English.
	Can use greetings, say thank you, agree and disagree in English (yes/no).
	Can say 4–5 rhymes, can sing 4–5 songs so that what is said or sung is comprehensible.
	Can repeat recurring words and expressions (chunks) in familiar stories/tales individually or with peers.
Reading	Can recognize the written form of familiar words.
	Can comprehend the meaning of familiar words.
	Can read words they have learnt aloud.
	Can spell aloud some of the words they have learnt.
Writing	Can copy the majority of familiar words.
	Can fill in missing letters in familiar words.
	Can write down letters to form words when dictated.

learning strategies are considered crucial; they are important across all skills and have to be developed systematically during the long process of learning English. Teachers should consciously focus on these strategies from the earliest stages of language development.

Learners should be able to

1. distinguish familiar words and expressions from unfamiliar ones;
2. guess meanings of words and expressions (in L1 and L2) in context by relying on their background knowledge of the world;
3. use visual and other contextual information for guessing meaning;
4. help their peers if they do not understand something;

Table 3 What can learners do at the beginner plus level?

Listening comprehension, speaking and interaction	Can follow simple instructions in English in familiar contexts and can respond through total physical response, using body language, facial expression, or by answering in a few words in English.
	Can participate in activities and tasks by closely following classroom instructions in English.
	Can comprehend the meaning of frequent expressions, requests, questions, descriptions and events in English.
	Can guess what is being described from hearing a short, simple definition/explanation in English and can say the word.
	Can follow short stories and tales in English with the help of illustrations.
	Can participate in 5–10 round games with peers.
	Can respond to questions in English by using a few words in English.
	Can join discourse in English by using a few words in English.
	Can express agreement, disagreement, and make requests with a few words in English.
	Can ask simple questions in English.
	Can say 5–10 rhymes and can sing 5–10 songs so that what is said or sung is comprehensible.
	Can repeat or utter recurring words and expressions (chunks) in familiar stories/tales with peers or individually.
	Reading
Can comprehend the gist of learnt words, expressions and sentences.	
Can read a familiar picture story aloud.	
Can read 4–5 familiar picture books.	
Writing	Can spell some of the familiar words.
	Can copy the majority of the familiar without mistakes.
	Can fill in missing letters in the familiar words.
	Can write down most of the familiar words after dictation with some help.
	Can write a few words about items in a picture.
	Can copy words from a list into the appropriate place in a short text (e.g., form, list).

5. ask for help if they do not understand something;
6. find in familiar texts (picture dictionary, story, description) what they cannot recall accurately;
7. check their own performances;
8. evaluate their own performances; and
9. self-correct their mistakes, if necessary.

Table 4 What can learners do at the elementary level?

Listening comprehension, speaking and interaction	Can follow classroom instructions accurately.
	Can ask for specific information if something is not clear.
	Can ask questions and ask for help if something is unclear.
	Can comprehend the gist of rhymes, songs and games.
	Can comprehend the gist and sequence of actions in stories, tales, cartoons, and films, with the help of visuals.
	Can comprehend peers' roles in role plays and can react to them.
	Can guess the meaning of new words and expressions in context.
	Can give short and appropriate answers to short questions in context.
	Can ask short, simple questions with a little help.
	Can tell a short story with the help of pictures and questions.
	Can act out a short dialogue with peers or with the teacher rarely switching to Hungarian.
Can ask a question or help peers when they are stuck.	
Reading	Can comprehend the gist of short, familiar texts.
	Can read a few familiar picture books aloud.
	Can read 5–10 picture books and stories.
	Can comprehend the gist of short descriptions, dialogues, and stories including some new words.
Writing	Can find specific information in a simple, unknown text.
	Can copy familiar words and short sentences correctly.
	Can write down most of the familiar words after dictation and can check if the spelling is correct.
	Can write down a short simple text quite accurately as it is being dictated.
	Can write simple, short sentences about items in a picture. Can fill in a form with personal data.

5.3 Target Culture–Related Areas

As the Hungarian NCC (2007) includes hints at what young learners should know about the target language cultures, it seemed reasonable to include some guidance in this domain at the three levels (Table 5).

5.4 Themes and Topics

The following themes and topics were typically found in teaching materials and considered relevant for developing diagnostic tests

Table 5 What young learners should be familiar with in the target language cultures

Beginner	Learners know some English rhymes, songs, games, stories, and tales.
	They know 1–2 holidays and customs related to L2 cultures.
Beginner plus	Learners know several English rhymes, songs, games, stories, and tales.
	They know 3–4 holidays and customs related to L2 cultures.
	They are familiar with a few objects, expressions, books, and places related to the L2 culture.
	They know that English speaking cultures are different from Hungarian culture in a few areas.
Elementary	Learners are familiar with a few objects, expressions, stories, tales, heroes, and places related to the L2 cultures.
	They know of many ways in which English speaking cultures are both similar to and different from Hungarian culture.

- The natural world: plants, animals, people
- Personal identification, appearance
- Family and friends
- Home, house, housework, hobbies, play and games
- School and study
- Time, days, months, seasons, weather
- Shopping and services
- Vehicles, transport, traffic, travel, holidays
- Daily routine, hobbies, sports, free time
- Professions and jobs
- Health and food
- Feelings and opinions
- Social events (parties, customs, holidays)
- Places: city, country, village, mountains, rivers, lakes and seas

5.5 *Types of Texts*

- Rhyme, poem, song, game
- Picture story, cartoon
- Fairy tale, adventure and detective story
- Sign in shops, markets, streets, parks
- Label, notice
- Advertisement, booklet
- List, instruction manual
- Menu, recipe
- Letter, card, email message, text message
- TV program, guide
- Textbook (excluding EFL)
- Newspaper, magazine,

- Website, blog
- Dialogue and conversation
- Telephone conversation
- Interview
- Oral description
- Announcement

5.6 Task Types Appropriate for Diagnostic Assessment of Young Learners

This final section includes the task types recommended for the assessment of young learners in their four skills. Some general principles were agreed on. All tasks should include an example (the first item). In all multiple matching tasks there are one or two more options than necessary. All multiple choice tasks include four options. Most tasks include six to nine items. No task should take longer than 5–7 min. All performances on tasks can be quantified. Children should get feedback on their performances right after taking the task. All tasks are appropriate for teaching as well as diagnostic assessment.

5.6.1 Listening Comprehension

A total of 26 task types were identified. Some are variations, for example, one version is multiple choice, and the other one is multiple matching. Some tasks integrate other skills with listening.

1. Listen and do. Listen to the instructions and do what you are asked to do. Voice on tape gives instructions and students act accordingly.
2. Listen and do. Listen to the instructions. Color the pictures according to what you hear.
3. Listen and do. Circle the things you hear in the instructions.
4. Listen and point. Point to the items you hear (separate pictures or realia placed in various places in the classroom or on a worksheet).
5. Listen and point: point to the items you hear in a larger picture (e.g., large picture showing scene with details).
6. Listen and tick what you hear: tick the items you hear on a worksheet (words or short sentences).
7. Listen to numbers and put them down.
8. Listen and *write* down words spelt out. (integrated with writing)
9. Listen to short definitions and choose which picture they match.
10. Listen to short definitions and guess what they mean by putting a number next to or crossing the item in a picture (large ones with details).
11. Listen to short dialogues and choose where the dialogues take place (multiple choice items of small pictures on of? places).

12. Listen to short dialogues and choose where the dialogues take place (multiple matching)
13. Listen to short dialogues and choose who are talking (multiple choice items of small pictures).
14. Listen to short dialogues and choose who are talking (multiple choice items of short texts).
15. Listen to picture descriptions. Choose what or who they are talking about in the pictures. Multiple matching of pictures.
16. Listen to picture descriptions. Choose what or who they are talking about in a picture. Multiple matching of words or expressions.
17. Listen to a picture description and look at short sentences about the picture. There is a mistake in every item. Correct the mistakes. (Integrating listening, *reading, writing.*)
18. Listen to short dialogues and choose the correct answers from options 1, 2, 3 or 4. (Items on specific information.)
19. Listen to a short dialogue and tick the things you hear.
20. This is a picture dictation task. Listen and draw a picture of what you hear.
21. Fill in chart, diary, timetable, or number according to the information you hear. *Write down words and numbers in context.*
22. Listen to short definitions and guess what they mean. Choose words from a list (multiple matching).
23. Listen to short definitions and guess what they mean. *Put down the words.*
24. Listen to a story and look at the pictures. Correct the mistakes you hear and put down the correct versions. (e.g., in text: three monkeys are going for a walk; in picture: two. In text man is happy, in picture unhappy.) *Writing words.*
25. Listen to a short story and look at the pictures. Match the pictures with what you hear by putting the number in the box in the picture.
26. Listen to a picture description. Something is wrong in every sentence. Correct the mistakes by filling in *not, but* *Writing words.*

5.6.2 Speaking

1. Look at this picture and answer my questions. (Is this a? Are there any ...? How many? What is this? What is the bear doing? Where is the?)
2. Tell a rhyme or sing a song (in small group, or pairs, or individually).
3. Here are some picture cards facing down. Guess what's on them. Ask questions. (Child is to guess what is in the picture cards not seen by asking, for example, Is it a fruit? Is it an animal? Is it green? Does it have two legs? Limited choice of items known to children. Another variation: instead of picture cards children guess what the objects are under a cover.)
4. Look at this picture. I'm thinking of one of thes (vehicles, plants, animals, people, objects...). Ask me questions and guess what I'm thinking of. (Is it a...? Is it yellow? Is it next to? Limited choice of items known to children... in context. E.g.: an object in a kitchen, a room in a house, a person in a crowded street or park, a fruit at a market...)

5. Look at these two pictures. One is mine, the other one is yours. They are not the same. There are X differences. Let's find them. (I start by saying e.g., In my picture there are two houses. How about your picture? or In my picture there are three. In my picture a boy is going home....) Task in pairs (first with teacher). Both of you can see both pictures.
6. Here are two pictures (facing down). One is mine, the other one is yours. They are not the same. There are X differences. Let's find them. Let me start: (e.g., how many cars are there in your picture? Person A asks a question, B answers it. Then B asks a similar question. How many dogs are in your picture? What color is the biggest....?). You cannot see one another's pictures.
7. Look at this picture (with several items like in a picture dictionary). Tell me five things you like to eat and five you don't. or Name four yellow things and three red items, or five animals and five objects... in the picture. (One- or two-word answers are expected.)
8. Short role play in pairs. (E.g., You are at the market. You have X pounds and you'd like to buy three things. Look at the picture and the prices. OR Act out role play on a topic or from a story. Exchange 4–5 utterances. E.g., shopping, asking the way, offering food at birthday party, packing for holiday, school, ...)
9. Ask and answer personal/interview questions in pairs. Look at your cards with a (famous) character on it. (Some data are written: Name, age, address, number of sisters, brothers, pets, hobbies, etc.: What's your name? How old are you? Where do you live?)
10. This is a board game played by 2–4 learners. They use two dice and a list of (11 or 36) questions (personal or quiz). Questions should be written one by one on numbered cards (random choice). Throw two dice and add up (2–12) or multiply (1–36) numbers on top of dice. Person throwing dice must answer the question of that number on a list. This can be a paired task or 3–4 students can take turns. *Reading* the questions is also part of the task. One person throws dice, other person reads question, next one answers, etc., for example:
 - What's your friend's name?
 - Could you spell your surname, please?
 - What's your favorite school subject?
 - What subject do you dislike if any?
 - What's your favorite food?
 - What are your hobbies?
 - How many sisters or brothers do you have?
 - What does your (older) sister/brother do?
 - What pets do you have?
 - What TV programs do you watch?
 - How often do you do sports? Etc.
11. This is a paired task. Think about a famous person. Introduce the person by telling five important things about them (their age, nationality, hobbies, where they live, etc.). The other person should guess who it is. Then it is his/her turn.

12. Students choose one picture (from picture dictionary) of a choice of, for example, six. They are asked the following questions: 1. Please, describe the picture you chose. What can you see in it? 2. Who are the people in the picture? 3. What are they doing? 4. How is this home (kitchen/garden/town/village/supermarket) similar to your home (kitchen, etc.)? 5. What are the differences between your home (kitchen, etc.) and the home in this picture?
13. This is a paired task. There is a list of 99 questions and slips of numbers with 1–99 on them. Students take turns picking numbers from slips facing down, read the question corresponding to the number on the list and they answer it. Then they take turns. It could be also used with an adult interlocutor.
14. Describing pictures to one another. Students work in pairs. They both look at the same nine pictures (for example about a girl's hobbies). They take turns and their partners need to point to the picture they describe (so listening and speaking are integrated in the task).
15. Tell a story shown in pictures. For example, nine small pictures show a story: The Story of a Giraffe Family. This is a paired or individual task. By describing the pictures the story unfolds.

5.6.3 Reading Comprehension

1. Match pictures and words. Read out the words as you match them. Pictures and words are printed on one page in random order. It can be an individual or a paired task.
2. Match picture cards and word cards. Read out the words as you match them. It is a paired task.
3. Read out the words on word cards. Paired task with turn taking.
4. Find words with similar meanings. Read the words and find their synonyms in a list.
5. Find opposites of the words. Read the words and find their opposites in a list.
6. Read out familiar short sentences under pictures in a picture story. Reading aloud task.
7. Look at pictures and match them with short texts describing them.
8. Read short definitions/descriptions and match them with words.
9. Read the sentences and match them with pictures from the story.
10. Read out short instructions on slips one by one. Your pair should act accordingly. Drink your tea! Brush your hair! (reading aloud task)
11. Read questions of a short dialogue. Match them with answers to them (multiple choice).
12. Read questions of a short dialogue. Match them with answers to them (multiple matching).
13. Read a short text with a title. Answer questions by finding specific information in the text. Multiple choice short answers.
14. Read short texts with no titles for holistic understanding of texts. Choose titles from four options.

15. Read a short text. Answer questions on specific information in the text. Write short answers.
16. Read a short gapped text with missing words (form, invitation, letter, story, description). Fill in missing words from given list. Multiple matching – more items than gaps.
17. Read text with missing phrases/expression. Fill in missing phrases from given list. Multiple matching.
18. Read text with missing sentences. Fill in missing sentences from given list. Multiple matching.
19. Match the titles of books with pictures on book covers.
20. Match titles of books, stories, films with short ads or descriptions on them (about 20–30 words). Multiple matching task.
21. Match quiz questions (where, why, what, who, which, how, how many) with answers. Multiple matching task.
22. Match public signs with where they can be found. Multiple choice or multiple matching tasks.
23. Match short texts on postcards with pictures on them (where they come from, pictures of places, what people are doing, etc).
24. Read a text and complete a timetable or chart with the information in the text.
25. Read a text and fill in the missing information in a picture, map, or diagram.
26. Draw lines between words in a list and things in a picture (e.g., a bathroom or market).
27. Choose pictures showing the place where short written dialogues take place – multiple matching.
28. Choose places (cinema, swimming pool, at home) where short written dialogues take place – multiple matching.

5.6.4 Writing

1. Copy words in categories. Look at the list of nine words. Copy the words under the category (umbrella) where they belong. E.g.: foods and drinks; plants and animals; black, white, other colors.
2. Look at pictures and words in random order (e.g., fruits). Copy the names of the fruits under the pictures.
3. Fill in missing letters in words (1 line = 1 letter): ele_ _ ant, hors _ , crocod_ l_ , do_ , etc. Choose letters from the list: g, e, p, e, h,
4. Fill in missing letters in words: no letters are given, but, for example, all are drinks or animals.
5. Write down ten words after dictation. All of them are colors or part of the body, etc.
6. Write down five short sentences after dictation (text is a story or description with a title). Every sentence is dictated twice, then all once more.
7. Look at a picture of a house/park..... Some animals/people are hiding there. Finish sentences by adding words.

8. Fill in words in gapped story or description. Choose from list of items. Multiple matching.
9. Fill in words in timetable, chart, shopping list, where a lot of info is in place, the rest of items should be chosen from list (multiple matching) e.g.: school subjects, breakfast, lunch, dinner.
10. Read a short text. Answer questions with specific information in the text. Write short answers.
11. Fill in personal data in a form. Short text is given on person whose data are to be filled in. Integrating reading.
12. Picture description: write short sentences about a picture. For example, what are children doing in a park? Write as much as you can about what they are doing.
13. Picture description: compare two pictures. Write about five differences.
14. Write a short personal letter/card in response to a letter/card worded similarly.
15. Put down some information after dictation. E.g., shopping list.
16. Error correction, based on pictures (reading integrated). Look at the pictures and the sentences. Something is wrong in every sentence, correct them.
17. Write down what animals/vehicles/foods/drinks/sports you can see in the pictures.

6 Conclusions and the Way Forward

The aim of this chapter was to share findings of a national project implemented in Hungary. At the beginning stage, we looked for sources to draw on and found some useful materials and ideas; however, it took a lot of work and effort to design and create what we finally managed to come up with. Now that we developed a framework, a list of can do statements, topics and task types, and by doing so we have learnt a lot of lessons, we assume that colleagues developing frameworks and tests for young learners may be interested in them and after critically reviewing them, some of these ideas might be useful and relevant in other situations. We hope some of the outcomes can be adopted in new educational contexts and readers may find them relevant not only for EFL but also for other foreign languages.

The chapter gave insights into the outcomes of a diagnostic assessment project, where an assessment for learning approach was applied; the tasks, however, could be also considered for other assessment purposes. The chapter presented the most important characteristics of young language learners and how they learn a FL; it also outlined the main principles of assessing children. As was shown, based on the framework and the lists of can do statements, text types and task types, over 200 new diagnostic tests were developed and piloted in the second phase of the project. Findings were published in English on various aspects of the piloting phase involving a large sample of young learners and their teachers of EFL in the first few grades of primary schools. Publications explored teachers' views on tasks that work (Hild & Nikolov, 2011), how the tests were piloted and the difficulty levels were estab-

lished (Nikolov & Szabó, 2011a, 2011b; 2012a) and children's feedback on the actual tests was also analyzed (Szabó & Nikolov, 2013). As a next phase these calibrated diagnostic tests are going to be made available to teachers for their classroom use in the online database.

In addition to these ideas, the framework and the task types could be used in teacher education programs to explore to what extent they would meet the needs of children and their teachers in various contexts. Also, the actual tasks could serve as excellent materials for small-scale classroom research; both in-service and pre-service teachers could experiment with them and explore how they work with their learners in the specific contexts and why. The tasks could be further developed and similar tasks could be designed and piloted on new topics, etc. Finally, yet another perspective is offered for further classroom research by asking learners after doing tasks about the extent to which they liked or disliked, were familiar with, and found the tasks easy or difficult. By involving learners in these discussions after completing tasks teachers may gain valuable insights into their learners' experiences, they may be able to tailor their teaching to their needs, and they may also develop their young language learners' self-assessment and autonomy.

Acknowledgements Special thanks go to members of the team for developing the can do statements, drawing up the list of topics and task types based on curricula, teaching materials and research findings. The team included twelve anonymous classroom teachers with over a decade of teaching experience in Hungarian classrooms and the following experienced teachers of young learners and experts on assessment and early language learning research: Lidia Bors, Judit Font, Gabriella Hild, Csilla Kiss, Réka Lugossy, Ildikó Pathó, Gábor Szabó, Zsófia Turányi (Hungary) and four international experts: Heini-Marja Järvinen (Finland), Lucilla Lopriore (Italy), Jelena Mihaljević Djigunović (Croatia), and Karmen Prizorn (Slovenia)

I am grateful to Jelena Mihaljević Djigunović for her helpful comments on the first draft of this chapter.

References

- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alexiu, 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). Clevedon/Avon: Multilingual Matters.
- Assessment Reform Group. (2002). *Assessment for learning*. Retrieved from <http://www.assessmentforlearning.edu.au/default.asp>
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.
- Bacsá, 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*. New York: Springer.

- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–71.
- Bors, L., Lugossy, R., & Nikolov, M. (2001). Az angol nyelv oktatásának átfogó értékelése pécsi általános iskolákban [A comprehensive evaluation study of the teaching of English in Pecs primary schools]. *Iskolakultúra*, 11(4), 73–88.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Csapó, B., & Csépe, V. (2012). *Framework for diagnostic assessment of reading*. Budapest: Nemzeti Tankönyvkiadó.
- Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19, 203–218.
- Csapó, B., & Szendrei, M. (Eds.). (2011). *Framework for diagnostic assessment of mathematics*. Budapest: Nemzeti Tankönyvkiadó.
- Csapó, B., & Szabó, G. (Eds.). (2012). *Framework for diagnostic assessment of science*. Budapest: Nemzeti Tankönyvkiadó.
- Csapó, B., & Zsolnai, A. (Eds.). (2011). *A kognitív és affektív fejlődés diagnosztikus mérése az iskola kezdő szakaszában*. Budapest: Nemzeti Tankönyvkiadó.
- Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon/Avon: Multilingual Matters.
- Curtain, H. A., & Dahlberg, C. A. (2010). *Languages and children – Making the match: New languages for young learners* (4th ed.). Needham Heights, MA: Pearson Allyn & Bacon.
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43, 393–415.
- DeKeyser, R., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). Oxford: Oxford University Press.
- Dörnyei, Z., Csizér, K., & Németh, N. (2006). *Motivation, language attitudes and globalisation: A Hungarian perspective*. Clevedon, England: Multilingual Matters.
- Dweck, C. (2006). *Mindset: The new psychology of success*. New York: Ballantine Books.
- Eurobarometer. (2006). *Europeans and their languages*. Brussels: European Commission.
- Eurobarometer. (2012). *Europeans and their languages*. Brussels: European Commission.
- García Mayo, M. P., & García Lecumberri, M. L. (Eds.). (2003). *Age and the acquisition of English as a foreign language*. Clevedon/Avon: Multilingual Matters.
- Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing*, 22, 337–354.
- Hild, G., & Nikolov, M. (2011). Teachers' views on tasks that work with primary school EFL learners. In M. Lehmann, R. Lugossy, & J. Horváth (Eds.), *UPRT 2010: Empirical studies in English applied linguistics* (pp. 47–62). Pécs: Lingua Franca Csoport. Retrieved from <http://mek.oszk.hu/10100/10158>
- Hung, Y.-J., Samuelson, B. L., & Chen, S.-C. (2016). The relationships 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*. New York: Springer.
- Kiss, C., & Nikolov, M. (2005). Preparing, piloting and validating an instrument to measure young learners' aptitude. *Language Learning*, 55, 99–150.
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. New York: Longman.
- Leung, C., & Scott, C. (2009). Formative assessment in language education policies: Emerging lessons from Wales and Scotland. *Annual Review of Applied Linguistics*, 29, 64–79.
- MacWhinney, B. (2005). A unified model of language development. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford: Oxford University Press.
- McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell Publishing.

- Medgyes, P., & Nikolov, M. (2010). Curriculum development: The interface between political and professional decisions. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 264–274). Oxford: Oxford University Press.
- Medgyes, P., & Nikolov, M. (2014). Foreign language learning and teaching in Hungary: A review of empirical research literature from 2006 to 2012. *Language Teaching*, 47(4), 504–537.
- Mihaljević Djigunović, J. (2001). Do young learners know how to learn a foreign language? In Y. Vrhovac (Ed.), *Children and Foreign Languages III* (pp. 57–71). Zagreb: Faculty of Philosophy.
- Mihaljević Djigunović, J. (2009a). Impact of learning conditions on young FL learners' motivation. In M. Nikolov (Ed.), *Early learning of modern foreign languages. Processes and outcomes* (pp. 75–89). Bristol, UK: Multilingual Matters.
- Mihaljević Djigunović, J. (2009b). Individual differences in early language programmes. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 199–225). Berlin/New York: Mouton de Gruyter.
- Mihaljević Djigunović, J. (2012). *Early EFL learning in context – Evidence from a country case study*. London: The British Council.
- Muñoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon/Avon: Multilingual Matters.
- National core curriculum (NCC). (2007). Budapest: Oktatási és Művelődési Minisztérium.
- National core curriculum (NCC). (2012). Budapest: EMMI.
- 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. (2001). A study of unsuccessful language learners. In Z. Dörnyei & R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 149–170). Honolulu, HI: The University of Hawaii, Second Language Teaching and Curriculum Center.
- Nikolov, M. (2002). *Issues in English language education*. Bern: Peter Lang AG.
- Nikolov, M. (2003). Angolul és németül tanuló diákok nyelvtanulási attitűdje és motivációja [Attitudes and motivation of English and German learners]. *Iskolakultúra*, XIII(8), 61–73.
- Nikolov, M. (2008). “Az általános iskola, az módszertan!” Alsó tagozatos angolórák empirikus vizsgálata [“Primary school means methodology!” An empirical study of lower-primary EFL classes]. *Modern Nyelvoktatás*, 10(1–2), 3–19.
- 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). Clevedon/Avon: Multilingual Matters.
- Nikolov, M. (2011). Az angol nyelvtudás fejlesztésének és értékelésének keretei az általános iskola első hat évfolyamán [A framework for developing and assessing proficiency in English as a foreign language in the first six years of primary school]. *Modern Nyelvoktatás*, XVII(1), 9–31.
- Nikolov, M., & Csapó, B. (2010). The relationship between reading skills in early English as a foreign language and Hungarian as a first language. *International Journal of Bilingualism*, 14, 315–329.
- Nikolov, M., & Józsa, K. (2006). Relationships between language achievements in English and German and classroom-related variables. In M. Nikolov & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 197–224). Pécs: Lingua Franca Csoport, PTE.
- 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. (2011a). Az angol nyelvtudás diagnosztikus mérésének és fejlesztésének lehetőségei az általános iskola 1–6. évfolyamán [Possibilities of developing English diagnostic tests for years 1–6 in the primary school]. In Csapó Benő & Zsolnai Anikó (szerk.) *A kognitív*

- és affektív fejlődés diagnosztikus mérése az iskola kezdő szakaszában* (pp. 13–40). Budapest: Nemzeti Tankönyvkiadó.
- Nikolov, M., & Szabó, G. (2011b). 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: Lingua Franca Csoport. Retrieved from <http://mek.oszk.hu/10300/10396>
- Nikolov, M., & Szabó, G. (2012a). 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* (pp. 347–363). Cambridge: UCLES/Cambridge University Press.
- Nikolov, M., & Szabó, G. (2012b). Assessing young learners' writing skills: A pilot study of developing diagnostic tests in EFL. In G. Pusztai, Z. Tóth, & I. Csépes (Eds.), *Current research in the field of disciplinary didactics* (Hungarian Educational Research Journal, Special Issue, Vol. 2, pp. 50–62). Retrieved from <http://herj.hu/2012/08/marianne-nikolov-and-gabor-szabo-assessing-young-learners%E2%80%9999-writing-skills-a-pilot-study-of-developing-diagnostic-tests-in-efl/>
- Papp, S., & Salamoura, A. (2009). *An exploratory study into linking young learners' examinations to the CEFR* (Research Notes, 37, pp. 15–22). Cambridge: Cambridge ESOL.
- Paradis, M. (2004). *A neurolinguistic theory of bilingualism*. Amsterdam: John Benjamins.
- Paradis, M. (2009). *Declarative and procedural determinants of second languages*. Amsterdam: John Benjamins.
- Pinter, A. (2006). Verbal evidence of task-related strategies: Child versus adult interactions. *System*, 34, 615–630.
- Pinter, A. (2007a). Benefits of peer-peer interaction: 10-year-old children practising with a communication task. *Language Teaching Research*, 11, 189–207.
- Pinter, A. (2007b). What children say: Benefits of task repetition. In K. Van den Branden, K. Van Gorp, & M. Verhelst (Eds.), *Task-based language education from a classroom-based perspective* (pp. 126–149). Cambridge: Cambridge Scholars Publishing.
- 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: Cito/EALTA.
- 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*. New York: Springer.
- Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.
- Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics*, 20, 213–223.
- Singleton, D., & Ryan, L. (2004). *Language acquisition: The age factor* (2nd ed.). Clevedon/Avon: Multilingual Matters.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
- 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* (pp. 7–21). Zagreb: FF Press. Retrieved from http://books.google.hu/books?id=&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge: Cambridge University Press.
- Swain, M. (2000). The output hypothesis and beyond: Mediating acquisition through collaborative dialogue. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 97–114). Oxford: Oxford University Press.

- Teasdale, A., & Leung, C. (2000). Teacher assessment and psychometric theory: A case of paradigm crossing? *Language Testing*, 17(2), 163–184.
- Ullman, M. (2001). The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, 4, 105–122.
- 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*. New York: Springer.
- Wong Fillmore, L. (1998). *Supplemental declaration of Lily Wong Fillmore*. Retrieved from <http://www.humnet.ucla.edu/humnet/linguistics/people/grads/macswan/fillmor2.htm>

Examining Content Representativeness of a Young Learner Language Assessment: EFL Teachers' Perspectives

Ching-Ni Hsieh

Abstract This study aims to provide content validity evidence for the new young language learner assessment—TOEFL Primary—a test designed for young learners ages 8 and above who are learning English in English as a Foreign Language (EFL) contexts. The test focuses on core communication goals and enabling language knowledge and skills represented in various EFL curricula. A panel of 17 experienced EFL teachers, representing 15 countries, participated in the study. The teachers evaluated the relevance and importance of the knowledge, skills, and abilities (KSAs) assessed by the reading and listening items of TOEFL Primary. Content Validity Indices (CVIs) (Popham, *Appl Meas Educ* 5(4):285–301, 1992) was used to determine the degree of match between the test contents and the target constructs and the importance of the KSAs assessed for successful classroom performance. Results showed that the majority of the items had an average CVI above the cut-off value of .80, indicating that the items measured what they were intended to measure and that the KSAs assessed were important for effective classroom performance, supporting the claim about using the test scores to support language teaching and learning.

Keywords Content validity • TOEFL Primary • Young learners • Language assessments • Teacher judgments • Language teaching

1 Introduction

Measuring and reporting content validity of newly developed tests is important because this type of validity evidence provides test users essential information regarding the extent to which test contents reflect the target constructs being

C.-N. Hsieh (✉)

Center for English Language Learning and Assessment, Research and Development,
Educational Testing Service, Princeton, NJ, USA

e-mail: chsieh@ets.org

© Springer International Publishing Switzerland 2016

M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*, Educational Linguistics 25, DOI 10.1007/978-3-319-22422-0_5

93

measured and the validity of the inferences drawn from the test scores (D'Agostino, Karpinski, & Welsh, 2011; Haynes, Richard, & Kubany, 1995; So, 2014; Yalow & Popham, 1983). The study reported here examines the degree of content representativeness within the context of a new young learner language assessment, TOEFL Primary, with the goal of providing an important piece of content validity evidence for the test.

As the number of young English language learners worldwide continues to grow, so too does the need for language assessments designed to target this population (McKay, 2006; Nikolov, 2016, in this volume). While several language assessments have been developed to serve the needs of these learners (e.g., Cambridge English: Young Learners English Tests; TOEFL Primary; TOEFL Junior), theoretical and empirical knowledge about the assessment of young language learners remains underdeveloped. For instance, relatively little is known about the target language use (TLU) domains for English communication among young learners. What is clear, however, is that language tasks designed for young learners need to take into consideration factors such as learners' shorter attention span (Robert, Borella, Fagot, Lecerf, & De Ribaupierre, 2009), memory capacity (Cho & So, 2014), longer processing time (Berk, 2012), developing literacy, and limited exposure to and experience of the world—factors that are distinct from those relevant to the assessments of adult learners of English as a Second (or Foreign) Language (ESL/EFL). Given these differences, it is critical for language test developers and researchers to better comprehend how the test contents of young learner assessments reflect and meet the communication needs of young learners and how individual characteristics of students should influence test design.

TOEFL Primary is a new young learner language assessment developed by Educational Testing Service (ETS). The test is designed for young learners ages eight and above who are learning English in EFL contexts. The test measures three English language skills: listening, reading, and speaking. Listening and reading are offered in two steps, i.e. Step 1 (low level) and Step 2 (high level), to reflect the wide range of language proficiency exhibited among the target population. The speaking test is designed for language learners at many different proficiency levels of English, from beginners to more proficient speakers, and thus is not separated into different steps. The test items of TOEFL Primary cover a set of communication goals, a range of difficulty, and various item types. The test is intended to support language teaching and learning by providing meaningful information for the test takers' current English ability. EFL teachers can use the test to guide their teaching goals, monitor student progress, and identify students' strengths and weaknesses in different areas of language use. The test scores can also be used for placement purposes if the test content corresponds to or is relevant to the content of the EFL curriculum that the students are exposed to. However, the test is not intended to support high-stakes decisions such as to inform admission decisions or to evaluate teachers' performances.

2 Literature Review

The link between test content and EFL curricula is an important facet in establishing content validity for tests that are developed to provide instructional support. Two studies that examined the relationships between test contents and course contents (Fleurquin, 2003; Wu & Lo, 2011) have specific implications for the current study. Fleurquin reported the process of developing and validating *Alianza Certificate of Elementary Competence in English* (ACECE), a standardized test of American English that measures young learners' English communication skills within the context of elementary schools in Uruguay. To examine content validity of the ACECE, the research team enlisted experienced EFL teachers to compare the grammar structures and vocabulary categories assessed in the test with the contents of three textbooks used with the target population in local schools. The comparison showed that the majority of the grammar structures and vocabulary assessed in the test matched those presented in the textbooks that the students had used during their school years, providing evidence to support the alignment between the content of the ACECE and the three textbooks. Specific comments regarding the test items and stimulus materials provided by the EFL teachers were also used to inform test revisions.

Wu and Lo (2011) investigated the relationship between a standardized English language proficiency test for young children, the Cambridge English: Young Learners English (YLE) Tests, and the EFL teaching practices at the elementary level in Taiwan. The study aimed to inform local teachers regarding whether the YLE tests were suitable for young learners in Taiwan. The researchers compared the Grades 1–9 Curriculum Guidelines published by the Ministry of Education in Taiwan and a popular series of English textbooks published by a local publisher with the content of the YLE. The comparison was conducted in six aspects: topics, grammar and structures, communication functions, competence indicators, vocabulary, and tasks. Results showed a moderate to high degree of alignment between the YLE and the local teaching practices with regard to the six aspects of the comparison and highlighted a gap between the two in terms of cultural differences between Taiwan and the UK as manifested in the wordlists introduced. Taken together, the use of expert teacher judgments in Fleurquin (2003) and Wu and Lo (2011) has proven useful in helping researchers and test developers determine content alignment between young learner language assessments and EFL curricula in different EFL contexts and identify aspects of misalignment to inform test revisions.

It needs to be noted that in content validation studies that use expert judgments, a criterion (i.e., cut-off point) is required to ensure the quality of the judgments. While both Fleurquin (2003) and Wu and Lo (2011) used expert teachers to evaluate the alignment between test content and local teaching practices, neither study employed a definite cut-off value, leaving open a determination of the test's content representativeness. Since one major purpose of content validation studies is to ensure that the test contents reflect what they are intended to measure, a criterion for making that decision is critical to represent the quality of the test content. The more

stringent the criterion is, the more confidence that can be placed in positive appraisals of the test content (Popham, 1992).

In this study, I examined the content representativeness of TOEFL Primary using a traditional content validity approach based on the computation of a Content Validity Index (CVI) (Davis, 1992; Lynn, 1986) with a predetermined criterion. The CVI approach entails a panel of expert judges evaluating whether the relevance of each test item on an assessment instrument is relevant to the target construct being measured. The percentage of items rated as relevant by each judge and the average of the percentages across the judges are reported as an indication of the degree of “content validity”, or more appropriately, content representativeness in this case. The use of CVIs to determine content representativeness is widely cited in test development literature for teacher licensure tests (Crocker, Miller, & Franks, 1989; Popham, 1992), nursing research (Davis, 1992; Polit & Beck, 2006) and social work research (Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003), but to the best of my knowledge, they have not been widely used for tests of second language proficiency.

3 Content Validation of TOEFL Primary

During the initial stage of test development of TOEFL Primary, the researchers and test developers at ETS had set out to conduct a two-stage process for establishing the content validity of the test (Lynn, 1986; Sireci, 1998, 2007). The first stage, or ‘Developmental Stage,’ involves the identification of the domain of content through a comprehensive review of relevant literature and domain analysis of language use in EFL classrooms—the TLU domain. The domain descriptions were enhanced by the development team’s review of EFL curricula and textbooks used in nine countries, including Brazil, Chile, China, Egypt, Japan, Korea, the Philippines, Qatar, and Singapore (Turkan & Adler, 2011). Results of the domain analysis helped define the construct of English communication for young learners. A set of communication goals that are unique to young learners’ communicative needs and the language knowledge and skills required to fulfill these communication goals are incorporated in the construct definitions. The communication goals targeted also helped test developers identify specific text types that young learners encounter in their EFL reading and listening materials and the various types of speaking activities that young learners engage in the EFL classrooms. A variety of test tasks associated with specific communication goals are developed for the test.

The second stage of content validation, the ‘Judgment/Quantification’ stage of content validation (Lynn, 1986), for TOEFL Primary is twofold, involving a teacher survey on the pilot-test items and a panel judgment of the operational test items—i.e. the current study. During pilot testing of TOEFL Primary, a teacher survey study was conducted at local testing sites where TOEFL Primary was piloted. The survey aimed to gather EFL teachers’ feedback on the importance and relevance of the set

of communication goals identified for construct definitions and the appropriateness and effectiveness of the task types proposed for young EFL learners. Results of the teacher survey, which contained the evaluations of test contents by 29 EFL teachers from Costa Rica, Egypt, Japan, Peru, and Vietnam, showed that the communication goals substantially reflected the communicative needs of young learners. The survey also revealed varying views regarding the effectiveness of the task types, which subsequently informed the subsequent refinement of the tasks (Hsieh, 2013).

The current study focused on the panel judgment of the TOEFL Primary operational listening and reading items in terms of their content relevance and the importance of the language knowledge, skills, and abilities (KSAs) assessed in these items for successful classroom performance. The study was informed by the body of literature that uses CVIs to determine the degree of content representativeness for newly developed language assessments. Predefined cut-off values suggested by the collective body of literature (e.g., Davis, 1992; Lynn, 1986) were adopted for determining whether test items were congruent with the constructs being measured and whether the KSAs assessed reflected those introduced in a number of EFL contexts. The use of CVIs to assess the degree of agreement among the EFL teachers has the benefit of allowing better comparability between the judgments gathered by different content validity studies.

The study aimed to address the following research questions:

1. To what extent do TOEFL Primary listening and reading test items reflect the target constructs as judged by EFL teachers?
2. What are EFL teachers' perceptions of the importance of the KSAs assessed by TOEFL Primary in their specific teaching contexts?

4 Method

4.1 Participants

A panel of 17 EFL teachers served as the expert judges in this study. The panel of judges was formed, to the extent possible with a relatively small sample, to have representation by gender, professional background, and geographic location. Participants were selected from a large pool of EFL teachers based on their expertise in young learner EFL curricula and professional experience. All teachers had experience teaching young learners similar to the target population for TOEFL Primary, i.e. ages eight and above. Fifteen countries (Brazil, China, France, Greece, Japan, Jordan, Kazakhstan, Mexico, Peru, Russia, Slovakia, South Korea, Spain, Sweden, and Vietnam) were represented. The teachers were between the ages of 25 and 52 (*Mean* = 38.9, *SD* = 7.3). Their years of teaching EFL ranged from 3 to 29 years (*Mean* = 14.9 years, *SD* = 7.0). Table 1 shows the demographic information of the teachers.

Table 1 Demographic information of the participating teachers

Educational background	N	%
College	5	29 %
Some postgraduate education	2	12 %
Master	8	47 %
PhD	2	12 %
Year of teaching		
Below 10 years	4	24 %
10–20 years	9	53 %
21–30 years	4	24 %
Age		
20s	2	12 %
30s	5	29 %
40s	8	47 %
50s	2	12 %
Gender		
Male	4	24 %
Female	13	76 %
Geographical region		
Asia & The Middle East	6	35 %
Europe	7	41 %
Latin America	4	24 %

4.2 Rating Materials

The rating materials used in this study consisted of operational listening ($N=57$) and reading ($N=57$) test items of TOEFL Primary. These items were carefully chosen by the test developers at ETS to cover all the targeted communication goals of TOEFL Primary, the full range of difficulty, and all item types (see Table 2). The number of items per item type reflected that of the operational form. The total number of the listening and reading items included in the study was larger than the number in an operational form because these items covered the two difficulty levels of TOEFL Primary. The inclusion of items from both steps was considered important to ensure a comprehensive coverage of the difficulty range of the test. Including more items in the study was also thought to produce more stable judgments overall. The speaking section was not included in the study due to time and resource constraints in data collection.

4.3 Instrument

A content alignment questionnaire for item evaluation was constructed by the researcher through consultation with ETS test developers and research scientists who were experienced with content alignment studies. The instructions to

Table 2 TOEFL Primary listening and reading items for evaluation

Listening item type	Communication goal	Step	N
Listen and match	Understand simple descriptions of familiar people and objects	1	7
Follow instructions	Understand spoken directions and procedures	1, 2	10
Question/response	Understand dialogues or conversations	1	6
Dialogue	Understand dialogues or conversations	1, 2	10
Social-navigational monologue	Understand short informational texts related to daily life	1, 2	10
Narrative set	Understand spoken narratives	2	8
Academic monologue	Understand expository monologues	2	6
Reading item type	Communication goal	Step	N
Match picture to word	Identify people, objects and actions	1	6
Match picture to sentence	Identify people, objects and actions	1	7
Sentence clues	Understand written expository or informational texts	1, 2	12
Telegraphic sets	Understand commonly occurring non-linear written texts (e.g. signs, schedules)	1, 2	8
Correspondence	Understand short personal correspondence	1, 2	6
Instructional texts	Understand written directions and procedures	2	6
Narrative sets	Understand simple, written narratives	2	8
Expository paragraph	Understand written expository or informational texts about familiar people, objects, animals, and places	2	4

participants during the alignment exercise, the questionnaire response formats and scales underwent multiple rounds of trials and revisions prior to data collection. The final survey instrument consisted of two subsections. Section I included seven parts, each corresponding to one listening item type. Section II included eight parts, each corresponding to one reading item type. The KSAs assessed in each item type were provided in the questionnaire to facilitate the evaluation process.

4.4 Procedures

The 17 EFL teachers were invited from their countries to ETS campus in Princeton, New Jersey, to participate in the study. Each teacher was supplied with (a) a background questionnaire that was used to gather the teachers' biographical information, (b) a test booklet that contained the 57 listening and 57 reading test items, (c) a copy of the scripts for the listening items, and (d) the content alignment questionnaire for the evaluation of the test items. Prior to the day of the content alignment exercise, all teachers took the TOEFL Primary test and reviewed documents on the test design framework and scoring guidelines to become familiar with the test constructs, design, and scoring criteria. On the day of data collection, the teachers first completed the background questionnaire and then were instructed to make

judgments on two aspects of the content representativeness of each item using the content alignment questionnaire. The two aspects were content relevance of and the importance of the KSAs assessed by the TOEFL Primary test items. In addition to the content alignment exercise, five teachers (from France, Jordan, Mexico, Peru, and Spain) agreed to participate in follow-up interviews that were conducted after the analyses of the rating data. The interviews focused on (1) the teachers' views about specific aspects of the test contents that the teachers considered less important or relevant to their own teaching practices and (2) how the teachers used the different types of texts and item types in their respective EFL classrooms.

4.5 Content Alignment Judgments

The two aspects of content alignment judgments the teachers were asked to perform are described as follows.

(1) Content relevance

The first judgment asked the teachers to evaluate the degree to which the content of each item reflected the target construct it is intended to measure. Congruent with Lynn's (1986) item relevance rating rules, judges were asked to provide the relevance ratings on a Likert scale with four possible responses: *no reflection*, *slight reflection*, *moderate reflection* and *strong reflection*. Responses of 'moderate reflection' and 'strong reflection' were regarded as indications of teachers' endorsement of the content relevance of the items, whereas responses of 'no reflection' and 'slight reflection' indicated the opposite. The responses were dichotomized in this fashion in order to facilitate summary evaluations.

(2) The importance of the KSAs assessed

The second judgment required the teachers to rate the importance of the KSAs required of young EFL learners for successful classroom performance in their own teaching contexts. The importance ratings, also on a 4-point Likert scale (Lynn, 1986), had four different labels: *not important*, *somewhat important*, *important* and *very important*. Responses of 'important' and 'very important' indicated teachers' agreement on the importance of the KSAs assessed, whereas responses of 'not important' and 'somewhat important' indicated the opposite. As with the content relevance ratings, the importance ratings were also dichotomized.

4.6 Analysis

To answer the research questions, individual ratings provided by the 17 judges were pooled and the CVIs for each item were calculated for evaluating the degree of content relevance and importance of the KSAs assessed in the TOEFL Primary test

items (Davis, 1992; Lynn, 1986; Polit & Beck, 2006). The analyses of the degree of content representativeness of the test items are described below.

(1) *CVIs for content relevance*

For the content relevance ratings, the CVI for each item was calculated by counting the number of judges who rated that item as either 'moderate reflection' or 'strong reflection' and dividing that number by the total number of judges. The CVI calculated for each item provided information about the proportion of judges who considered an item as content relevant. The CVIs for the listening and reading sections were defined as the proportion of items on the section that achieved a rating of 'moderate reflection' or 'strong reflection' across all judges. The CVIs for listening and reading sections were derived, respectively, by averaging the CVIs across the 57 items for each section.

(2) *CVIs for the importance of the KSAs assessed*

For the importance of the KSAs assessed, the CVI for each item was calculated by counting the number of judges who rated the item as either 'important' or 'very important' and dividing that number by the total number of judges. The CVI calculated for each item provided information about the proportion of judges who considered the KSAs assessed by an item as important for successful classroom performance. The CVIs for the listening and reading sections were defined as the proportion of items on the section that achieved a rating of 'important' or 'very important' across all judges. The CVIs for listening and reading sections were derived, respectively, by averaging the CVIs across the 57 items for each section.

To determine the degree to which TOEFL Primary test items reflect the target constructs and assess the important KSAs required of young learners, a CVI of .80 was used as the acceptable criterion, following Davis (1992). This criterion is widely used in the literature for determining content representativeness of new assessments (e.g., Rubio et al., 2003). This cut-off value indicates that, when a total of 17 judges are considered, at least 14 agree that the items reflect the intended target constructs or that the KSAs assessed are important for successful classroom performance.

5 Results

5.1 Results of the Content Relevance Ratings

Descriptive statistics of the content relevance ratings and the average CVIs for each item type are provided in Table 3. As the table shows, all listening item types had an average CVI above .80. The CVI for the Listening section was .95, clearly above the cut-off criterion. Similarly, all the reading items and item types had a CVI above the cut-off value of .80. The CVI for the Reading section was .95, indicating excellent content relevance.

Table 3 Descriptive statistics and average CVIs for content relevance

Listening item type	Mean	S.D.	CVI
Listen and match	3.66	0.18	0.94
Follow instructions	3.89	0.69	0.97
Question/response	3.45	0.22	0.94
Dialogue	3.48	0.12	0.95
Social-navigational monologue	3.55	0.13	0.93
Narrative set	3.72	0.12	0.94
Academic monologue	3.77	0.07	0.97
Reading item type	Mean	S.D.	CVI
Match picture to word	3.62	0.05	0.89
Match picture to sentence	3.74	0.15	0.95
Sentence clues	3.71	0.13	0.96
Telegraphic sets	3.51	0.14	0.95
Correspondence	3.73	0.11	0.96
Instructional texts	3.74	0.13	0.97
Narrative sets	3.68	0.12	0.93
Expository paragraph	3.79	0.03	1.00

5.2 Results of the Importance of the KSAs Assessed

Descriptive statistics of the importance ratings and the average CVIs for each item type are provided in Table 4. The table shows that six listening item types had an average CVI above .80, with the exception of ‘Academic Monologue.’ The ‘Academic Monologue’ item type is only present in Step 2 of TOEFL Primary. The item type requires test takers to listen to a monologue spoken by a teacher or another adult instructing academic content to students. The test takers then answer three multiple-choice comprehension questions. These questions assess the students’ abilities to understand spoken informational texts and require test takers to have knowledge of organization features of expository texts and the ability to understand key information in a monologue.

A similar degree of agreement among the judges is seen in the Reading section. The majority of the reading item types had a CVI above .80, with the exception of ‘Telegraphic Sets’ that had a borderline CVI of .79. The ‘Telegraphic Sets’ item type is present both in Step 1 and Step 2 of TOEFL Primary. This item type asks test takers to answer multiple-choice questions by locating the relevant information in telegraphic texts in which language is presented in single, phrasal, and short sentence form. Commonly used stimulus materials include posters, menus, schedules, and advertisements. The slightly lower CVI of .79 was considered negligible given that the majority still rated the KSAs assessed in the ‘Telegraphic Sets’ important.

To summarize, the results of the importance of the KSAs assessed by TOEFL Primary indicate high agreement among the judges. The Listening and Reading sections both had an average CVI of .89, suggesting that the majority of the teachers

Table 4 Descriptive statistics and average CVIs for the importance of the KSAs assessed

Listening item type	Mean	S.D.	CVI
Listen and match	3.55	0.22	0.94
Follow instructions	3.55	0.14	0.92
Question/response	3.37	0.18	0.82
Dialogue	3.55	0.07	0.96
Social-navigational monologue	3.61	0.09	0.90
Narrative set	3.70	0.11	0.95
Academic monologue	3.26	0.05	0.72
Reading item type	Mean	S.D.	CVI
Match picture to word	3.69	0.05	0.91
Match picture to sentence	3.76	0.12	0.97
Sentence clues	3.61	0.14	0.92
Telegraphic sets	3.79	0.93	0.79
Correspondence	3.48	0.11	0.84
Instructional texts	3.50	0.09	0.86
Narrative sets	3.68	0.12	0.97
Expository paragraph	3.49	0.07	0.88

considered that the KSAs assessed were important for their respective language teaching contexts.

6 Discussion

This study used CVIs as a research methodology to evaluate the degree of content representativeness of TOEFL Primary. A representative panel of experts was convened to evaluate the degree of match between the test construct and the content of the listening and reading items of the test and to evaluate the importance of the KSAs assessed. The expert teachers' judgments were used as the criterion on which the content-related evidence of validity was based. Results of the study suggest that TOEFL Primary test content largely reflects the target construct being measured and covers the important domains of language knowledge and skills EFL learners are required to possess in order to perform successfully in EFL classrooms.

The content alignment exercise performed by the expert judges identified one listening item type, 'Academic Monologue,' that had slightly lower agreement among the judges, warranting further discussion. As described earlier, the "Academic Monologue" items assess test takers' ability to understand expository texts in a lecture and are more difficult items for the target population. These items were perceived to be less important may be because the listening input was relatively long and for younger learners or lower-proficiency students, the cognitive load of the stimulus materials posed might be overwhelming. It may also be the case that the "Academic Monologue" is designed for learners with higher proficiency level—a

level that is higher than the one that the participating teachers were familiar with or currently teaching and thus was considered less important or relevant to their given contexts. Follow-up interviews with the EFL teachers lend a hand to explain the results seen here. One Peruvian teacher, who had 21 years of experience teaching beginner to intermediate English for young learners, indicated that her students had limited exposure to this type of listening input and thought that the academic monologues were too demanding for her students. She said: “We do not have that kind of exercise in the textbook or any other listening task we use in class; we consider this kind of exercise a bit demanding for our students who do not have access to that kind of input neither in their schools nor in their daily lives.”

Other teachers interviewed generally had a positive view about the inclusion of the academic monologues; however, three suggested that the choice of topics should take into consideration young learners’ age and life experience. A French teacher, who had 16 years of experience teaching beginner to intermediate young EFL learners, commented that:

My students are never exposed to this kind of listening, except when it has to deal with the culture of an English speaking country, such as the life of Nelson Mandela, the religious wars in Ireland, the pilgrim fathers, the constitution in 1776, etc., but not things about insects or for example the earth. Or it would be very general, like not how a volcano works, but the different types of natural catastrophe that you can experience. That is to say, the topic should not be too technical.

This comment indicated that the French teacher’s students, in fact, had exposure to Academic Monologues; however, they were not familiar with the topics included in TOEFL Primary. While this comment highlights the importance of selecting topics that are accessible for young learners who have limited exposure to complex or abstract concepts, it needs to be noted that the teachers’ perceptions of the topic choice might have been influenced by the two academic monologues given to them for evaluation, since both of them were science-related topics. TOEFL Primary encompasses a wide range of topics that represent a variety of disciplines, both in social and natural sciences. The teachers’ views about the topic choice would have been different if different topics had been chosen. Another interesting point worth discussing relates to the French teacher’s remark on introducing topics such as a prominent historical figure from South Africa or the constitution of the United States. These topics, albeit culturally relevant in the French context, may appear to be less familiar for young EFL learners in different parts of the world or EFL contexts.

The teachers’ comments also bring out an important issue in the content design of young learner assessments—topic effects. Whereas the majority of the teachers considered that the Academic Monologue measures what it is intended to measure, the topics of the monologues appear to impact how the teachers perceived the importance of the KSAs assessed with respect to their teaching contexts. This result suggests that there might be a topic effect on the perceived difficulty of task types and potentially on test performance—an effect that can introduce construct-irrelevant variance (Cho & So, 2014). The impact of topics on test performance thus warrants further investigation to inform the choice of topics for the academic monologues.

In terms of research methodology, the investigation suggests that the use of CVIs and an acceptable standard for the CVIs are useful in estimating the degree of content representativeness of newly developed young learner language assessments. On the basis of the results obtained and previous research (Davis, 1992; Lynn, 1986), it appears that content validation of young learner language assessments can be performed by a judiciously selected panel of expert judges who are familiar with the target population and that the experts' judgments can be analyzed using the CVI approach. Emphasis needs to be placed, however, on the careful adoption of a cut-off point that can be used to determine a good degree of content alignment.

7 Limitations and Suggestions for Future Research

A few limitations of the study need to be pointed out. First of all, while the panelists were experienced, representative EFL teachers judiciously selected from varying EFL contexts, the sample size remains small and thus the findings might only apply to the participating teachers' contexts. Future research in validating content representativeness of newly developed young learner language assessments should include expert judges with more diverse nationalities and larger sample size so as to ensure the generalizability of the study results. Secondly, this study evaluated the reading and listening items of the TOEFL Primary test. The computer-delivered speaking test was not included in the evaluation, leaving open the question of the content representativeness of the speaking tasks and the importance of the speaking communication goals for young EFL learners. Subsequent research should investigate the content representativeness of the speaking tasks so that a more comprehensive evaluation of the TOEFL Primary test can be made available to interested EFL teachers and test users. In addition, future research should also investigate whether the mode of test delivery, i.e. paper-based versus computer-delivered, plays a role in how young language learners process input materials and test prompts in order to inform test design. Finally, the study used information from the EFL teachers' judgments of the test items. Other sources of information (e.g., empirical response data) were not available at the time of data collection; however, they should be considered as potential data sources in the future.

8 Conclusion

Results of the study have provided an important piece of empirical evidence to support the content validity of TOEFL Primary and the intended uses of the test. The KSAs assessed by TOEFL Primary listening and reading items were judged to be important and relevant to the content of the different EFL curricula the panelists were familiar with. This finding corroborates with findings from the domain analyses of EFL textbooks conducted in the initial stage of test development and the

results of the teacher survey discussed earlier. The multi-stages of test validation have yielded convergent results, consolidating the claims made about the test uses by providing meaningful feedback to support language teaching and learning. In addition, this study presented an evaluative process that can be applied to investigate content representativeness of similar language assessments. Equally important, it suggests a significant role for EFL teachers in the development of new tests for young English language learners.

References

- Berk, L. E. (2012). *Child development*. London: Pearson.
- Cho, Y., & So, Y. (2014). *Construct-irrelevant factors influencing young English as a foreign language (EFL) learners' perceptions of test task difficulty* (Research Memorandum No. RM-14-04). Princeton, NJ: Educational Testing Service.
- Crocker, L., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing the fit between test and curriculum. *Applied Measurement in Education, 2*(2), 179–194.
- D'Agostino, J., Karpinski, A., & Welsh, M. (2011). A method to examine content domain structures. *International Journal of Testing, 11*, 295–307.
- Davis, L. L. (1992). Instrument review: Getting the most from your panel of experts. *Applied Nursing Research, 5*, 194–197.
- Fleurquin, F. (2003). Development of a standardized test for young EFL learners. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 1*, 1–23.
- Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247.
- Hsieh, C.-N. (2013, September). *Establishing domain representations for a large-scale language assessment for young EFL learners*. Paper presented at the Midwest Association of Language Testers, Michigan State University, East Lansing, MI.
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382–385.
- McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.
- Nikolov, M. (2016). 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*. New York: Springer.
- Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*, 489–497.
- Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*(4), 285–301.
- Robert, C., Borella, E., Fagot, D., Lecerf, T., & De Ribaupierre, A. (2009). Working memory and inhibitory control across the life span: Intrusion errors in the Reading Span Test. *Memory & Cognition, 37*(3), 336–345.
- Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94–104.
- Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*(4), 299–321.
- Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481.

- So, Y. (2014). Are teacher perspectives useful? Incorporating EFL teacher feedback in the development of a large-scale international English test. *Language Assessment Quarterly*, 11(3), 283–303.
- Turkan, S. & Adler, R. (2011). *Conceptual framework for the assessment of young learners of English as a foreign language*. Unpublished manuscript. Educational Testing Service, Princeton, NJ.
- Wu, J., & Lo, H.-Y. (2011). The YLE tests and teaching in the Taiwanese content. *Research Notes*, 46, 2–6.
- Yalow, E. S., & Popham, W. J. (1983). Content validity at the crossroads. *Educational Researcher*, 12(8), 10–21.

Developing and Piloting Proficiency Tests for Polish Young Learners

Magdalena Szpotowicz and Dorota E. Campfield

Abstract This chapter describes the stages of design of a bespoke pen-and-paper assessment of listening and reading comprehension administered for 10-year-old learners of English as a foreign language in Polish primary schools. Test creation is followed, from initial construct identification through to pilot and careful item analysis leading to final choice of items with the best psychometric qualities. Particular attention is paid to the many challenges to creating a useful large-scale test for measurement of children's foreign language in the context of varied course materials and learning experiences. Critical importance of the child perspective to inform test construction and administration is discussed. Despite the limitations of a closed-ended pen-and-paper format, the result was a child-friendly and attractive assessment. It emphasised authentic language and the type of communication children might expect to meet in everyday situations. It was hoped to encourage exposure to longer stretches of text.

Keywords Assessment • Instructed child foreign language learning • Primary schools • Item analysis • Cognitive interviews

1 Introduction

A bespoke pen-and-paper assessment of listening and reading comprehension for 10-year-old learners was delivered in 2011 as part of a national, empirical study on Polish school effectiveness. A representative sample of over 4700 children from 172 state schools was tested. The aim of this study was to assess English language abilities that children had learnt during their compulsory primary school education. These abilities were assessed twice. First, after Grade 3 (age 9–10), the first phase of primary education and then towards the end of the second phase in Grade 6 (age 12–13) – the concluding phase of primary education. The study, carried out by the Educational Research Institute, was intended to provide evidence for

M. Szpotowicz (✉) • D.E. Campfield
Educational Research Institute, Warsaw, Poland
e-mail: m.szpotowicz@ibe.edu.pl; d.campfield@ibe.edu.pl

recommendations to the Ministry of Education, schools, teachers, parents and pupils concerning foreign language education.

The first assessment of young learner language achievement, at age 10, is the focus of this chapter. It demonstrated the many challenges faced in its measurement and the creation of a bespoke pen-and-paper test for children aged 10. This chapter describes this daunting task, its division into phases, starting from the lengthy process of conceptualization with initial construct identification through stages of design, co-operation with artists, piloting, revision of items and tasks, to the development of pilot and administration, leading to the final choice of test items with the best psychometric parameters. A particular challenge was to ensure age suitability of the test, demanding test creators' appreciation of young learners' developing cognitive and foreign language literacy skills. Pre-pilot meetings referred to as *cognitive laboratories* were held with children of target age to try the tasks, describe their experience and share their opinions. Their contributions highlighted the critical importance of the child perspective and informed the construction of the final test.

2 Language Test Development for Young Learners – Challenges

Children's developmental characteristics together with their low level of foreign language knowledge are key obstacles to developing reliable tools for valid measurement of children's achievement. Deciding factors for test item format and conditions should be strongly determined by the stage of children's cognitive and emotional development (Schaffer, 2004). Cognition is the process of knowing and thinking which integrates reception, storage and processing of information received through the senses. Cognitive processes also include perception, awareness, judgment, the understanding of emotions, memory and learning (Ashman & Conway, 1997, p. 41). In testing children's abilities, attention is the most prominent cognitive factor. Its role in the decoding of information is critical. Attention is defined as the "awareness and sensitivity to objects or events that are occurring (...)" and which enter and leave focus and is intimately bound to interest and selectivity (p. 71).

By the time they start school, children have developed voluntary attention which allows them to focus on classroom tasks. Involuntary attention, dominant in earlier years, is still, however, easily triggered by internal or external stimuli such as noise, light, colour, hunger and tiredness, and may quickly distract children from a set task. When children between ages 6 and 8 are engaged in a single learning task, the maximum time for focused attention during instruction is up to 15–20 min duration, providing the task is engaging and commands their interest (Wesson, 2011). Research in cognitive development shows that attention which is controlled and directed towards a goal is more influenced by age than attention that is triggered by a stimulus or spontaneous exploration of the environment (Enns & Trick, 2006). The older the child, the more motivation they have to remain focused (Bredenkamp

& Copple, 1997). This propensity is an important signal, conditioning initiation into formal testing.

Other important cognitive factors requiring consideration in language test development include the ability to retrieve items from memory (e.g. words, numbers) and correct interpretation of the test layout and symbols used (e.g., icons). Perception is yet another important aspect of cognition at this age. As Vernon and Vernon (1976) state, children's ability to notice and recall details from a picture is greater than their ability to interpret the whole picture. Therefore, test items should favour a series of smaller pictures over a large picture, in which children might become lost.

Affective characteristics are also critical to test performance. Although children's attitudes towards a foreign language are generally positive (Mihaljević Djigunović & Lopriore, 2011; Mihaljević Djigunović & Vilke, 2000), motivation to participate in language tasks is related to classroom atmosphere and the sense of security achieved by the rapport established with the teacher and other learners. Test administration and test characteristics, which do not mimic regular daily school activities and thus do not engender procedure and task familiarity, are likely to cause stress, result in apathy or even loss of motivation. To avoid this, a test might be supervised by the class teacher or, if considered inappropriate, other teachers should be present during the test. A familiar teacher, present during externally administered tests might in many cases re-establish children's sense of security and this provides solid grounds to justify their participation.

Among the challenges to the development of proficiency tests for children is their language content (see Hsieh, 2016 in this volume). This is largely determined by the curriculum and course books used. In Poland, the National Curriculum (2008) consists of several descriptors formulated as expected learning outcomes at every stage of school education. The document was designed to be suitable for all foreign languages and does not list language items for a target language. The list of topics to be covered within each stage is available for all stages, with the exception of stage one (age 6–8). Table 1 shows the expected learning outcomes for foreign language education at stage 1 (age 9).

In Poland, as in many other European countries, child target language exposure is often limited to school. Contact with the foreign language outside school, through television, digital media or native speakers is sporadic (Muñoz & Lindgren, 2011, 2013). For this reason, language competence is largely circumscribed by course book content. For young learners, the content of course books is usually planned around common topics while the choice of lexical items and phrases is often determined by the storylines used (Rixon, 1999). This results in relatively few lexical items common between course books used nationally (Alexiou & Konstantakis, 2007; Kulas, 2012). The absence of a common point of reference manifests itself in a situation in which children's lexicon varies from one school to another, depending on choice of course book. It is, therefore, rather difficult to describe a common core of items shared by course books for a child population of the same age.

Rate of development for literacy in the mother tongue is important in determining how foreign language skills and achievement can be tested. In Poland, it is recommended that reading and writing should not be taught before children are aged

Table 1 Expected learning outcomes in a foreign language at educational stage 1 (age 9) in the National Core Curriculum (MEN, 2008, p. 216)

A pupil who has accomplished 3 years of FL instruction (age 9)	
Listening	distinguishes between words which sound similar
	recognizes everyday phrases and can use them
	understands the gist of short stories told with the help of pictures and gestures
	understands the gist of simple dialogues in picture comic strips (also in audio and video recordings)
Speaking	responds verbally and non-verbally to simple instructions
	asks questions and responds using formulaic phrases, says rhymes, chants and sings songs, names objects in the learning environment and describes them, participates in drama activities
Reading	understands the gist of dialogues in picture comic strips
	understands simple words and sentences in reading tasks
Writing	copies words and sentences
Non-linguistic skills	uses picture dictionaries, readers and multimedia
	cooperates with peers

6–7. Since ability to read and write in a foreign language follows the development of literacy in L1, children are introduced to reading and writing in a foreign language a few years later, usually when they are aged 8–9. Before this age neither mother tongue nor foreign language skills are formally tested. Development of L1 and L2 literacy can be compared for listening and reading at the age of 9. Table 2 shows that age 9 achievement targets in the mother tongue are considerably higher than for the foreign language (Table 1). The foreign language skills of young learners at this age are closer to those acquired in the mother tongue 2 years earlier (Table 2).

The difference between expected learning outcomes for mother tongue and the foreign language highlights the later onset of literacy in L2. This poses an obstacle to parallel test design for mother tongue and a foreign language. Since literacy in L2 is less developed, tests and tasks may necessarily appear ‘childish’ and below learners’ levels of cognitive ability. For example, while children are exposed to longer written instructions and passages of text in their mother tongue, in the foreign language they are only ready to respond to short sentences supported by pictures or icons which they may conceive as more appropriate for preschool.

In view of these key considerations, the challenges of test item development for large-scale measurement of children’s foreign language need to be regarded from the perspective of test usefulness which is “an overriding consideration in designing, developing and using tests” (Bachman, 2004, p. 5). According to Bachmann and Palmer (1996), this engenders vital qualities, including: reliability, construct validity, authenticity, interaction, impact and practicality. McKay (2006) notes that these qualities should be observed from the design phase. Each is discussed below from the perspective of test item development for children aged 9.

To reduce compromising reliability of large scale testing for children’s language skills, as in the example presented in this study, the administration stage for the test

Table 2 Learning outcomes in the mother tongue for educational stage 1 – translation of the National Core Curriculum (MEN, 2008)

A pupil who has completed 1 year of mother tongue instruction (aged 7)		A pupil who has accomplished 3 years of mother tongue instruction (aged 9)
Listening	pays attention to peer and adult contributions and is willing to understand them	listens attentively and can respond appropriately to the information obtained
Speaking	communicates their reflections, needs and feelings in a clear way	makes contributions a few sentences long, tells short stories, describes objects and people
	addresses the interlocutor in a respectful manner, speaks to the point, asks and answers question, adjusts their tone of voice to the situation	participates in conversations, asks and answers questions, presenting their personal point of view, expanding lexis and syntax
	participates in conversation about family, school and literature	pays attention to register of the conversation, uses correct pronunciation, stress and intonation in affirmative, interrogative and negative sentences, uses pleasantries
Reading	understands the sense of coding and decoding information, understands simplified pictures, pictograms, signs and headings	reads and understands age-appropriate texts and draws conclusions
	knows all letters of the alphabet, reads and understands short and simple texts	selects specific information from texts, referring to young learner dictionaries or encyclopaedias as required is familiar with genres such as: greetings, invitations, announcements, letters or notes and can respond appropriately
Writing	writes short, simple sentences, copies, writes from memory	writes stories a few sentences long, letters, greetings and invitations
	writes clearly and follows the rules of handwriting	produces clear and legible handwriting pays attention to grammar, spelling and punctuation rules
		copies and writes text from memory and can formulate individual contributions

demands rigorous attention. Among the requirements for test procedures for language learners of English as a second language recommended by Butler and Stevens (2001, p. 413), some were particularly apposite to the present study. These included: testing spread over several sessions, administration to small groups in separate rooms, breaks during testing, native language instructions given orally, questions read aloud in English, answers inserted directly in a specially prepared test booklet and the instructions explained.

Construct validity should be ensured by extensive literature review covering child socio-psychological and cognitive development, foreign language learning at an early age and local teaching and assessment practices (McKay, 2005; Taylor & Saville, 2002). Test developers should acquire knowledge of the constructs to be assessed, supported by in-depth analysis of curricula and course books (Inbar-Lourie

& Shohamy, 2009). Taylor and Saville stress the primacy of spoken over written language with respect to young learners – hence the focus on oral/aural skills in tests for young learners, such as the Cambridge Young Learners' English Tests.

Task authenticity, defined as the “degree to which test tasks resemble target language use (TLU) tasks” (Carr, 2011, p. 314) is easier to achieve during informal classroom assessment than in large-scale external tests. To select authentic tasks appropriate for young learners in a national context, test item writers need an appreciation of the tasks used during lessons, offered by course books and other materials supplied by teachers or materials, such as comic strips or cartoons, which children may read or look at in their spare time.

McKay (2006) asserts that only interactive tasks which require children to use the language knowledge and skills that are being assessed can provide useful evidence for inference of children's level of language competence. In a pen-and-paper test, listening and reading skills can be assessed if the format of the tasks and content are familiar through prior classroom exposure.

Espinoza and Lopez (2007) give a critical overview of current assessment measures for young English language learners and point out the scarcity of appropriate standardized tests.

When testing young learners it is vital to ensure positive impact and to avoid children – the test-takers – experiencing any negative consequences. According to Messick's (1989) work on validity theory, “consequences of tests must be sufficiently positive to justify the use of the test”. Carr (2011, p. 55) argues that washback, the effect of a test on teaching and learning, is the most commonly discussed aspect of impact. In high-stakes tests washback may include the curriculum, materials, teaching approaches and how students prepare for tests. “Trying to plan tests that seem likely to cause positive *washback* is important, because teachers will wind up teaching to the test, at least to some extent” (Carr, p. 55).

Social consequences should also be considered when designing external tests for young learners, especially with regard to test fairness and ethical considerations. According to *Kunnan's Test Fairness Framework* (2004), apart from being valid, a test should be free from bias (e.g., standard setting and analysis of differential item functioning), ensure uniform security for administration and provide equal access to students (e.g., familiarity with equipment, conditions and the opportunity to learn from the test) (cited in Carr, 2011, p. 155). With reference to ethical considerations, anonymity in test administration is crucial and needs to be guaranteed by design of suitable test procedures at the planning stage. It is paramount that neither children nor their teachers can be identified either during transport or coding of scripts or later from the database. The most delicate issue, however, concerns publication of test results to be shared with teachers, schools or authorities. Reporting requires tact and extreme care to present the results in an informative and useful way without risk of any detrimental washback on learners or their teachers.

3 Context and Research Questions

3.1 *The Context of the Study*

The aim of the present study was to assess children's foreign language abilities after completion of the first stage of foreign language education in primary school, Grade 3 (age 10). To conform to this, the research population was defined as those pupils who had completed the first phase of primary education and who at the beginning of the study had just started Grade 4. These children started school in 2008 at the age of 7 when English as a foreign language was made compulsory in primary schools. Since town size has been shown to be a significant factor in educational research in Poland, to obtain a representative sample of the population, a stratified random sampling framework was adopted to reflect the range of settlement size from cities and large towns, through market towns serving farming populations to villages. As a result, 172 primary schools were randomly selected. In schools with one or two Grade 4 classes, all pupils were selected for the study, whilst in schools with more than two Grade 4 classes, two classes were randomly selected. This sampling procedure resulted in 4717 pupils qualifying for the study frame.

The pen-and-paper test was administered to the full study sample to assess listening and reading comprehension. The choice of these two skills for assessment was informed mainly by practical considerations; since it is possible to assess them using pen-and-paper tests which, given the sample size, was deemed practically and logistically feasible (Szpotowicz & Lindgren, 2011). Written production skills were assessed in the second phase of the study when pupils were at the end of Grade 6 (age 12, not reported in this chapter). Oral production skills were not assessed but an Elicited Imitation task was carried out on a sub-sample of 665 children (Campfield, *in preparation*).

The constructs for listening and reading comprehension were suggested by the National Foreign Language Curriculum (Ministerstwo Edukacji Narodowej (MEN), 2002, 2008) and the European Language Portfolio for children aged 6–10 (Pamuła, Bajorek, Bartosz-Przybyło, & Sikora-Banasik, 2006). For children completing the first phase of primary foreign language instruction, listening comprehension was defined as:

- (a) ability to comprehend lexical items (e.g., names of foods, animals, rooms and items of furniture, body parts, sport and leisure activities) and simple everyday expressions (e.g., classroom language),
- (b) ability to follow the general gist of simple dialogues supported by visual prompts/materials.

Reading comprehension was defined as:

- (a) ability to comprehend single words and simple everyday expressions,
- (b) ability to follow the general gist of simple texts, such as stories.

3.2 *Research Questions*

The study reported here aimed to address the following questions:

- What is the level of listening and reading comprehension exhibited by children who started learning English as a compulsory school subject in 2008?
- Which school- and home-related factors influence these abilities?

4 Method

The specific focus of this chapter is the description of the various stages of design for the pen-and-paper listening and reading comprehension tests, through the pilot stage to the final choice of test items with the best psychometric parameters.

4.1 *Participants*

The research population were 10-year old children who had completed Grade 3 and were just starting Grade 4. The study materials were piloted on a convenience sample of the target age group. The pilot sample was drawn from three geographic areas: the North-East, South-East and central Poland, covering radii of 50 km from the biggest town in each area, principally for economies of travel and cost for researchers. Within each area, primary schools were selected to reflect the socio-economic character of the area: eight schools in the North- and South-East and six schools in central Poland. This resulted in selection of 22 schools from larger cities, smaller towns as well as market towns serving the farming population. Care was taken to ensure that no schools were at the extremes of the socio-economic or academic ability spectrum. Since in the course of their research careers the researchers involved in this study had established contact with these schools, this encouraged them to be willing to participate in the pilot. From the 22 schools chosen for pilot, 42 Grade 4 classes were selected. A total of 829 pupils took part.

4.2 *Materials*

The design and development of the pen-and-paper test followed the preparation of an assessment task specification formulated with reference to Carr (2011, p. 50) and McKay (2006). The final goal of the study was to formulate recommendations concerning foreign language instruction for the Ministry of Education, school heads, teachers, parents and pupils. The aim of the assessment, therefore, was to generate potential for a large positive impact on the acquisition of foreign language by young

learners with all effects judged as being desirable and using a test considered fair by all stakeholders.

To satisfy the criterion of fairness, it was important that (a) children had been previously exposed to the proposed types of assessment task and (b) the target language used was drawn from familiar vocabulary and structures. Therefore, for the test to be fair, the assessment tasks had to reflect children's classroom experience. However, a positive *washback* effect was also an important aim for the assessment. For this reason, the specification required task developers to place emphasis on authentic language and turn of phrase and use listening material which was as realistic as possible. To reiterate, the aim was to be able to describe the extent to which children had understood words and simple expressions used in situations they might expect to encounter every day.

Test items were constructed within the Institute by a team of experienced test developers, researchers with experience in child second language acquisition, language teaching for young children and teacher training. The team included a native speaker of British English who also monitored that authenticity of language and turn of phrase was satisfied. An internal and an external expert on language testing were consulted on all materials on a continuous basis as an integral part of the task development process.

The team of item developers were working according to a set of jointly-drawn guidelines, such as authenticity of language and delivery, in the case of the listening material and the avoidance of incorrect English, contrived or peculiar expressions and trick questions. The language and contexts were expected to be universally familiar, requiring unambiguous interpretation. Furthermore, responses to items could not be made on the basis of single lexical items. The test materials had to be conceptually and visually pleasing with clear and ample instructions supported by sufficient examples. Finally, test items needed to be at appropriate levels of difficulty to allow them to potentially function as anchor items for the second assessment, at the end of Grade 6 (age 12, not reported in this chapter).

Item construction was preceded by the analysis of vocabulary and structures in the English language textbooks approved by the Polish Ministry of Education and available on the market in the autumn of 2010 for Grades 2 (age 8–9) and 3 (age 9–10) of primary school (Kulas, 2012). This analysis demonstrated great variance between textbooks in terms of both the range and commonality of vocabulary but allowed the selection of 177 lexical items common to all textbooks. Rixon (1999) had commented on the paucity of common vocabulary between children's textbooks which bears scant resemblance to what would be expected for learners in the target language environment.

In the present study it was not possible to obtain a measure of the frequency of exposure to each of the 177 lexical items because the frequency of a word's appearance in any book does not impute its frequency of use in the classroom. To obtain this data it would be necessary to conduct a large observation study. In the absence of knowledge about exposure, piloting at a later stage was expected to be the best predictor for suitability of choice of vocabulary.

The list of common vocabulary and language structures compiled as a result of textbook analysis formed the basis for item development. However, this common core was not the sole source of language for task construction, since during item construction the authors used some individual lexical items outside the common list but believed to feature in the first years of English at school. Additionally, these lexical items outside the common list were not specifically instrumental to the understanding of test items but provided necessary language for item construction.

Test writers were guided by two considerations in item construction. Language contained had to be close to what children were likely to have heard in the course of their instruction. Equally important was the desire to emphasise authentic language and realistic communication to assess the extent of children's ability to comprehend the spoken exchanges or simple texts they might meet in everyday situations. Care was taken for tasks to reflect such types of communication and present language in appropriate contexts. Therefore, the tasks took the form of short dialogues and brief descriptions with which children could conceivably have been engaged during school. The emphasis on authentic language and realistic communication aimed to encourage and reinforce classroom practice and the types of task aimed to encourage exposure to longer stretches of text.

Two tasks were prepared to assess listening and three to assess reading comprehension. To ensure variety, one task to assess listening comprehension was *multiple-choice* and the other was of the *true/false* type. Reading comprehension was assessed by *multiple-choice*, a *picture with text matching* and *title and text matching* tasks. Two versions of the *multiple-choice* tasks were constructed and four for *picture matching with text* and *title and text matching*.

Given the participants' age and the level of L2 literacy expected to have been reached after 3 years of exposure in instructional settings, listening and reading comprehension were to be assessed without requiring written responses. Therefore, two artists with experience of illustrating materials for children were engaged to prepare supporting illustrations for the tasks. For this age group, illustrations were also considered good promoters for motivation to complete the task. Children were required to mark their responses by circling letters, labelling illustrations or sentences in the case of multiple-choice tasks, crossing the right box in the case of the true/false tasks and ordering sentences in the correct sequence for pictures with text or titles with text matching. One illustrator prepared materials for the listening and the other for reading comprehension.

Initial versions of tasks were assessed by children of the target age group in a number of meetings with small groups of children held in three different regions of the country. These meetings, referred to as *cognitive laboratories*, were fundamental to the process of task construction and are, therefore, described in the section below. They provided information on children's understanding, perception of the language and the visual materials or types of tasks. These findings identified aspects of tasks for modification or to be rejected in view of children's reactions. Table 3 shows task versions that progressed to the pilot stage following the cognitive laboratories.

Table 3 Piloted versions of listening and reading comprehension tasks with number of items in each task

Instrument	Pilot version	Type	Number of test items
Listening 1	1	Multiple choice	19
	2		
Listening 2	1	True/False (<i>Family at home</i>)	11
	3	True/False (<i>In the park</i>)	
	4	True/False (<i>In the classroom</i>)	
Reading 1	1	Multiple choice	18
	2		
Reading 2	1	Picture and text matching (<i>The story of cat and mouse</i>)	10
	2	Picture and text matching (<i>Computer</i>)	
	4	Picture and text matching (<i>TV</i>)	
Reading 3	1	Title and text matching (<i>Too many sweets</i>)	5
	2	Title and text matching (<i>Play with animals every day</i>)	
	4	Title and text matching (<i>Holiday hobby</i>)	

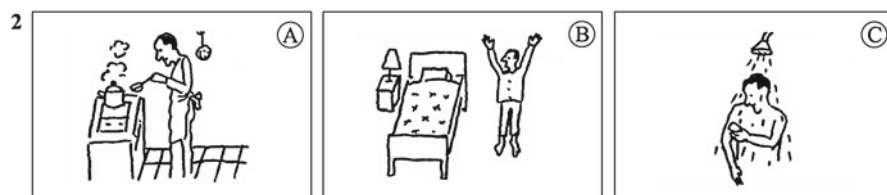


Fig. 1 Example of listening comprehension items in task 1: multiple choice

In the first listening task, children listened to an utterance or a brief exchange and were asked to indicate which of the three illustrations best fitted what they had heard (Fig. 1). In the second listening task, children looked at an illustration depicting a lively scene and heard utterances or brief dialogues requiring them to identify whether what they heard was a true representation of the scene (Fig. 2). The tasks were prepared in a way which avoided possible guessing based on familiarity with any single individual word.

Translation of the instruction in Polish: *Indicate which picture matches the recording. You will hear the recording twice .*



nr	True	False
1		X
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		

Fig. 2 Example of listening task 2 – true/false (*In the park*)

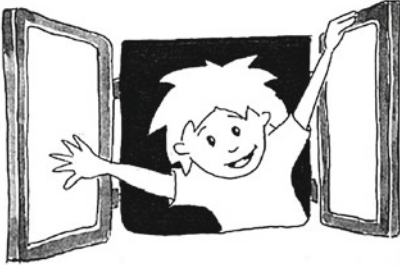
Translation of the instruction in Polish: *Look carefully at the scene. Listen to the sentences or brief dialogues and mark the appropriate box according to whether what you hear is True or False with a cross (x). You will hear the recording twice.*


Materials for the listening comprehension tasks were recorded by a male and female pair of native British English teachers of children with relevant studio experience. Recordings were made using a normal speaking voice and natural intonation. Care was taken to ensure that the recorded material was delivered with the stress, rhythm and intonation of natural British English.

In the first of the three reading comprehension tasks children were presented with three sentences and a picture to illustrate one of these sentences (Fig. 3). The second reading task presented a brief story using 11 consecutive cartoon-like illustrations (Fig. 4). Below the sequence of pictures, sentences or brief exchanges/dialogues were presented in the wrong order, ten matched the illustrations and one extra text did not match any of the illustrations. The task was to match sentences with the illustrations.

Translation of the instruction in Polish: *There are three sentences below each picture. Choose the sentence which describes the picture and tick the box next to it.*

Translation of the instruction in Polish: *Look carefully at the pictures in the story. There are 10 pictures in the correct order. Match the sentences with the pictures. Write the number of the picture next to the correct sentence. There are 11 sentences, so one is extra.*

4 

5 

What a boring game!

What a lovely day!

What terrible weather!

Go and brush your teeth!

Dinner's ready!

Let's go for a ride!

Fig. 3 Example of reading comprehension task 1 (multiple choice)

In the final reading task children were presented with five brief texts with eight possible titles to match to these texts (Fig. 5). Two examples were given: one as an example of a correct match and the other an example of a title that did not match any of the texts, marked appropriately as '0'. With eight titles to choose from, the task offered five items. This task was included following the advice of the external expert and after much deliberation by the team of authors. The rationale for including this task was twofold. First, it allowed for the assessment of a reading sub-skill: understanding the main idea. Additionally, as with the second reading task (picture and sentence matching to follow a story), the aim was to introduce an important *wash-back* effect on classroom practice to encourage teachers to expose young learners to stretches of text. Particular effort was made to ensure that such texts were interesting, age-appropriate and as with all other tasks, responses required reading of the whole text and could not be guessed from individual words.

Although the authors were aware of the need to avoid item interdependence, this was not always possible, given the narrow range of options (see Figs. 4 and 5). There were difficulties allowing for task variety without including some requiring reordering of sentences to match a story line or the titles with texts. It was hoped that additional items provided with these tasks helped mitigate this shortcoming in the last two reading tasks.

Additionally and encouraged by Nikolov and Szabó (2012, also see Nikolov, 2016 in this volume) each task was followed by three multiple choice items to enquire about how participants rated task difficulty, familiarity and attractiveness (see Fig. 6). The aim was to find out how children themselves reacted to the tasks, to assess perception of task features in relation to ability to tackle the challenge.

1 Mum says: 'Please, come and have something to eat.' But Tom is very busy.

2 It is time to go to bed.

3 Tom looks at the dog. It's too late to go for a walk!

4 Anna is brushing her teeth. It's too late to play.

5 Dad is reading a newspaper. It's too late to play.

6 It's evening. Toms says to his Mum: 'I'm hungry!'

7 Tom is at home. He is playing on the computer.

8 Dad goes into Tom's room and says: 'Let's go and play football!'

9 Tom's little sister Anna asks him to play with her.

10 Tom is sad. Nobody wants to play. It's too late.


Fig. 4 Example of reading comprehension task 2 (picture and text matching)

1

One day in the summer we were going on a picnic. Mum and I were ready. Then Dad walked out of the garage and said: "I'm really sorry but the car will not start". "Oh no!" I cried. "Don't worry. We can go on our bikes to a new picnic place." said Mum. We cycled to a beautiful park and spent all day fishing and playing games!


2

In this photo I look scared! I'm




3

It's a photo of my mother's birthday



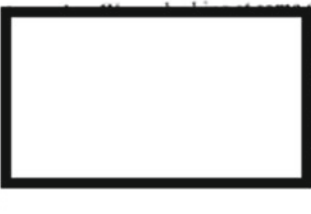
4

I don't like myself in this photo, but it's funny. I'm tired and a bit angry



5

In this photo Kate and I are sitting in



- My family on our camping holiday
- A holiday hobby
- 1 A bad start and a good ending
- Not my plan
- I'm playing with my friends
- A happy summer day
- Holidays by the water
- 0 My dog on holidays

14

Fig. 5 Example of reading comprehension task 3: title and text matching

Po wykonaniu zadania 1 zastanów się i zaznacz czy zadanie było:

1 (A) łatwe	(B) ani łatwe ani trudne	(C) trudne
2 (A) znane	(B) trudno powiedzieć	(C) nieznanne
3 (A) podobało mi się	(B) nie wiem	(C) nie podobało mi się




Fig. 6 Example task of task evaluation for children

5 Results

5.1 *Pre-pilot Stage: Cognitive Laboratories*

Since children's perspectives and opinions were considered vital to the creation of suitable test materials, pre-pilot cognitive laboratories with target-age children were organised. A cognitive laboratory aims at reconstructing possible problems with interpretations of instructions and questions, evaluating tasks and the level or sources of difficulty to complete the test. It is organised in the form of a cognitive interview (Beatty & Willis, 2007), involving the administration of draft survey questions while collecting additional verbal information to evaluate the quality of responses the questions generate. The procedures most often used are based on two approaches (Beatty & Willis, 2007). In the first approach the researcher's role is "to facilitate the participants thought processes" (p. 289) and to follow a strict think-aloud protocol which the researcher records. The other approach is internally varied, including a group of methods, referred to as *probing* and derives from the practice of intensive interview followed by probes. The researcher asks participants about specific items in a test or questionnaire. These questions may be flexible to allow exploration of opinions or structured for comparability of results between different researchers.

The Beatty and Willis (2007) review describes the advantages of both approaches, yet they see more benefits of *probing over thinking aloud*. The chief drawback of the latter approach is that less able participants more frequently become confused and less tolerant of the procedure (Redline, Smiley, Lee, DeMaio, & Dillman, 1998). This is an important consideration with child participants who tend to require individual attention.

In this study the cognitive laboratories were in the form of interviews which followed a relatively strict protocol but allowed some flexibility, including asking children for additional explanation. The aims were to explore how children

- understood instructions: to ensure they had been formulated in an age-appropriate and comprehensible way
- responded to test items: in order to estimate their level of difficulty
- felt about the illustrations: in order to check if the style and aesthetics appealed to young learners' tastes
- commented on the difficulty and user-friendliness of the whole test and individual items.

Sample selection aimed to obtain interviews with children of varying abilities in English. The 36 children chosen were 9 years old and attended schools in three geographically distinct Polish regions (Podlasie, North-Eastern, Mazowsze, Central, Dolnośląskie, South-Western). Schools were located in rural, urban and suburban areas with varying socio-economic characteristics. School and parental consent for the interviews was previously obtained.

Interviews were carried out by three researchers following the same procedure and took place with groups of four to six children in quiet classrooms. Children

were presented with the tasks sequentially and separately, so that they could attempt to complete each task and were able to comment immediately. The researcher noted the times children needed to complete each task. The same *probe* procedure was used with all participants. It involved the following steps:

- The researcher introduced herself and explained the children's role as advisors for the creation of tasks for other children which would be used as teaching and test materials.
- Copies of tasks were distributed and children were encouraged to attempt the tasks.
- After they completed each task the researcher asked questions and recorded answers. Children were first asked to respond spontaneously and those who did not volunteer were approached individually and asked to share their opinions.

The questions asked during interviews were as follows:

1. Was the task easy or difficult? What made it easy or difficult?
2. Was the task interesting or boring? What made it interesting or boring?
3. Did you like the illustration, its layout and design of the page?
4. Were the instructions clear?
5. Would you change anything in the task? What and how?

On reflection on one's performance in language tasks and self-assessment techniques used in assessing young language learners (see also Butler 2016 and Nikolov 2016, both in this volume).

5.2 Key Outcomes from Cognitive Laboratories – Problematic Tasks

The value of the findings from cognitive interviews cannot be overestimated. It showed that although researchers and test writers were experienced with the age-group, tasks demanded some radical changes. Some types of tasks were abandoned, others were removed from the test directly after the cognitive interviews and those that remained were further tested during the pilot.

Task type: *title and text matching (reading comprehension)*

The main challenge with any jumbled text is that the way one item is answered influences the other items. If a student marks one answer incorrectly, they immediately block two possible answers with this response (the correct option, which remains empty and can only become an incorrect response to another item, and the incorrect one, which prevents a correct response to another item). In this way the items are not independent and their relationship reduces test validity. Since children can rarely focus on a text for more than a few minutes, the necessarily short text does not provide enough material for many suitable items. As a result, a reading task, providing only four or five items cannot offer high reliability. The children

often tried to guess which title matched a text without reading it and sometimes they found a few key words which were sufficient to provide the correct answer without the need to understand the whole text.

Task type – *picture matching with text (reading comprehension)*

The task in which children matched jumbled speech bubbles to scenes in a comic strip and which seemed to be both age-appropriate and interesting, emerged as a serious challenge to develop. The text often appeared ambiguous and sometimes one speech bubble matched more than one picture. On other occasions children could number the jumbled text for a story without looking at the comic. As with the task described above, the problem of related items remained.

Task type: *Marking statements about one picture as true or false (listening comprehension)*

This task presented a relatively complex picture containing many elements and a few people, e.g., a living room or a classroom. Next to the picture there was a chart with item numbers and spaces to indicate the truth of the statements about the illustration which children listened to in the recording. Although seemingly age-appropriate, the task was confusing and was of low reliability. Primarily, it required quick aural and visual processing of information (recording to picture). Although the recording of each statement was played twice, some children needed longer to respond.

5.3 *Cognitive Interviews – The Benefits*

Beyond observing children's immediate reactions to particular types of tasks, cognitive interviews provided a unique and invaluable opportunity to collect

- feedback on the clarity of instructions (order, language used)
(e.g., it was evident that children did not know the word *paws* which, although it was not key to understanding, completely distracted them, making them focus on what they did not understand)
- feedback on procedures (tolerable length, estimated time of performance)
(the interviews revealed differing response times and various strategies and learning styles, e.g., risk-takers and risk-avoiders)
- comments on the ambiguity of picture-text relationships (in matching sentence to picture two sentences seemed to suit one picture): *Two sentences are OK for the last picture* “It’s time to go to bed” and “Tom is sad. Nobody wants to play. It’s too late.” *he is sad, but it is late, so it is time to go to bed, so this is not a good item, you know?* (about reading task 2 in Fig. 2)
- comments on the transparency and aesthetics of the illustrations: *There should be no posters with text in Polish – it’s an English test.* (comment about a picture of a classroom in listening task 2)

- children’s practical advice for improving the items (e.g., changing vocabulary items which determined comprehension of the whole reading passage): *I didn’t need to read the whole text, just the first two sentences. It was enough to know these two words.*
- corrections of inconsistencies between pictures and texts: *Grandpa in the picture is not wearing a jacket which we heard in the recording, but a sweater!*

The extracts below show selected reactions and opinions expressed spontaneously during the cognitive interviews.

1. A boy who read the following text in reading task 3 in the cognitive laboratory reacted as follows:

Text: “Who are you going to write about?” asks Mark. “Bella, my sister. She is my best friend” answers Suzy. “That’s nice!”

The boy (genuinely surprised with the above text):

A sister who is the best friend? I’ve never heard of anything like that before.

2. A girl’s reaction to the artist’s illustration of a sentence describing a child doing her homework:

The girl cannot be doing her homework. If she is sitting at the computer, she must be playing computer games.

5.4 Pilot Procedure

A letter with a broad description of the study and its aims was sent to heads of the schools that agreed to take part in the pilot. Parents and caretakers were also sent an information letter and were asked to consent to their child taking part in the study. The school heads were made aware that participation in the pilot was anonymous and confidential in that no information specific to a particular child could be easily traced back to that child and that no person other than the researcher was to be present during the test or able to see any element of it. It is worth pointing out that performance on tasks, the reliability of which the pilot served to assess, could not form the basis for pupil assessment, although some useful general suggestions could be made in the form of constructive feedback.

Four staff from the Educational Research Institute supervised the pilot during May 2011 after an internal training session. Training was intended to ensure that the guidelines and procedures were followed in the same way at all schools. This training was a prelude for training of test administrators recruited specially for the main study for whom a training video and simulation scenarios were prepared. In the pilot, each version of the tasks shown in Table 4 was administered at least 320 times.

Researchers were instructed to avoid planning pilot sessions on busy school days which might be predicted as likely to introduce distraction or disturbance. Testing during lessons immediately before lunch was also to be avoided, although it was important that no child was hungry, thirsty, upset in any way or needed the toilet.

Table 4 Pilot test versions

Task	Test version												
	A	B	C	D	E	F	G	H	I	J	K	L	M
Listening 1	2 ^a	1	1	1	1	2	2	1	2	1	2	1	2
Listening 2	4	4	3	3	1	3	1	4	1	3	4	3	1
Reading 1	2	2	1	2	1	2	1	2	1	2	1	1	2
Reading 2	2	1	4	2	1	1	4	2	4	1	4	2	4
Reading 3	2	4	1	4	4	1	4	1	2	2	4	2	1

^aNumbers in columns refer to task versions shown in the second column in Table 3

Researchers (during the pilot) and administrators (during the main study) were encouraged to adopt the role and demeanour of a facilitator, supporting children through the experience, being helpful and friendly, smiling and looking at the children when talking to them, establishing eye contact and immediate rapport. While they were asked to administer the test efficiently, they were also requested to avoid looking officious, behaving formally or creating an exam atmosphere. This included not dressing in a way that children might associate with authority.

Information the children received about the test itself and particularly about their roles was considered vital to the success of the assessment. It was important to thank them for agreeing to take part and emphasise their importance as helpers in the research since their participation would provide information aimed to improve foreign language learning for all school children in the country. The research aims were explained to them in age-appropriate language.

Whilst there may be exceptions, the general climate in Polish schools encourages competitiveness between children who are used to a degree of continuous assessment, having their work graded and often being compared to their peers. It was important, therefore, to emphasise that this was not the aim of this research and that the children's performance would not be similarly judged, nor would they receive any points or marks for their performance. They were encouraged, however, to do their best, without being upset if they found something difficult. They were asked to respond to each test task reasonably quickly, to the best of their ability, before proceeding to the next. It was suggested that they could return to any problematic items at the end, i.e., they should not spend too long on one question since they could return to parts of the test they found more difficult. They were told how long the test would take, that it was not a race and that there would be plenty of time to answer every question. Since the children might not have done a test like this before, they were encouraged to understand the task first and look at the questions carefully before answering. As a result of the pilot, it was decided that in the main study a training exercise of about 10 min would be used to introduce children to the test (see [Appendix](#)).

Children were asked not to talk during the test but to raise their hand if they had any questions or still found aspects of tasks unclear. It was stressed that since only what they could do themselves was of interest, they should not be tempted to look at what other children were doing. For reasons of timing and logistics, the pilot was

administered in intact classrooms, with seating traditionally arranged, pupils sitting in pairs at desks arranged in two or three rows. For each pair, classroom boxes for storing materials were used as makeshift divides between children. The aim was to discourage them looking at how others were responding. However, the pilot showed that some children found it difficult to resist the temptation. During the main study participants sat individually, reducing the possibility of copying.

Each task began with an example. The tasks were administered in the sequence shown in Table 3. The two listening tasks were sequenced with all children working at the same pace. A single repetition of all listening material was played to guarantee redundancy deemed necessary for this age group. For the pilot, the entire test comprising all five tasks took approximately 45 min. In the main study the test was administered in two sessions, each lasting 30 min with a 10-min break between them. The first session consisted of a 10-min training test, followed by the two listening tasks and the second contained the three reading tasks. Children who finished the test earlier were asked to check their answers when possible, turn the paper over and stay in the room until the end of the session. Five minutes before the end they were gently reminded of the time remaining.

Some pilot sessions were in the presence of the class teacher whilst in others the researcher was alone. During the pilot, it was found that for the main study the class teacher should be present, introduce the person administering the test, help with supervision and deal with any discipline problems arising. The one proviso was that the class teacher should not be their English teacher. In this case another teacher familiar to them would assist.

6 Results of Pilot Study

Table 5 demonstrates the sequence of events followed leading to the final version of the test.

Following the pilot, the theoretical framework applied to design the measurements of ability relied on Item Response Theory (IRT) as guidance for suitability of candidate tasks. IRT yielded detailed descriptions of the relationship between pupils' ability and the likelihood of their being able to approach the task items. Descriptions of item difficulty and their discrimination indices suggested a task construction which ensured discrimination between pupils of different levels of ability over the expected ability range. It was important that items avoided ceiling effects and also to offer the weakest pupils an opportunity to derive a sense of achievement from the assessment. A sufficient number of items of appropriate difficulty were required to measure ability in the second study phase, when the same pupils would be tested again at the end of Grade 6.

The aim of the pilot was to (a) assess psychometric characteristics both of tasks and items, (b) obtain reliability indices for all tasks and test versions and (c) evaluate the task administration procedures intended for the main study. The task versions (see Table 3) were organised into 13 possible test versions (see Table 4) with

Table 5 Test development sequence

Stages of test development and administration		Additional tasks
1	Test conceptualisation	
2	Course book analysis (common vocabulary and structures)	
3	Selection of types of tasks	Consultation with external experts
4	Test plan and specification	
5	Recruitment of illustrators	
6	Evaluation of sample drawings for listening and reading items	Consultation with external experts
7	First versions of test items	Consultation with external experts
8	Initial cognitive laboratories	
9	Correction following first laboratories	Consultation with external experts
10	Correction and modification of test items	Sampling design consultation
11	Cognitive laboratories following modification	Recruitment of schools
12	Assembling final pilot versions	Audio recordings
13	Proofreading	
14	Copying and posting tests to schools	Training of test administrators
15	Pilot-test administration	
16	Recording pilot-test data	
17	Analysis (IRT and CTT)	
18	Selecting items for the final test	Consultation with experts
19	Assembling final test	
20	Final proofreading of test	

Table 6 Pilot reliability indices: test versions (Cronbach's alpha and IRT Rasch modelling)

Task	Test version												
	A	B	C	D	E	F	G	H	I	J	K	L	M
Cronbach's alpha	.60	.63	.76	.69	.76	.71	.68	.64	.57	.80	.20	.61	.78
Person reliability (Rasch)	.50	.72	.64	.81	.72	.66	.52	.81	.80	.78	.58	.60	.55
Item reliability (Rasch)	.99	.99	.99	.99	.98	.99	.99	.99	.99	.99	.99	.99	.99

each child taking one test comprised of two listening and three reading comprehension tasks.

Reliability analysis was carried out using both Classical Test Theory and Item Response Theory (IRT) with the use of Rasch modelling in Winsteps v. 3.74. Reliability indices were obtained for individual tasks and for the 13 test versions (A to M, Table 6). Cronbach's alpha ranging from .60 to .70 is considered 'acceptable' and from .70 to .90, 'good' for low-stakes testing. Table 6 shows that some sets of tasks, i.e., test versions, demonstrated good reliability indices. The *person reliability index* represents the replicability of rank order that could be expected if the sample of participants were given another set of items measuring the same construct

whilst the *item reliability index* indicates the replicability of item ranking that could be expected if the same items were given to the same-sized sample with different participants behaving in the same way (Wright & Masters, 1982). Table 6 demonstrates that all sets of tasks had very high item reliability indices but in some cases considerably lower person reliability indices, suggesting that learners were guessing or that their responses were influenced by other children's responses.

Apart from providing reliability indices, IRT allowed assessment of

- (a) the extent to which each item difficulty matched participant ability,
- (b) how well each item fitted the single parameter Rasch model by providing *infit* and *outfit* values,
- (c) the behaviour of distracter items,
- (d) difference between expected and observed item measures, with an additional map, allowing unexpected responses (an indication of possible guessing) to be identified,
- (e) *differential item functioning* (DIF) demonstrating the extent to which different sample sub-sets (e.g., boys and girls) responded differently to certain items.

This analysis allowed suitability of each item for measurement to be assessed, indicating items that needed modification or rejection.

To illustrate the usefulness of IRT analysis, Fig. 7 shows the Person/Item map for one version of the first listening task (version 1 of the multiple-choice Listening 1 task in Table 3). Participants are placed on the left of the dividing line, from less able at the bottom to more able placed towards the top of the map. The items are placed on the right, from the easiest at the bottom to more difficult to the top of the map. The mean measure of item difficulty at 0.00 logit was only slightly lower than the mean measure for person ability, suggesting a good match between task difficulty and participant ability. Ability ranged from -3 to $+4$ logits, whilst item measures ranged from -1.26 to $+2.03$. This suggests that there were participants whose ability exceeded the difficulty of most difficult items and some whose ability fell below the difficulty of the easiest items. The map allows identification of these items and to assess the number of participants outside the task range. In the case of this version of the first listening task, the map shows that almost everyone answered item 3 correctly, whilst items 5, 8 and 10 were difficult. The map illustrates how 6 % of children in the upper range of ability were above the range of the test, i.e., over scale, and almost 3 % of children were below the ability required for the easiest item.

As a result of the analysis, two items were removed from this task: a difficult item 10 and item 18, of average difficulty. Although the *infit* and *outfit* values for all items fell within the range of 0.5–1.5 which, according to Linacre (2012), is deemed productive for measurement, both items had the highest *infit* and *outfit* values: 1.12 and 1.26 for item 18 and 1.10 and 1.27 for item 10. According to Classical Test Theory, these items also had the lowest discrimination values: .08 for item 18 and .12 for item 10, suggesting that both qualified for rejection or substantial change. Additionally, item 10 was scored correctly by a number of participants whose scores were otherwise weak.

In addition to the first version of the multiple-choice listening comprehension task, modified by the items discussed above, as a result of detailed pilot item analysis, the following tasks were selected for the final test:

- (a) the fourth version of the true/false task '*In the classroom*'
- (b) the first version of the multiple-choice reading comprehension task reduced by two items
- (c) the first version of the picture and text matching reading comprehension task '*The cat and mouse story*'.

All pilot versions of the third reading comprehension task (title and text matching) were rejected and a new version of the task was constructed and piloted with 20 children of the target age group. Time considerations did not permit a larger sample for this second pilot.

7 The Final Test

Following the pilot and re-piloting of certain items, the finished product could be regarded as not only the task versions demonstrating the best reliability and pupil differentiation but also the plan and instructions for test administrator recruitment and training, the procedures, collection of scripts, coding and quality control. Analysis of the nationwide test was to follow a strategy similar to the one employed to assess the candidate versions. The same statistical tools and methods for item analysis were to be used. The same criteria were to be applied to items as in the pilot, since on a larger scale anomalies might be observed which would not be visible at the smaller pilot scale. Final dissemination of the findings is planned to coincide with a conference together with a published report written with all stakeholders in mind. Sound database design is needed for the final results and associated contextual data. The tools required for this should be based on relational database technology to allow the use of SQL to select subsamples of pupil and teacher data according to chosen selection criteria.

8 Conclusions

This chapter described some solutions to the problems associated with the creation of a large-scale language test designed, piloted and administered to young learners as part of an empirical study. Beyond the general difficulty of ensuring the usefulness of a language test from the perspective of the young learners, the team of test developers faced the following challenges: (1) How to create interesting and age-appropriate test items from a very limited volume of common vocabulary; (2) How to reconcile learners' well-developed cognitive skills with their low level of foreign language knowledge in order to create test materials; (3) How to encourage willing

participation and an ensuing sustained high level of intellectual engagement with a test from which there would be no tangible reward for individuals. In other words, how to ensure that participants try their best throughout the test; and finally, (4) What message and what type of organisation would best assure this.

Several aspects of the design process need to be particularly emphasised. The first is the careful analysis of items using IRT to ensure a choice with the best psychometric qualities. The second is the enormous value of cognitive laboratories to obtain young learners' perspectives on planned tests. These interviews cast doubt on many adult assumptions about the visual and linguistic content of the test, thus saving resources and ensuring the effectiveness and adequacy of the subsequent pilot. Cognitive interviews with the target age group are vital at pre-pilot stage for any similar assessment. Finally, administration of a mass-delivered test must be homogeneous and conducted in a sympathetic manner likely to encourage children to cooperate and try their best without fear. Since researchers do not necessarily have experience of this type of test conditions, it should not be assumed that they would share the same image of their role. Therefore, the importance of well-planned training and preparation should be intrinsic to planning for the study, for which simulation and authentic videos should complement explanation.

Considering the scale and the complexity of such a task, careful planning and execution of all steps in the process are vital to its success, possible only through good will, trust and cooperation between all the players at all levels in the process.

9 Need for Future Research

This study has highlighted the importance of the child perspective in terms of linguistic, visual and pragmatic content of test item, the need for target-age group consultation and careful piloting of items and test procedures. Future research should give attention to these aspects of large-scale measurement of children's foreign language and attempt to explore ways how such measurement could better account for the variety of lesson content, course materials and learning experiences of young foreign language learners in instructional settings. Full verification of assessment should include follow up, particularly of outliers.




Appendix

TEST SZKOLENIOWY




Zadanie 1 Słuchanie

Zaznacz rysunek, który przedstawia to, co usłyszysz w nagraniu. Zamaluj literę przy rysunku. Każde nagranie usłyszysz 2 razy.

1

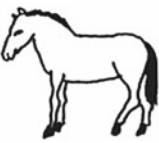


(A) 	(B) 	(C) 
--	--	--

1

(A) 	(B) 	(C) 
--	--	--

Jeżeli się pomyliłz przekreśl błędną odpowiedź krzyżykiem i zaznacz poprawną!

1



(A) 	(B) 	(C) 
---	---	--

Zadanie 3 Czytanie

Z podanych trzech zdań wybierz jedno, które opisuje obrazek.

Zaznacz wybrane zdanie.

Jeżeli się pomyliłz przekreśl błędną odpowiedź krzyżykiem i zaznacz poprawną!

1 	(A) It's a cup. (B) It's a flower. (C) It's a cake.	1 	(A) It's a cup. (B) It's a flower. (C) It's a cake.
--	---	--	--

References

- Alexiou, T., & Konstantakis, N. (2007, July). Vocabulary in Greek EFL young learners' course books. Paper delivered to ESCR Seminar: *Models and concepts, practical needs and theoretical approaches in modelling and measuring vocabulary knowledge*. Swansea University, Swansea, Wales.
- Ashman, A. F., & Conway, R. N. F. (1997). *An introduction to cognitive education*. London: Routledge.
- Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, UK: Cambridge University Press.
- Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford, NY: Oxford University Press.
- Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2), 287–311.
- Bredenkamp, S., & Copple, C. (Eds.). (1997). *Developmentally appropriate practice in early childhood programs*. Washington, DC: National Association for the Education of Young Children.
- 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*. New York: Springer.
- Butler, F. A., & Stevens, R. (2001). Standardized assessment of the content knowledge of English language learners K-12: Current trends and old dilemmas. *Language Testing*, 18(4), 409–427.
- Campfield, D. E. (in preparation). *Function words and lexical difficulty – Using Elicited imitation to study child L2*.
- Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford, NY: Oxford University Press.
- Enns, J. T., & Trick, L. M. (2006). Four modes of selection. In E. Bialystok & F. I. M. Craik (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 43–56). New York: Oxford University Press.
- Espinoza, L. M., & Lopez, M. L. (2007, August). *Assessment considerations for young English language learners across different levels of accountability*. Paper prepared for The National Early Childhood Accountability Task Force and First 5 LA. Retrieved from <http://www.first5la.org/files/AssessmentConsiderationsEnglishLearners.pdf>
- Hsieh, C.-N. (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*. New York: Springer.
- Inbar-Lourie, O., & Shohamy, E. (2009). Assessing young language learners: What is the construct? In M. Nikolov (Ed.), *Contextualizing the age factor: Issues in early foreign language learning* (pp. 83–96). New York: Mouton de Gruyter.
- Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. J. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona conference, July 2001* (pp. 262–284). Cambridge, UK: Cambridge University Press.
- Kulas, K. (2012, July). *The selection of vocabulary for EFL lower-primary school textbooks*. In 10th Teaching and language corpora conference, The Institute of Applied Linguistics, University of Warsaw, Warsaw, Poland.
- Linacre, J. (2012). *Practical Rasch measurement*. Retrieved from www.winsteps.com/tutorials.htm
- 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.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5–11.
- Mihaljević Djigunović, J., & Lopriore, L. (2011). The learner: Do individual differences matter? In J. Enever (Ed.), *ELLiE. Early language learning in Europe* (pp. 43–60). London: British Council.

- Mihaljević Djigunović, J., & Vilke, M. (2000). Eight years after: Wishful thinking vs. the facts of life. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 67–86). Pécs, Hungary: University Press Pécs.
- Ministerstwo Edukacji Narodowej (MEN). (2002). Rozporządzenie Ministra Edukacji Narodowej i Sportu z dnia 26 lutego 2002 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół (Dz. U. z 9 maja 2002 r. Nr 51, poz. 458).
- Ministerstwo Edukacji Narodowej (MEN). (2008). Rozporządzenie Ministra Edukacji Narodowej z dnia 23 grudnia 2008 r. w sprawie podstawy programowej wychowania przedszkolnego oraz kształcenia ogólnego w poszczególnych typach szkół. Dz.U. nr 4 z dn. 15 stycznia 2009. Warszawa, Poland: Kancelaria Prezesa Rady Ministrów.
- Muñoz, C., & Lindgren, E. (2011). Out-of-school factors: The home. In J. Enever (Ed.), *ELLiE. Early language learning in Europe* (pp. 103–124). London: British Council.
- Muñoz, C., & Lindgren, E. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, 10, 105–129.
- Nikolov, M. (2016). 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*. New York: Springer.
- 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* (pp. 347–363). Cambridge, UK: UCLES/Cambridge University Press.
- Pamuła, M., Bajorek, A., Bartosz-Przybyło, I., & Sikora-Banasik, D. (2006). *Europejskie portfolio językowe dla dzieci od 6 do 10 lat*. Warszawa, Poland: Centralny Ośrodek Doskonalenia Nauczycieli.
- Redline, C., Smiley, R., Lee, M., DeMaio, T., & Dillman, D. (1998). Beyond concurrent interviews: An evaluation of cognitive interviewing techniques for self-administered questionnaires. *Proceedings of the section on survey research methods* (pp. 900–905), Alexandria, VA: American Statistical Association. Retrieved from https://www.amstat.org/sections/SRMS/Proceedings/papers/1998_155.pdf
- Rixon, S. (1999). Where do the words in EYL textbooks come from? In S. Rixon (Ed.), *Young learners of English: Some research perspectives* (pp. 55–71). Harlow, UK: Longman.
- Schaffer, H. R. (2004). *Introducing child psychology*. Oxford, UK: Blackwell.
- Szpotowicz, M., & Lindgren, E. (2011). Language achievements: A longitudinal perspective. In J. Enever (Ed.), *ELLiE. Early language learning in Europe* (pp. 125–143). London: British Council.
- Taylor, L., & Saville, N. (2002). *Developing English language tests for young learners* (Research Notes 7, pp. 3–6). Cambridge, UK: UCLES.
- Vernon, H., & Vernon, M. (Eds.). (1976). *The development of cognitive processes*. London: Academic.
- Wesson, K. (2011). *Attention span revisited*. Retrieved from <http://sciencemaster77.blogspot.com/2011/01/attention-spans-revisited.htm>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

The Development and Validation of a Computer-Based Test of English for Young Learners: Cambridge English Young Learners

Szilvia Papp and Agnieszka Walczak

Abstract This chapter summarises the rationale for the development and validation work that took place over 2.5 years before the launch of the computer-based (CB) format of the *Cambridge English Young Learners* English tests (YLE). Several rounds of trials were carried out in a cyclical way, in a number of different locations across various countries, to ensure data was collected from a representative sample of candidates in terms of geographical location, age, L1, language ability, familiarity with the YLE tests, and experience of using different computer devices – PC, laptop and tablet. Validity evidence is presented from an empirical study, using a convergent mixed methods design to explore candidate performance in and reaction to the CB YLE tests. Regression analyses were conducted to investigate which individual test taker characteristics contribute to candidate performance in CB YLE tests. The results indicate that CB delivery presents a genuine choice for candidates in line with the Cambridge English ‘bias for best’ principle. Positive feedback from trial candidates, parents and examiners suggests that CB YLE tests offer a contemporary, fun, and accessible alternative to paper-based (PB) YLE tests to assess children’s English language ability.

Keywords Young learners • Computer-based assessment • English language • Test taker characteristics • Effect on performance • Regression analysis

1 Introduction

This chapter contributes to the volume’s theme by describing how a computer-based (CB) format of *Cambridge English Young Learners* tests (YLE), an existing suite of exams for young learners, was developed, piloted and validated. YLE tests were developed for primary and lower secondary school learners of English between the

S. Papp (✉) • A. Walczak

Research and Validation Group, Cambridge English Language Assessment, Cambridge, UK

e-mail: Papp.S@cambridgeenglish.org; szilvia@ecreationtech.co.uk;

Walczak.A@cambridgeenglish.co.uk

ages of 7 and 12. The tests are available in three levels: Starters, Movers and Flyers, set at levels pre-A1 to A2 of the Council of Europe's *Common European Framework of Reference* (CEFR, Council of Europe, 2001). YLE tests measure achievement in four skills in 3 papers: (a) Listening, (b) Speaking, (c) combined Reading and Writing. Candidates receive a certificate that indicates their level of success in the test through showing a number of shields for each section of the test. The maximum number of shields awarded for each section is five so a candidate could score a maximum of fifteen shields per test. Achieving five shields indicates a very strong performance on the test. A score of three shields indicates that candidates are performing at the level intended by the test. In order to provide motivation to the young children taking the test, all candidates are awarded at least one shield for each section. It is not possible to 'fail' the test.

First, we discuss the rationale for developing a CB version of the tests. Next, we discuss what methodological issues were considered in the trialling and validation of the CB format. Then, in a mixed methods enquiry, we report on some validation evidence generated by investigating how candidates' performances are related to individual differences (age, gender and preference for, and experience of, computer use), and what candidates, parents and observers said about CB YLE tests. The convergent mixed methods design allows us to triangulate the results and consider evidence from various sources to mutually inform our interpretations of the data.

2 Why Develop Computer-Based Tests for Young Learners?

Cambridge English Language Assessment endeavors to ensure that language tests support the work of the wider education communities and their policies in which they are taken. This is especially relevant for the *Cambridge English Young Learners* tests where target candidates are learners of English in primary and lower secondary schools between the ages of 7–12.

One of the policies in many primary and secondary education systems is to include information technology (IT) objectives as part of national schools curricula. For instance, in Hong Kong, the *General Studies for Primary Curriculum* (Hong Kong Special Administrative Region Education Bureau Information Services Department, 2014; Hong Kong Special Administrative Region Government Education Bureau n.d.) advises Hong Kong schools to adhere to the current strategy on IT in education. The aims of that strategy are for children to become information literate, to become competent in using IT both for learning and in daily life. For Hong Kong schools the aim is to integrate IT into teaching and learning across the whole curriculum, and for parents the stated aim is to make IT accessible to their children and to help develop their information literacy. The provision of a computer-delivered version of YLE clearly supports these wider educational goals.

These aims reflect the fact that children's formal and informal learning is increasingly mediated through technology. Teachers around the world routinely ask children to use the internet to help complete school projects or homework.

Children regularly ask to borrow their parents' smartphones, tablets or laptops so they can play educational games and apps. In this way, technology is increasingly integrated into children's day-to-day learning and everyday lives. Offering children the opportunity to take tests on computers reflects this integration and this shift is likely to impact on children's expectations of assessment.

As a result of the ubiquity of digital media, technologies and services since the turn of the millennium, a new generation of learners has grown up who are variously labelled 'digital natives' (Prensky, 2001) or 'new millennium learners' (OECD/CERI, 2008). There is thus a growing generational difference in the frequency and experience of computer use between young learners and their older counterparts (Becker, 2000; Pedró, 2006, 2007; Rideout, Vandewater, & Wartella, 2003).

This may have an effect on young learners' attitudes, ability and confidence in taking tests in paper-based (PB) or CB delivery mode. These young learners interact with digital technologies from a very early age, have more experience and thus familiarity with technological devices, and therefore feel more comfortable dealing with them in comparison with older learners or adults.

However, when considering this new generation of learners, it needs to be remembered that not all young people have access to digital technologies and there may be differences based on socio-economic status, geographical area and gender. Parental and peer attitudes and social values, as well as individual preferences based on learning styles and strategies, will also have an impact on how subgroups and individuals relate to digital media within this generation (OECD/CERI, 2008). The development of CB YLE has offered an opportunity to investigate such attitudes and preferences regarding PB and CB tests.

Such an investigation demonstrates the Cambridge English commitment to providing tests that are both fair and useful to schools, parents and children. Developing a CB version of YLE means a greater choice of test dates for schools, faster results for parents and a test that better reflects how children are learning today. According to the Cambridge English 'bias for best' principle (Jones & Maycock, 2007, p. 12), tests in different formats or modes of delivery present equality of opportunity, i.e. the opportunity to select the test format that offers children the best prospect of performing at their best.

Therefore, the development of CB and computer-adaptive assessments for young learners is a particularly promising enterprise for testing agencies and examinations boards such as Cambridge English Language Assessment. Technology and CB assessment have the potential to inform teaching and unobtrusively monitor and guide learning (Bennett, 1998; Jones, 2006; Tymms, Merrell, & Hawker, 2012). Technology, especially if the assessment is adaptive, can help turn the teaching, learning and assessment cycle into a truly integral system, where assessment has the role of 'feeding into' teaching and learning, which in turn shape subsequent assessment in an iterative fashion. Thus adaptive CB testing has the potential to ensure that assessment genuinely supports teaching and learning – an ethical imperative – while at the same time providing the right amount of challenge for young learners for their learning to be efficient and successful, keeping them engaged and motivated. Adaptive assessment can also provide the information needed by parents on children's progress over time.

Cambridge English has produced CB tests in CD-ROM format since 1995, for instance, the adaptive *CB BULATS* (Business Language Testing Service) or the *QPT* (Quick Placement Test). Cambridge English initially used the CB format in low-stakes testing, typically for shorter tests that were not certificated and where the test administration was not supervised (Jones, 2000). However, higher-stakes tests have also been delivered in CB format, including *Cambridge English Skills for Life*, *Teaching Knowledge Test* (TKT), *Cambridge English Key* (KET), *Cambridge English Preliminary* (PET), *Business English Certificate* (BEC) *Preliminary* and *Business English Certificate* (BEC) *Vantage*. CB delivery of the *for Schools* versions of KET and PET was introduced in April 2010, only a year after their launch in PB format in March 2009. Similarly, soon after *Cambridge English First for Schools* was launched, its CB format was introduced in March 2012. The development of CB delivery of *Cambridge English Young Learners* started in late 2011, with a series of trials carried out between 2012 and 2014. These CB tests are computer-mediated linear tests. Cambridge English continues to be engaged in developing a range of computer-adaptive tests, such as *Business Language Testing Service* (BULATS), and various placement tests in progress (e.g., *Cambridge English Placement Test* (CEPT), *Cambridge English Placement Test for Young Learners*).

Before developing a CB version of a test to offer an alternative to the PB delivery mode, test providers need to carry out research to investigate comparability of the two delivery methods, which we discuss in the next section.

3 Methodological Issues in the Validation of Computer-Based Tests for Young Learners

3.1 *The Case for Comparability Studies*

The extent to which PB and CB formats of a test measure the same trait determines whether they can replace each other (Clariana & Wallace, 2002; McDonald, 2002; Neuman & Baydoun, 1998; Pommerich, 2004; Pomplun, Frey, & Becker, 2000; Wang, Jiao, Young, Brooks, & Olson, 2007; Zandvliet, 1997). Jones and Maycock (2007, p. 11) note that the goal of comparability studies can be to inform test users that

1. the PB and CB format can be used interchangeably
2. they differ to some extent for practical reasons inherent to the PB and CB formats
3. their designs differ so that one may be considered better than the other for some purposes.

In all comparability studies between Cambridge English PB and CB test formats, Rasch modelling has been used as a measurement tool. Item banking techniques

generally ensure that when items are made available for use in a CB test, their difficulty is known as they have been calibrated (i.e. their difficulty has been estimated) on a scale. Thus, it is possible to compare the difficulty of items in the two formats (Jones & Maycock, 2007, p. 11).

In experimental conditions, where the two test formats are completed one after the other, the sequence effect may produce variations in performance due to fatigue, inattention, etc. Hence test order is always controlled for in a counterbalanced research design.

In order to gauge attitudinal and preference data on each delivery format, candidates are usually asked to fill in a questionnaire or to participate in a focus group covering their perception of test difficulty in the two formats, the appropriateness of the length of the test, and background variables such as their attitudes (likes and preferences) as well as their familiarity with, ability, experience and confidence in using computers (Jones, 2000; Maycock & Green, 2005).

Candidate perceptions, preferences and attitudes are revealing as they reflect the extent to which candidates feel at ease with either format and which format they feel allows them to best demonstrate their language ability. Research, however, has found that these perceptions, preferences and attitudes tend not to have an effect on candidate scores in either format (Jones, 2000; Maycock & Green, 2005; O'Sullivan, Weir, & Yan, 2004; Taylor, Jamieson, Eignor, & Kirsch, 1998).

3.2 *Are Young Learners Different?*

As indicated in the introduction, young learners may perform differently from adults in CB tests and their attitudes may also be different to them. Younger candidates are more familiar with keyboard technology than their adult counterparts, as pointed out by Hackett (2005) before the launch of *CB PET*. Among 190 trial candidates aged 20 or under, most (67 %) found typing as easy or easier than having to write by hand in *PET* Writing Parts 2 and 3. Candidates found on-screen reading easier than on paper (46 % vs 25 %). Listening individually through headphones was preferred (by 87 %) to listening from a CD in a group. It is interesting to compare the *CB PET* candidates' overall preferences (63 % for CB Reading & Writing and 83 % for CB Listening) with adults' views in the CB IELTS trials running at the same time (Maycock & Green, 2005). Adult candidates perceived computer familiarity (i.e., good computer and keyboard skills) to be an advantage in the CB format of IELTS. Out of 882 candidates aged 16 or over, only 33 % said they can type faster than write, 48 % said they can handwrite faster than type, and 19 % claimed to have the same speed in both formats. Despite this, over half of the IELTS candidates preferred the CB format.

The argument that younger candidates are more computer literate than adults can be taken further in relation to children. Computer use has become so widespread among school learners that nowadays the issue may be not lack of familiarity with using computers but lack of familiarity with using paper and pencil. Students may

be more familiar with reading and typing on the computer than with reading and writing on paper, due to the frequency of online activities in learners' lives. Russell and colleagues have repeatedly found that students in US schools perform better on computers (e.g., Russell & Haney, 1997). This has led them to consider whether writing on paper is less of a 'real world' task (cf, Chapelle & Douglas, 2006; Lee, 2004; Li, 2006). It is worth noting that some findings in European schools differed from this: Endres (2012) found that, while 12–16 year-old Spanish learners of English tend to use computers for leisure and informal communication, they do not use it as much for schoolwork and homework.

Apart from the design features common to all comparability studies noted above, studies among young learners need to use methods of enquiry familiar to and widely accepted by early childhood professionals. Thus, all methods used in the validation of CB YLE were modelled on "best practices", complying with relevant ethical guidelines on research with children (e.g., British Educational Research Association, 2011; British Psychological Society, 2009; Economic and Social Research Council, 2012; European Commission Information Society Technologies, n.d.; National Association for the Education of Young Children, 2009; Social Research Association, 2003). For instance, it was considered that children may need help filling in questionnaires even if delivered in their L1. A focus group discussion may be more appropriate. Alternative ways of eliciting data from children were also considered (e.g., see Sim, Holfield, & Brown, 2004; Sim & Horton, 2005). For those children who may not feel comfortable responding verbally, drawing may be an alternative way of eliciting responses (Wall, Higgins, & Tiplady, 2009). In addition, individual debriefing interview sessions may be more suitable with younger children where open-ended questions allow children to respond using their own words (Barnes, 2010a, 2010b).

4 Development and Validation of the CB Version of *Cambridge English Young Learners Tests*

4.1 *Trial Methodology*

To create the CB version of *Cambridge English Young Learners tests* Cambridge English spent 2.5 years developing and trialling different versions to ensure the tests are as intuitive, accessible and user-friendly as possible. During the development phase, the CB YLE tests were trialled in a number of different locations (China, Hong Kong, Mexico, Spain, Argentina, Italy, Turkey, Macau) with over 1800 candidates. The aim was to gain a sample as representative as possible in terms of age, gender, language ability, familiarity with the YLE tests, familiarity with computers etc. Ensuring that CB YLE was trialled in several cultural and educational contexts (state and private) across different L1s with a wide range of candidates enhances the generalizability of the results. After each trial, test results were analysed and

feedback was collected from about 650 candidates on questionnaires, as well as about 64 observers (participating examiners, test administrators, ushers, teachers or external observers) on checklists and surveys. Each time adjustments and improvements were made in light of the findings and then the tests were trialled again. A constant feature throughout the development and trialling process was the effort made to ensure that the CB test was comparable to the PB test. The focus throughout was confirming that, like the PB tests, the computer-delivered version would provide an accurate, consistent and reliable measure of children's language ability.

In all CB YLE development trials a convergent mixed-methods research design was used (Creswell & Plano Clark, 2011). Use of mixed methods allows the merger of quantitative (exam performance data) and qualitative data (information on context, setting, participants gathered through questionnaires, testimonials, focus group interviews, surveys). The various types of evidence from these data sources were used in a convergent design to mutually inform each strand of enquiry and to triangulate results (Creswell & Plano Clark, p. 118). Use of merged results produces better understanding and mutually confirms findings, and ultimately provides validation and validity evidence for CB YLE. Data types, sources, and analyses used in the CB YLE trials are displayed in Fig. 1.

4.2 Candidate Profile in This Study

In this study, we report on a set of regression analyses to investigate what learner characteristics impacted on achievement in PB and CB YLE tests during the trials. Only those candidates for whom we had both questionnaire and CB YLE test performance data are included. Trial candidates in China, Hong Kong, Argentina and Macau are not included as they were not asked to provide data on their background and attitudes to CB testing in the trials. Table 1 shows the total number of candidates in *Starters*, *Movers* and *Flyers* in the regression analyses. The table presents the breakdown of candidate numbers in percentages by country. Altogether 135 candidates from Mexico participated in the *Movers* and *Flyers* trials (forming 22 % of all candidates taking part in the trials). Mexican candidates were not included in the *Starters* trial as the CB YLE *Starters* test was still in development at the time of the Mexico trial. Altogether 219 Spanish, 136 Italian, and 120 Turkish candidates took part in the trials at all levels (making up 36 %, 22 %, and 20 %, respectively, of the total candidates in the trials).

Table 2 shows gender distribution of candidates at each level in the study sample. In *Starters* and *Movers* male and female candidates are nearly equally distributed. In *Flyers*, there were more female candidates than males.

Candidates were given the test in paper-based (PB) and computer-based (CB) format. Table 3 shows percentage of the total trial candidates at each level with scores in both PB and CB tests.

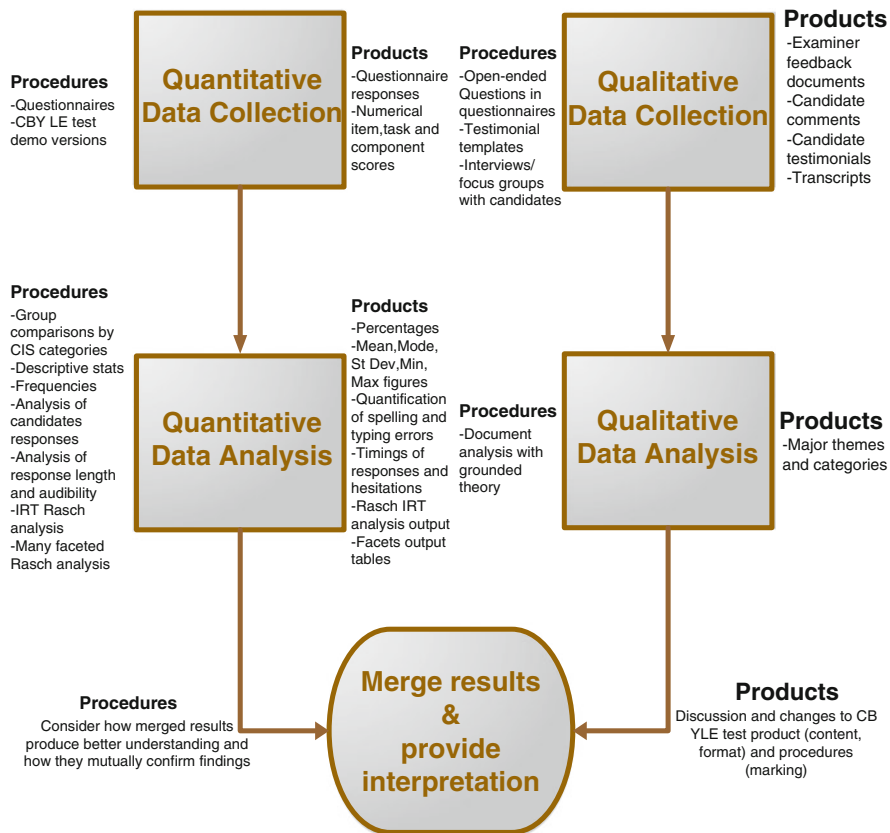


Fig. 1 Mixed method research design procedures and products in CB YLE development trials (Adapted from Creswell & Plano Clark, 2011, p. 118)

Table 1 Percentage of candidates per exam level and country

	Mexico (%)	Spain (%)	Italy (%)	Turkey (%)	Total
Starters	–	33	49	18	216
Movers	31	38	10	21	214
Flyers	38	37	5	19	180
Total	22	36	22	20	610

Table 2 Percentage of male and female candidates at each level

	Male (%)	Female (%)
Starters	52	48
Movers	44	56
Flyers	38	62

Table 3 Percentage of candidates who took a CB and PB test

	% of candidates who took CB and PB
Starters	27 %
Movers	65 %
Flyers	75 %

Table 4 Percentage of *Movers* and *Flyers* candidates by device

	IPAD (%)	PC (%)
Movers	26	21
Flyers	20	37
Total	23	28

Table 4 shows the percentages of candidates who took the YLE test on an ipad (tablet) and on a PC. Combining *Movers* and *Flyers*, 23 % of candidates (N=90) took the test on an ipad, while 28 % of candidates (N=110) took the test on a PC.

4.3 Data Analysis

Mixed method research involved the following specific steps in the CB/PB YLE comparability study:

1. Correlations among PB and CB scores by level (*Starters*, *Movers*, *Flyers*) and component (Listening; L, Reading & Writing; RW, Speaking; S).
2. Regression analyses (Fox, 2002, 2008) on combined exam score data and candidate background and attitudinal data collected through questionnaires (332 candidates in Mexico, Spain, Italy and Turkey, see Appendix A for candidate questionnaire).
3. Analysis of verbal feedback and drawings in questionnaires and testimonials provided by trial candidates and their parents (126 candidates and their parents from Hong Kong, Mexico, Spain).
4. Analysis of trial observer feedback (64 observers from Hong Kong, Spain, Italy, see Appendix B for observer checklist).

Table 5 summarises the techniques of data collection and analysis in the PB/CB YLE comparability study.

The regression analyses reported below used both quantitative (exam score data, candidate background information) and qualitative data (experiential and attitudinal data related to computer use and CB tests). The following variables were used in the quantitative regression analyses:

- Dependent variables:
 - total score in CB test for *Starters* (Model 0),
 - total score in PB test for *Flyers* (Model 1),

Table 5 Overview of research areas, data types and sources, instruments and analyses in PB/CB YLE comparability studies

Research area	Data type	Data source and instruments	Data analysis
1. Candidate and parent attitude to CB YLE	Quan	1. Candidate questionnaires	1. Frequencies of questionnaire
	Qual	2. Focus group interviews with trial candidates	2. closed responses
		3. Candidate and parental testimonials	3. Regression analyses
		4. Candidate drawings	4. Document thematic analysis
		5. Photos and video footage of trials	
2. Observations from trials	Quan	1. Observer checklist	1. Frequencies of checklist closed responses
3. Examiner attitude to CB YLE	Qual	2. Summary of observer verbal comments for action	2. Document thematic analysis
	Quan	1. Examiner survey	1. Frequencies of closed responses
4. Candidate background information	Qual	2. Soft feedback in reports and emails	2. Document thematic analysis
	Quan	1. Candidate background data elicited in questionnaires and testimonials	1. Analysis of background data in questionnaires and testimonials
		2. Candidate information sheet (CIS) provided on exam entry forms	2. Analysis of CIS data
			3. Descriptive statistics
5. Candidate exam performance (CB and PB)	Quan	1. Recorded candidate responses (L, RW) (CB response files and PB scripts)	1. Candidate written response analysis (typing errors, spelling)
	Qual	2. Transcription of audio-recorded Recorded candidate responses (L, RW) (CB response files and PB scripts)	2. Analysis of transcripts for hesitation and examiner feedback/support
			3. Timings of candidate responses

6. Scoring and marking (CB and PB)	Quan Qual	1. Marking keys (L, RW) 2. Test scores at item, task and component level (L, RW) (CB and PB) 3. Speaking examiner scores (at criterion level and total) (S)	1. Document analysis 2. Descriptive and Classical analysis of score data, e.g., facilities, discrimination (CB and PB) 3. IRT Rasch analysis (PB L, PB RW) 4. Correlations (PB vs CB) 5. Regression analyses 6. Multi-faceted Rasch analysis with Facets (S)
7. Changes in CB YLE test content and marking during test development	Qual	1. Test development and trialling procedural documents 2. Successive demo versions of CB YLE tests	1. Procedural document analysis 2. Documentation on changes made to test content and test delivery systems (e.g., entry portal, examiner portal)

Caption: Quan = quantitative data type, Qual = qualitative data type

- total score in CB test for *Flyers* (Models 2,3,4),
 - total score in CB test for *Flyers* and *Movers* (Model 5)
- Explanatory variables:
- Age of candidates on the day of the test (in years)
 - Gender ('female' used as a baseline for comparison in regression analyses)
 - Number of years candidates have been taking English language classes (Years of English instruction)
 - Preference for exam delivery (response categories are 'on paper' as a baseline for comparison, 'no difference' and 'on computer')
 - Frequency of computer use ('only at weekends' as a baseline, 'every day', 'once or twice a week')
 - Purpose of computer use (using computers for English homework, for playing games, for email/chat; for other activities)
 - Type of computer at home (Desktop (PC/Mac) as a baseline; Desktop/Laptop, Desktop/Tablet, Desktop/Tablet/Laptop, Laptop, Tablet and Tablet/Laptop)

The variables elicited through the questionnaires were included in the hope that they would offer some insight into the differences in performance on PB and CB YLE tests across age, gender, frequency and purpose of computer use among young learners. We did not ask trial candidates what actually they do when they use computers, how they use IT resources available to them, and to what extent their use of computers is linked to additional exposure to English. We are aware of the limitations of the explanatory variables used in the study. However, they provide some insight into the link between performance on a computer-delivered test and computer use, as we explain below.

5 Trial Results: Quantitative

How candidate performance in PB and CB YLE is related to individual characteristics was investigated by addressing two the following research questions:

Research Question 1: Are the scores on PB and CB comparable?

In order to investigate whether the PB and CB YLE tests are comparable, the data was analysed from two perspectives:

- firstly, for the relationship between scores in the PB and CB YLE tests;
- secondly, to see which variables explain candidate performance in the PB and CB YLE test for each level (*Starters*, *Movers*, *Flyers*).

Research Question 2: Does candidate performance in the CB test vary according to the type of device on which the candidates took the CB test?

To answer these research questions, we employed a series of regression analyses to explore the effect of background, experiential and attitudinal variables on candidate performance, controlling for factors other than the effect of delivery mode on candidate performance.

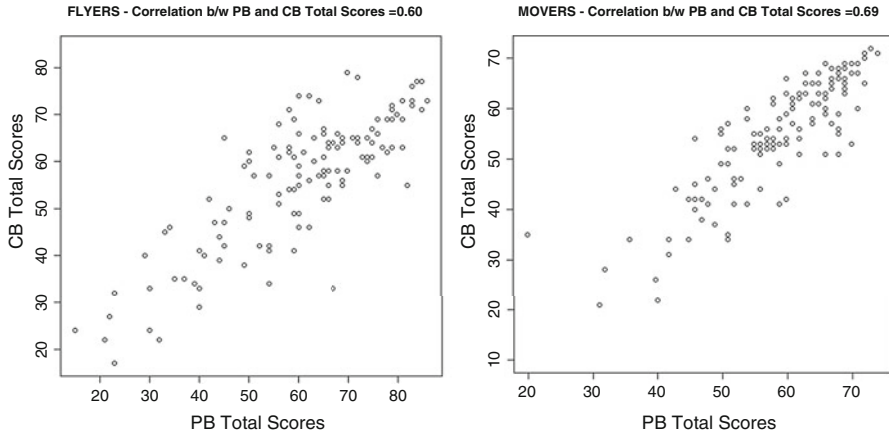


Fig. 2 Correlation between PB and CB total scores for *Flyers* and *Movers*

5.1 Relationship Between PB and CB Scores

First the relationship between the total scores in PB and CB exams was investigated. This analysis is based on PB and CB scores for *Movers* and *Flyers* (N=274). Figure 2 shows that for *Flyers* the correlation was 0.60, and for *Movers* it was 0.69. This provides evidence of the extent of comparability between PB and CB YLE tests during the trials. Please note that during the trials candidates had not been familiarised with the computer-based delivery of YLE, so these otherwise modest correlations were encouraging for when sample practice tests (now freely available as apps on AppleStore) were made available for candidates. These offer guidance on how to take CB YLE and provide advance practice on functionality for candidates.

5.2 Variables Explaining Candidate Performance in PB and CB Tests

Figures 3 and 4 show the distribution of CB total scores for *Starters* and *Flyers* by country. The data is presented in the form of boxplots. Boxplots show the distribution of data for each category. The rectangles show the distribution of data from the 1st to 3rd quartile, where the bottom side of the rectangle represents the 1st quartile (25th percentile) and the upper line represents the 3rd quartile (75th percentile). The thick black horizontal line shows the median in the data. The vertical dashed lines – the whiskers – show the range of the data. Outliers are indicated with dots beyond the whiskers.

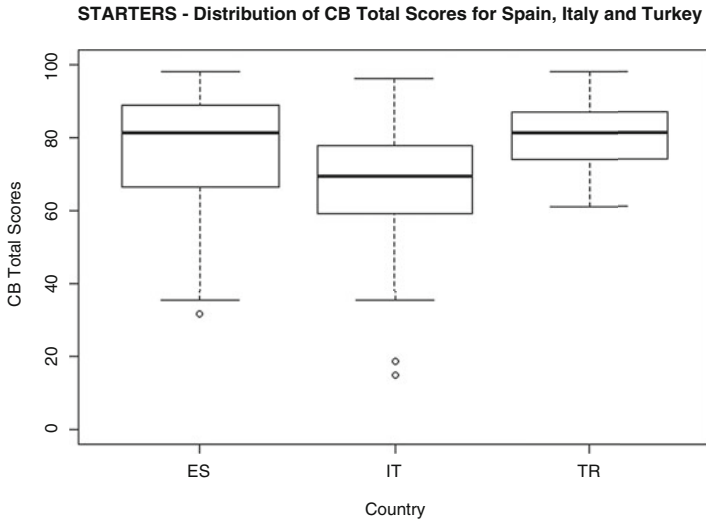


Fig. 3 Distribution of CB scores by country in *Starters*

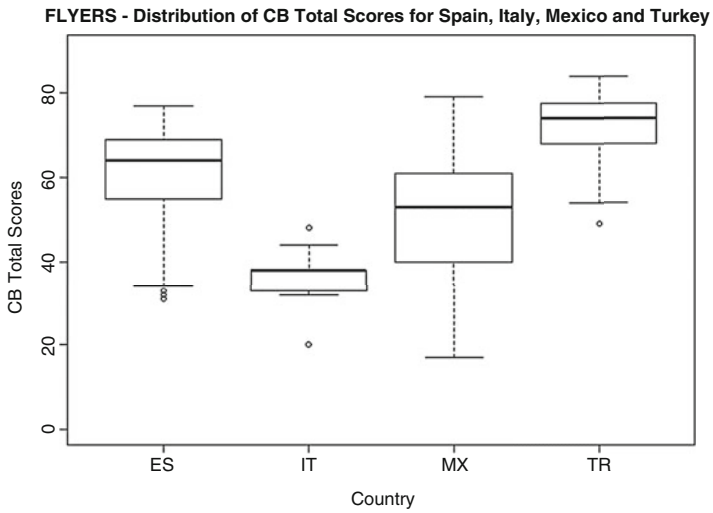


Fig. 4 Distribution of CB scores by country in *Flyers*

5.2.1 Country

There were candidates from Spain, Italy, Turkey taking CB *Starters*, plus candidates from Mexico taking CB *Flyers*. Figure 3 shows that *Starters* candidates' CB scores vary slightly according to country. There is evidence of differences between countries in results of general educational assessments among school learners (e.g., Merrell & Tymms, 2007; Tymms & Merrell, 2009). When we look at the range of scores in Fig. 4 for CB *Flyers*, we see that Turkish candidates perform the best in the sample, followed by candidates from Spain. In order to account for differences in performance across countries we included dummy variables for each country in all the regression analyses below.

Next, we report the results of regression analyses carried out on the data for *Starters* and *Flyers*, with the aim of identifying variables that explain candidate performance in each test. The dependent variable in these regressions is (1) total score on the PB test and (2) total score on the CB test.

5.2.2 Age

To investigate the effect of age on PB and CB test performance, scores and candidate age were plotted against each other. As can be seen in Fig. 5 for *Flyers* there was a clear curvilinear relationship between age and scores. This shows that the older the candidates are after age 11, the lower their scores are in both PB and CB tests during the comparability trial. The target candidature for *Cambridge English Young Learners* is up to age 12. Here we see evidence that candidates older than age 11 and a half may have been inadvertently affected by motivational and affective variables: they may not have taken the tests seriously. Due to the curvilinear relationship between candidate age and performance on PB and CB scores in *Flyers*, the regression analyses include a variable Age Squared to account for this.

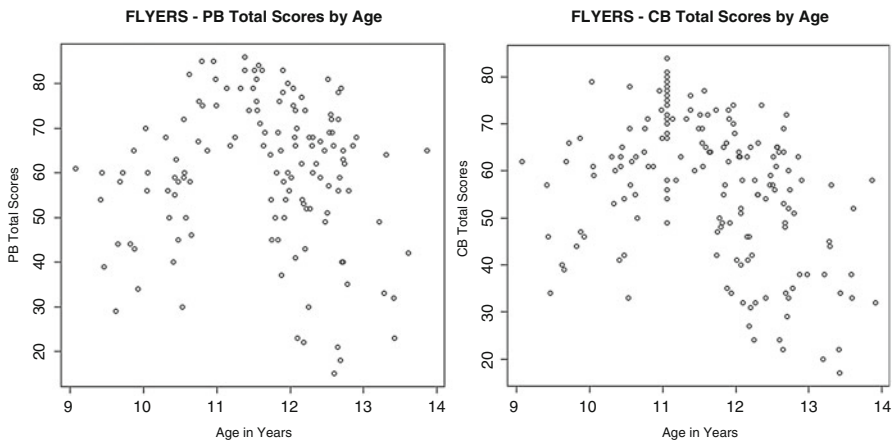


Fig. 5 PB and CB total score by age for *Flyers*

5.2.3 Gender, Years of Instruction, Computer Preference

According to Brown and McNamara (2004), the relationship between gender and test performance is not linear. Historically, in PB YLE, gender tends to affect test performance. Girls tend to achieve slightly higher than boys in terms of average score in each skill and at each level. A slightly higher standard deviation for boys indicates a wider spread of ability among boys as compared with girls in each skill at all levels. The PB/CB YLE comparability trials provided an opportunity to check for the effect of gender on candidate performance in the CB version of the tests.

First, we investigated the influence of age, gender, years of instruction and preference for delivery mode on candidate performance in the CB test for *Starters*. The variable ‘Preference for delivery mode’ describes candidate preference for delivery mode for taking an exam – either on paper, on computer or no difference. In the model we used preference on paper as the baseline for comparison for other groups. The graphs in Appendix C illustrate the effects of all regression analyses presented in this section. As Table 2 shows, in *Starters*, years of English instruction have a statistically significant effect on CB scores – the longer candidates have been receiving English instruction the better they perform in the *Starters* CB test. Table 6 shows that the effects of gender and age are not statistically significant – there seems to be no difference in the performance of male and female candidates and there are no differences in performance across age. The results show, however, that performance of *Starters* in the CB test is affected by the preference to take exams on computers rather than on paper. Candidates who prefer to take the exam on computer perform significantly better than candidates who prefer to take the exam on paper (the magnitude of the effect is 3.05). Graphical effect plots can be seen in Appendix C for all regression analyses.

To investigate which characteristics explain candidate performance in PB and CB in *Flyers* two models were tested. In the first model, the effects on candidate performance in the PB test were investigated while in the second model the performance in the CB test was investigated.

Table 6 Individual level effects on CB total test scores in *Starters*

Model 0: CB total scores				
	Estimate	Std. error	t value	Pr(> t)
(Intercept)	205.98**	67.97	3.03	0.00
Age in years	-30.24	15.86	-1.91	0.06
Age in years squared	1.53	0.91	1.69	0.09
Gender Male (baseline: Female)	0.21	2.53	0.08	0.93
Years of English instruction	1.84**	0.69	2.66	0.00
Preference for delivery mode (baseline: On paper)				
No difference	6.53	4.36	1.50	0.14
On computer	3.05**	4.24	0.72	0.47

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘.’ 1

N = 161

Multiple R-squared: 0.14, Adjusted R-squared: 0.11

F-statistic: 4.26 on 6 and 154 DF, p-value: 0.0005314

As Table 7 shows, in *Flyers*, the same variables affect candidate performance in PB and CB exams. Years of English instruction is the major factor affecting PB and CB scores – the longer candidates learn English the better they performed on the PB and CB tests. With each additional year of English instruction candidates scored 1.89 (± 0.49 standard error) points higher in the PB test (Model 1). In the CB test (Model 2) the magnitude of this effect was 1.45 (± 0.41 standard error).

We can also see that candidates who prefer to take the exam on computer scored significantly higher in both exams (PB and CB), similar to what we found in *Starters*. The effect of this variable is high – candidates who prefer to take the test on computer scored 10 points higher (± 3.3 std. error) in the PB test than candidates who prefer to take the exam on paper. Also in the CB exam candidates who prefer to take the test on computer scored 9 points higher (± 2.7 std. error) in the CB tests than candidates who prefer the exam on paper. Candidates who prefer taking the test on computer perform significantly better both in the PB and CB tests – this finding suggests a special characteristic of those candidates. However, this was not measured in the study. Even after we control for frequency of computer use and reason for computer use (as presented in Models 3 and 4 below), the effect of preference for computer use persists. We suspect that candidates who prefer taking exams on computer have a trait in common that we do not capture in this analysis as we do not have relevant variables to measure and explain it. This finding would be interesting to explore in further studies why candidates who prefer taking a test on a computer perform better than other candidates in both delivery modes.

Interestingly, boys performed significantly better than girls in the CB *Flyers* test. Boys also performed better than girls in the PB test but this difference is not statistically significant. The tendency of boys scoring higher in the CB test is also manifest at *Starters* level, even though the effect does not reach statistical significance. Again, the fact that boys perform better than girls cannot be explained by the frequency and purpose of computer use. In Models 3 and 4 the gender effect remains significant even after we controlled for both purpose and frequency of computer use.

Both models in *Flyers* account for a considerable amount of variance in candidate performance – 48 % in the PB exam and 47 % in the CB exam.

5.2.4 Frequency of Computer Use, Reason for Computer Use, Type of Computer at Home

Table 8 displays the results of Models 3 and 4 in *Flyers* where we investigated the effects on CB scores of the following individual background variables and preferences:

- (5) Frequency of computer use
- (6) Reason for computer use
- (7) Type of computer at home.

Model 3 includes individual background variables and frequency of computer usage. The results in Table 4 in Model 3 show that the frequency of computer use does not influence candidate performance on CB *Flyers*, whereas years of English

Table 7 Effects of individual background variables on PB and CB total test scores for *Flyers*

	Model 1: PB total scores					Model 2: CB total scores				
	Estimate	Std. error	t value	Pr(> t)		Estimate	Std. error	t value	Pr(> t)	
(Intercept)	-106.07	124.68	-0.85	0.40		-182.75	111.05	-1.65	0.10	
Age in years	32.00	22.97	1.39	0.17		46.78*	20.46	2.29	0.02	
Age in years squared	-1.59	1.05	-1.51	0.13		-2.34*	0.94	-2.49	0.01	
Gender (baseline: Female)										
Male	4.91	2.48	1.97	0.05		5.38**	1.93	2.79	0.01	
Years of English instruction	1.89***	0.49	3.82	0.00		1.45***	0.41	3.56	0.00	
Preference for delivery mode (baseline: On paper)										
No difference	6.66	3.47	1.92	0.06		5.07	2.88	1.76	0.08	
On computer	10.05***	3.30	3.05	0.00		9.11**	2.75	3.31	0.00	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

N=115

Adjusted R-squared: **0.48**

F-statistic: 16.11 on 7 and 106 DF, p-value: 2.8e-14

N=148

Adjusted R-squared: **0.47**

F-statistic: 17.36 on 8 and 138 DF, p-value: <2.2e-16

Table 8 Models 3–4 Effects of individual background variables and individual preferences for *Flyers*

	Model 3: CB total scores				Model 4: CB total scores			
	Estimate	Std. error	t value	Pr(> t)	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-180.67	111.85	-1.62	0.11	-76.63	117.94	-0.65	0.52
Age in years	46.48*	20.58	2.26	0.03	26.96	21.77	1.24	0.22
Age in years squared	-2.32**	0.94	-2.46	0.02	-1.43	1.00	-1.43	0.15
Gender (baseline: Female)								
Male	5.31**	1.97	2.70	0.01	5.28*	2.04	2.59	0.01
Years of English instruction	1.42***	0.41	3.42	0.00	1.33***	0.42	3.17	0.00
Preference of delivery mode (Baseline: On paper)								
No difference	5.02	2.90	1.73	0.09	4.55	2.99	1.52	0.13
On computer	9.29***	2.77	3.35	0.00	8.20***	2.91	2.82	0.01
Frequency of computer use (Baseline: Only at weekends)								
Every day	-2.28	2.82	-0.81	0.42	-2.01	2.87	-0.70	0.48
Once or twice a week	-1.05	2.71	-0.39	0.70	-1.42	2.76	-0.51	0.61
Using computers for English homework								
Using computer for playing games					-0.05	2.00	-0.03	0.98
Using computer for email/chat					0.90	2.65	0.34	0.73
Using computer for other activities					0.67	2.21	0.30	0.76
					3.00	2.25	1.34	0.18

(continued)

Table 8 (continued)

	Model 3: CB total scores				Model 4: CB total scores			
	Estimate	Std. error	t value	Pr(> t)	Estimate	Std. error	t value	Pr(> t)
Type of computer at home (Baseline: Desktop(PC/Mac))								
Desktop(PC/Mac)/laptop					-8.75	11.47	-0.76	0.45
Desktop(PC/Mac)/tablet					0.26	5.37	0.05	0.96
Desktop(PC/Mac)/tablet/laptop					2.63	4.14	0.64	0.53
Laptop					1.60	2.82	0.57	0.57
Tablet					6.00*	2.89	2.08	0.04
Tablet/Laptop					-0.23	3.67	-0.06	0.95

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

N = 148

Adjusted R-squared: **0.47**

F-statistic: 13.84 on 10 and 136 DF, p-value: <2.2e-16

N = 147

Adjusted R-squared: **0.48**

F-statistic: 7.58 on 20 and 125 DF, p-value: 1.11e-13

instruction and preference for taking exams on computer do. There is also an effect of gender – boys score significantly higher than girls on the CB *Flyers* test. We also see an age effect – the older the candidates the better they score on the CB *Flyers* test, but this effect reverses at a certain point (the curvilinear effect of age noted earlier). This model explains 47 % of variance in candidate performance.

In Model 4 we included additional individual preference variables – the reason for computer use and type of computer at home. The results show that, apart from the variables that were significant in Model 3, candidates who only have a tablet at home perform significantly better than candidates who have a PC. Model 4 explains 48 % of variance in candidate performance.

5.2.5 Type of Computer Used at the Exam

Since CB YLE is available on PC, laptop and tablet, in order to make sure that the type of computer used for the test does not affect performance in CB YLE, we investigated whether using an iPad or PC creates a difference to candidates' total score in the CB test. For this, a combined *Flyers* and *Movers* dataset was used in order to gain a considerable number of observations.

Figure 6 shows descriptive statistics of candidates who took the CB exam on an iPad and a PC. The figure shows that the median score for candidates taking the CB

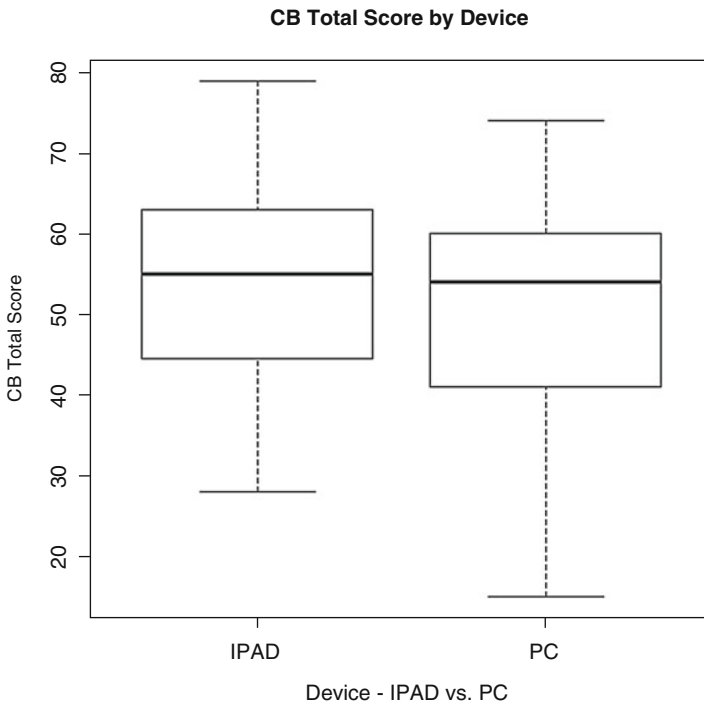


Fig. 6 CB Total score by computer device

Table 9 Effect of computer device on CB total scores for *Movers* and *Flyers*

	Model 5: CB total scores			
	Estimate	Std. error	t value	Pr(> t)
(Intercept)	-112.06	80.36	-1.39	0.17
Age at test date	30.70*	14.73	2.08	0.04
Age at test date squared	-1.44*	0.67	-2.14	0.03
Gender Male	3.66	2.0	1.83	0.07
Years of English Instruction	1.06*	0.45	2.35	0.02
Preference for delivery mode				
No difference	3.50	3.61	0.97	0.33
On computer	5.40	3.32	1.63	0.11
Frequency of computer usage				
Every day	2.29	3.21	0.71	0.48
Once or twice a week	-2.35	3.14	-0.75	0.46
Reason for computer use				
English homework	3.83	2.11	1.81	0.07
Games	0.27	2.56	0.11	0.92
Email/chat	-0.17	2.17	-0.08	0.94
Other	1.45	2.31	0.63	0.53
Type of computer at home				
PC/laptop/tablet	-0.48	2.66	-0.18	0.86
Tablet	2.43	2.26	1.08	0.28
Device used PC	-5.23	2.97	-1.76	0.08

Signif. codes: '****' 0.001 '***' 0.01 '**' 0.05

N = 203

Adjusted R-squared: **0.23**

F-statistic: 4.067 on 16 and 150 DF, p-value: 2.033e-06

test on an iPad is close to the median score for candidates that took the exam on a PC. Whether the difference in performance between those two groups is statistically significant was then tested in a regression analysis, presented below.

The computer device variable was introduced in the regression model with individual characteristics of the candidates and variables on computer usage. The dependent variable here is the total score in the CB test. As Table 9 shows, there is no statistically significant difference in CB total scores between candidates who took the test using an iPad and those using a PC when individual characteristics of candidates and their preferences for computer usage are controlled for.

6 Trial Results: Qualitative

6.1 Candidate and Parental Testimonial Feedback

Of the 322 candidates who completed the questionnaire, altogether 126 candidates and their parents from Hong Kong, Mexico and Spain gave testimonials (in their L1 or English) during the trials. In the testimonials candidates were asked three questions in their L1:

1. Did you enjoy taking the Cambridge English Young Learners test on computer? Why?
2. What did you like most about the test?
3. Would you recommend the Cambridge English Young Learners test on computer to your friends? How would you describe it to them?

On a similar form, candidates' parents were asked the following parallel questions in their L1:

1. Why did your child take the Cambridge English Young Learners test on computer?
2. What did your child like most about taking the test on computer?
3. Would you recommend the Cambridge English Young Learners test on computer to other parents? Why?

All feedback from trial candidates and parents was overwhelmingly positive, confirming the suitability of the CB delivery mode for the target candidature.

Candidate feedback indicates the CB YLE exams are very popular among young learners, as exemplified by their comments translated into English. In addition to verbal comments in questionnaires and testimonials, candidate pictures and related written comments add another perspective on their views and experiences of taking CB YLE.

These additional qualitative sources of evidence (i.e., testimonials from candidates and parents, and verbal and graphical comments from candidates) were carefully examined for common themes emerging. They were categorised by the same candidate background variables (i.e., age, gender) that were investigated by the statistical analysis from the questionnaire data. This was done in order to look for confirmation of findings or interpretation of the results, as is conventionally done in a mixed methods design. Below we exemplify some of the recurring themes emerging with typical candidate, parental and observer comments and candidate drawings.

6.1.1 CB YLE Is Innovative

Candidates and their parents especially appreciated the innovative nature of the computer-based exam delivery and the new technology involved:

“I enjoyed taking the test on computer, because it’s more interactive. I liked that the questions were oral. I would recommend it, and say: take it, it’s nice.”

(*Movers* trial candidate, boy, age 8, Mexico)

“I enjoyed taking the test on computer, because of the technology it uses and its effectiveness. I like the most that it was on an iPad. I would recommend it to my friends, as it represents a step forward for exams.”

(*Flyers* trial candidate, boy, age 12, Mexico)

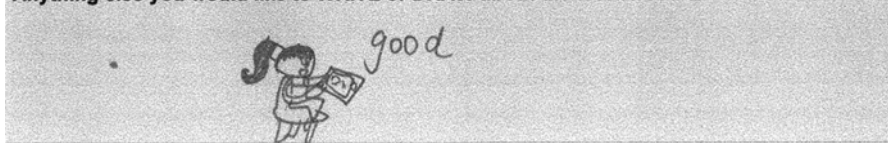
“I enjoyed it because I’ve never done an exam on a computer. I liked the speaking questions the most. I would recommend it to friends because it is very fast.”

(*Starters* trial candidate, boy, age 10, Spain)

“I think it’s an innovative method that is going to help her in the future. My child enjoyed the interaction with the computer the best. I would recommend it to other parents because children are becoming more familiar with this type of technology.”

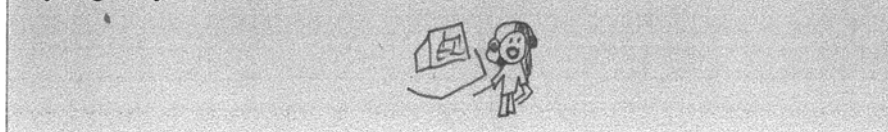
(Parent of *Movers* trial candidate, girl, age 8, Mexico)

Anything else you would like to WRITE or DRAW about the test on PAPER or COMPUTER?



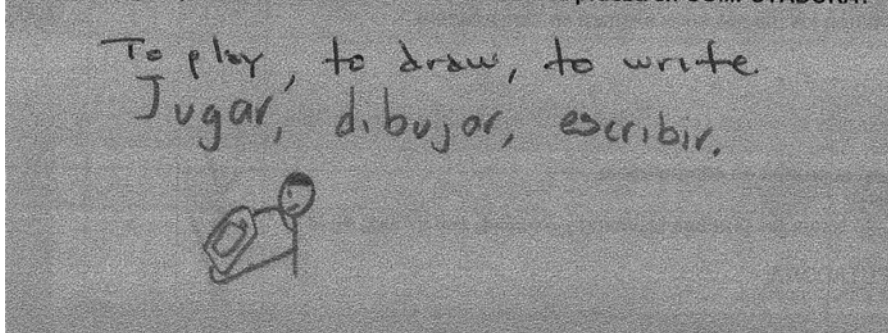
Starters trial candidate, girl, age 7, Hong Kong

Anything else you would like to WRITE or DRAW about the test on PAPER or COMPUTER?



Starters trial candidate, girl, age 8, Hong Kong

¿Algo más que quieras ESCRIBIR o DIBUJAR sobre la prueba en COMPUTADORA?



Movers trial candidate, boy, age 10, Mexico

6.1.2 CB YLE Is Fun and Motivating

Candidates and their parents thought the CB YLE tests are fun and enjoyable and game-like, and therefore have a strong motivational effect on children. Some observers' comments confirmed this:

“I enjoyed taking the test because it was funny and very entertaining. I liked the Speaking test the most.”

(*Flyers* trial candidate, girl, age 9, Mexico)

“I like it – it’s quicker and more fun. To tell you the truth I liked all of it, but if I had to choose one part it would be the speaking. I would recommend it to my friends, I would tell them: try it, it’s fun and not boring!”

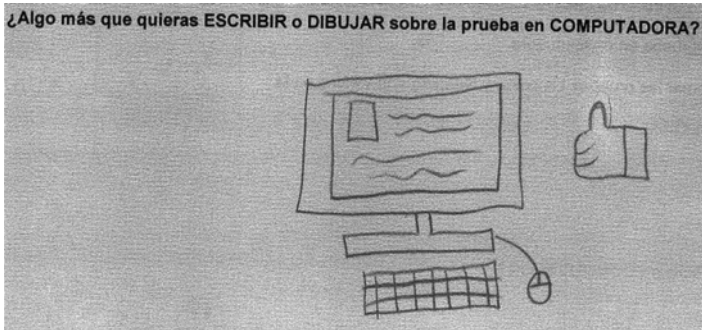
(*Movers* trial candidate, girl, age 11, Spain)

“I liked it because it was like a game and fun. I would tell my friends to do the tests because they are like games and are fun.”

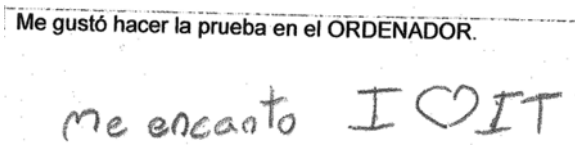
(*Starters* trial candidate, girl, age 8, Spain)

“I enjoyed taking the test on the computer – it’s very fun. I would tell my friends to do the exam because it’s fun, cool and entertaining.”

(*Starters* trial candidate, boy, age 8, Spain)



Flyers trial candidate, girl, age 12, Mexico



Starters trial candidate, boy, age 9, Spain

“My child took the test to gain more knowledge. She said it was like a game and as a mother I have seen more motivation with the computer and overall.”

(Parent of *Starters* trial candidate, girl, age 8, Spain)

“Children like computers! It’s funnier.” (observer’s comment)

“Children’s comments ranged from ‘more modern’ to ‘fun’. (observer’s comment)

6.1.3 CB YLE Tests Are at the Right Level of Difficulty

Candidates said that the level of the CB YLE tests is appropriate even though challenging for some:

“Yes I enjoyed it, because it is not easy and not too hard, it just right.”

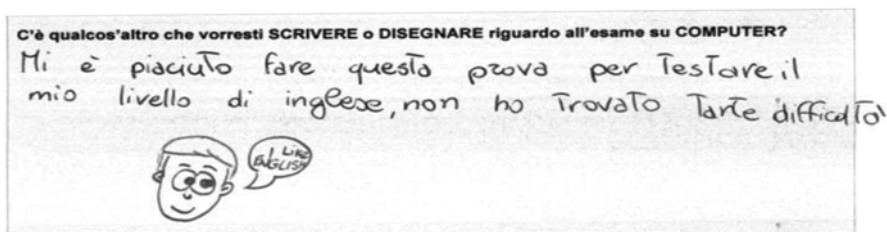
(*Flyers* trial candidate, boy, age 8, Hong Kong)

6.1.4 CB YLE Helps Students Learn English

Candidates thought that the test helps them learn English:

“I enjoyed taking the test because it was easy and fun and helped my English. I liked the Speaking test the most. I would recommend it to my friends and I would describe it to them like this: Cambridge English test is really good, it will help your English a lot.”

(*Flyers* trial candidate, boy, age 8, Hong Kong)



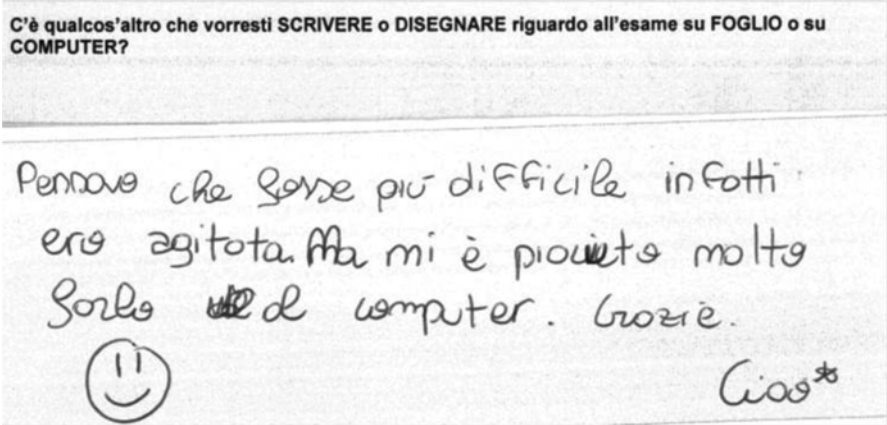
Starters trial candidate, girl, age 10, Italy: ‘I enjoyed doing this test to test my level of English, I didn’t find it that difficult’

6.1.5 CB YLE is Child-friendly

Candidates mentioned that the CB format reduces the stress conventionally associated with tests:

“I enjoyed taking the exam on the computer because you don’t get as nervous and it is more fun. The best bit was the listening exercise. I would recommend it to my friends because it’s a difficult exam that’s fun at the same time.”

(*Movers* trial candidate, girl, age 9, Spain)



Starters trial candidate, girl, age 10, Italy: 'At first, I thought it was more difficult and I was nervous. But I enjoyed it very much doing it on the computer.'

6.1.6 CB YLE Is Made by Cambridge English

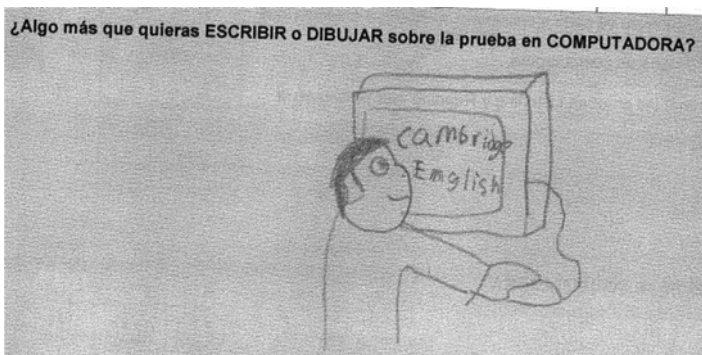
Candidates also alluded to the reputation of Cambridge English:

"I was a bit nervous during the test especially at the beginning of the oral test, but after I calmed down. During the test I was constantly thinking how much it was an honour to do this test for Cambridge."

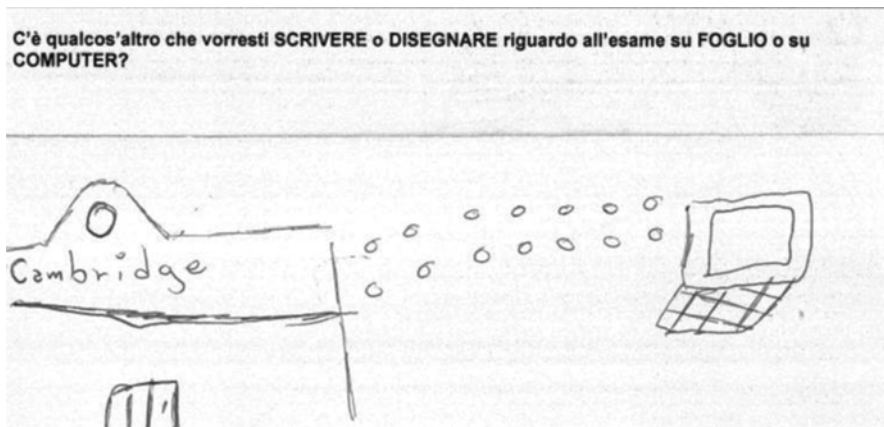
(*Starters* trial candidate, boy, age 10, Italy)

"I enjoyed the computer exam, it was like a game – it was fun. I would tell my friends to take the exam because it's from Cambridge and they study a lot for this."

(*Flyers* trial candidate, boy, age 11, Spain)



Flyers trial candidate, boy, age 12, Mexico



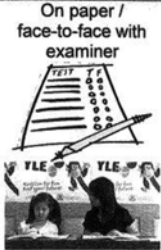

Starters trial candidate, boy, age 10, Italy

6.1.7 CB YLE Is Environment-friendly

Both children and parents recognized CB tests' beneficial effect on the environment:

“The Teacher recommended that my children try the exam. They enjoyed the test because it was easier to correct yourself if you make a mistake, and it’s more comfortable than the paper-based exam. I would recommend it then because the children enjoyed it, and I think it’s more environmentally-friendly than on paper.”

(Parent of *Starters* trial candidates, aged 10, Spain)

Your preferences about the YLE test on PAPER or COMPUTER				
		On paper / face-to-face with examiner 	No difference	On computer 
46	I found the test easier ...	3	3	3
	Why?	I bon't want cut tree		Yes

Starters trial candidate, girl, age 5, Hong Kong

6.1.8 CB YLE Helps Checking Language Learning Progress

Parental testimonials are a rich source of information to explain why parents would prefer their child to take YLE on the computer. Parents see the value of the CB test in checking their children's progress in learning English:

“Our child took the test because we would like to know his knowledge in English so that we can continue to help him in the future. Evaluating people's knowledge is the only way of guaranteeing quality in their knowledge and education.”

(Parent of *Movers* trial candidate, boy, age 9, Mexico)

“Our child took the test because it seemed a good experience and you could learn how good your child is with language. She liked the listening exercises because you can hear really well with the headphones, it's easier to concentrate.”

(Parent of *Flyers* trial candidate, girl, age 11, Spain)

6.1.9 CB YLE Helps Candidates With Special Educational Needs

One parent mentioned the educational value of the CB YLE test for her child with special needs:

“Since my son suffers from ADD, it is difficult for him to take regular exams that do not take into account the added difficulties that his attention disorder and hyperkinesis represent in terms of writing activities. I would recommend the test to other parents, because there is a wide variety of children with special needs among those taking the exam and it might be the most suitable option for many of them.”

(Parent of *Flyers* trial candidate, boy, age 12, Mexico)

The qualitative analysis revealed that younger candidates (aged 12 and under) and boys showed slightly more explicit positive attitudes towards the new CB format. Parental and candidate feedback was also confirmed by observes.

6.2 Observer Feedback

Observers also made some general comments on their checklists. Again, this source of evidence was used to inform the interpretations of the findings from the other sources of evidence in the trials.

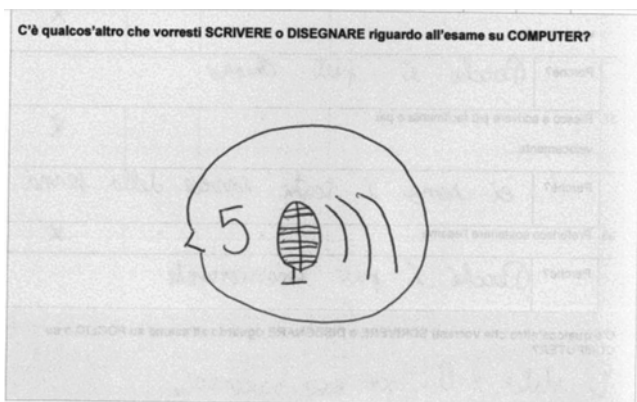
6.2.1 CB YLE Is Enjoyable

Observers confirmed that children enjoyed speaking in the CB YLE test:

“Children seemed to be quite confident in Speaking to a computer.”

“They seemed to be very happy speaking to the computer!!”

“In general, children were very comfortable with the CB YLE Speaking tests.”



Starters trial candidate, girl, age 10, Italy

6.2.2 CB YLE Is Engaging

According to observers, the very high level of engagement that children exhibited in CB YLE tests can be attributed to the following features:

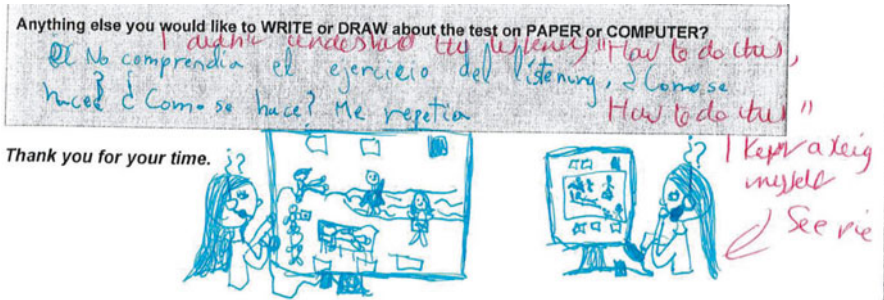
“Children seemed to be engaged and motivated by the pictures, sound and interactive activities.”

“In general computer based is more fun as the candidates enjoy using computers and it's more visual”.

6.2.3 CB YLE Is Best on Tablet

Observers confirmed that children are very capable of using computers, and they especially like using ipads/tablets. However, feedback from candidates and observers were very useful in improving the tests during the development phase:

“In general there were no problems. With some practice all the small problems that the students had could be ironed out.”

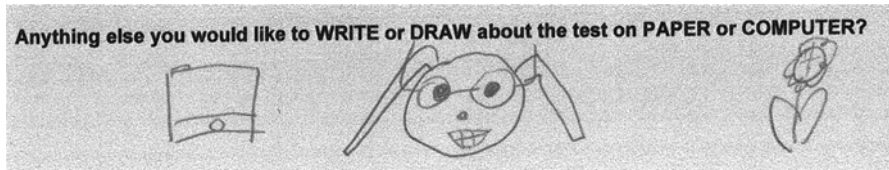


Movers trial candidate, girl, age 10, Spain

There was a clear preference for tablet delivery among candidates, which was confirmed by observer comments:

“In general, the candidates used the hardware capably and interacted well with the software. Engagement levels were high and they clearly enjoyed doing the tests.”

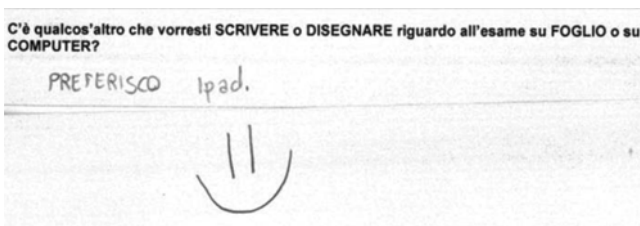
“They have no problems managing PCs and iPads at all. The candidates were happier when they were told they could do the exam with iPads.”



Starters trial candidate, girl, age 7, Hong Kong



Movers trial candidate, girl, age 10, Mexico

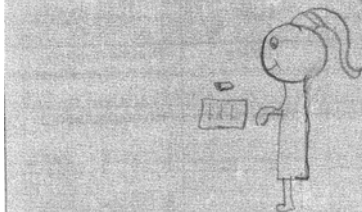


Starters trial candidate, boy, age 10, Italy

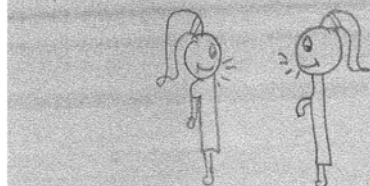
6.2.4 CB YLE May not Be Everyone's Choice

Of course, some candidates still prefer the paper-based YLE. Some candidate opinion was divided between paper and computer-based delivery as the following drawings indicate, mainly by girls:

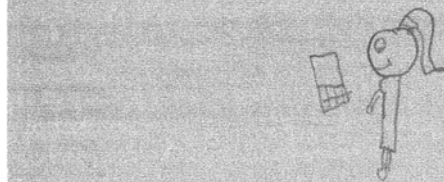
¿Algo más que quieras ESCRIBIR o DIBUJAR sobre las pruebas de Listening y Reading and Writing en papel?



¿Algo más que quieras ESCRIBIR o DIBUJAR sobre la prueba de Speaking EN VIVO con un EXAMINADOR HUMANO?



¿Algo más que quieras ESCRIBIR o DIBUJAR sobre la prueba de Speaking en ORDENADOR?



¿Algo más que quieras ESCRIBIR o DIBUJAR sobre las pruebas de Listening y Reading and Writing en ORDENADOR?



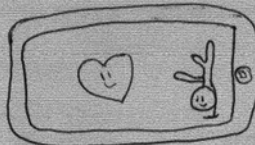
Starters trial candidate, girl, age 9, Spain

¿Algo más que quieras ESCRIBIR o DIBUJAR sobre la prueba de Speaking EN VIVO de YLE con un EXAMINADOR HUMANO?

Me cayó bien.
He/she was nice



¿Algo más que quieras ESCRIBIR o DIBUJAR sobre la prueba en COMPUTADORA?



I liked speaking to the computer.
Me gustó hablar con la computadora

¿Algo más que quieras ESCRIBIR o DIBUJAR sobre las pruebas de Listening y Reading and Writing en PAPEL?

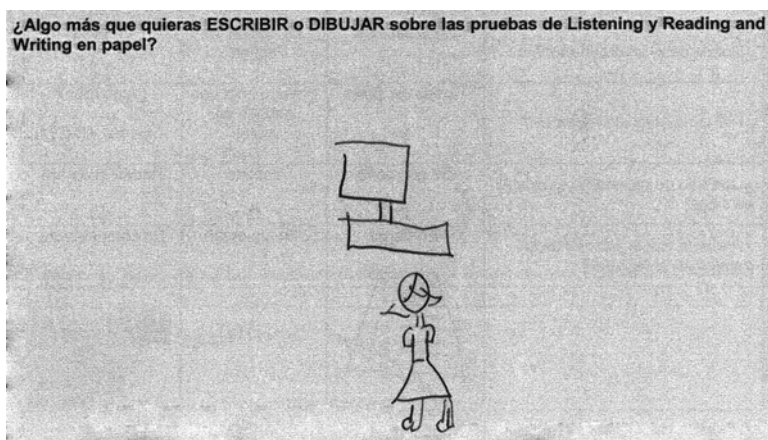
Me gustó muchísimo 😊
I liked it too much.



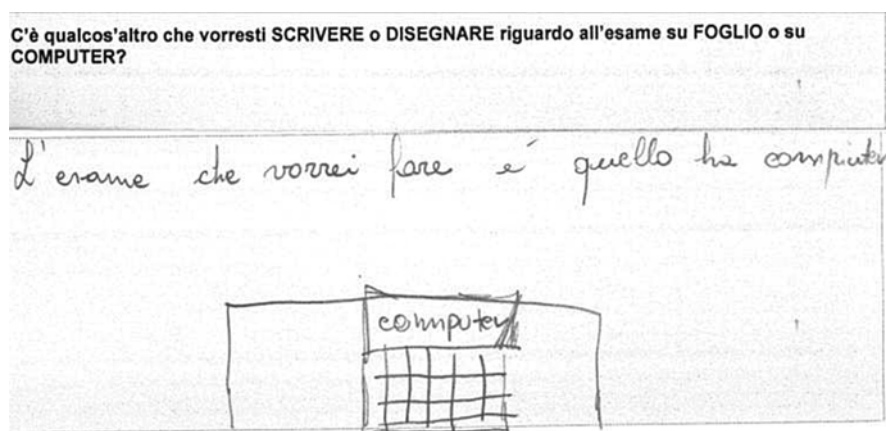
Movers trial candidate, girl, age 10, Mexico

6.2.5 CB YLE Is Popular

However, some candidates categorically preferred computers:



Starters trial candidate, girl, age 8, Spain



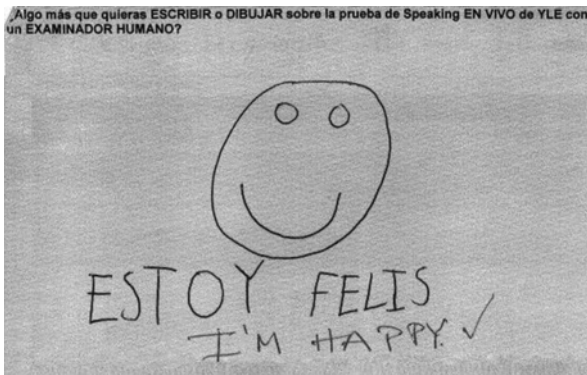
Starters trial candidate, girl, age 10, Italy "The exam I would like to do is on the computer"

6.2.6 CB YLE Trials Offered a Positive Experience

The CB YLE trials provided an overwhelmingly positive experience to children, as illustrated in the following drawings:



Movers trial candidate, girl, age 10, Mexico



Movers trial candidate, boy, age 9, Mexico

Gracias por tu tiempo.
Fue divertido



Movers trial candidate, girl, age 11, Spain

6.2.7 Conclusion

The results of the mixed methods validation study we reported on show that paper-based and computer-based versions of *Cambridge English Young Learners* are comparable alternatives and present a genuine choice for candidates to choose the exam delivery mode they feel most comfortable with. Regression analyses have shown that the number of years of English language instruction is the main factor in explaining candidates' performance both in PB and CB tests, which is in line with expectations. Candidates who prefer to take the test on computer performed significantly better both in PB and CB versions than those who prefer PB tests. This may be related to personal and motivational characteristics that this study did not explore. This result may also be related to the other interesting finding that boys were found to perform better than girls in the trials. We can speculate that perhaps this is a result of a set of personality and affective factors such as enthusiasm for computers combined with an effective use of computers and the internet to benefit from extra exposure to the English language. This interpretation was corroborated by the data collected in the testimonials as well as the verbal and pictorial feedback from questionnaires. Candidates who revealed positive attitudes to the novelty and game-like nature of the new test format tended to show stronger performance.

Importantly, during the trials, no statistically significant difference was observed in CB exam performance between candidates who took the test using an iPad and those who used a PC, confirming that which device candidates take the test on will not have an effect on their performance. However, it was very clear from the children's feedback that they prefer touch screen devices (iPads/tablets) to mouse operated devices (laptops and PCs).

In sum, overwhelmingly positive feedback from trial observers, candidates, and parents indicates that CB delivery presents a contemporary, fun, accessible and alternative way to assess children's language ability. In addition, CB YLE tests capture invaluable response and performance data for the on-going review and quality assurance of both the test material and assessment criteria employed by Cambridge English to assess children's English language ability.

The development of computer-based assessments provides young learners with an opportunity to choose the format that they prefer: PB/face-to-face or CB. Following the 'bias for best' approach that Cambridge English subscribes to, YLE candidates are allowed to choose whichever format (PB or CB) they want to take YLE tests to demonstrate the best of their language ability. The purpose of test use will determine which format is chosen: whether candidates' language skills are to be demonstrated on the computer or in a PB/face-to-face test.

7 Limitations and Future Research

In spite of the wide range of countries CB YLE was trialled in, at the time of this study data was available for analysis from only four countries. This may have an effect on the generalizability of the findings to the whole YLE population which is taken in 86 countries in the world.

In the future it would be worth exploring what causes the difference in performance between boys and girls in paper-based and computer-based language tests. Research on L2 learning in a CLIL approach also found that gender differences are cancelled out between boys and girls. It would be interesting to investigate what contributes to boys' improved attitude and motivation towards L2 learning and improved performance in these studies.

As this study only looked at self-reported computer use, further investigation could be conducted using objective measurement of children's computer use in relation to language learning. Exploratory research reported on in this volume and elsewhere in the emerging literature on young learners' English language development could be complemented by more empirical studies isolating and controlling for intervening factors to better understand the causal relationship between variables and their effect on learning outcomes.

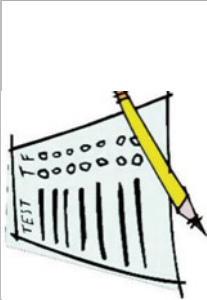

In the future, further impact studies need to be conducted to investigate reasons for choice of delivery mode (paper-based or computer based) by test takers, parents, teachers, school heads and policy makers.

Finally, this study has shown that on-going validation studies need to be carried out throughout various phases of CB test development for young learners. This 'change as usual' perspective is important in order to keep up with the changing nature of the effect of technology on learning, teaching and assessment. As Bennett (1998) has predicted, with the increasing role of technology in assessment, the boundaries between learning, teaching and assessment will ultimately be blurred, and assessment will truly be part of the teaching and learning processes, unobtrusively monitoring and guiding both (Jones, 2006).




Appendix A: Candidate Questionnaire (English Version)




CB YLE CANDIDATE QUESTIONNAIRE									
Dear CB YLE Candidate									
We would like to ask you a few questions about you and the YLE test you took. Your responses will allow us to make sure the test is working properly. Your answers will be strictly confidential. Please tick <input type="checkbox"/> the boxes or WRITE or DRAW your answers as appropriate.									
1.	What is your full name?		Surname		Given name				
2.	What is your candidate ID?								
3.	What is your mother tongue?								
4.	Are you a boy or a girl?	Boy	⊖	⊖	Girl	⊖			
5.	How old are you?	5 or younger	6	7	8	9	10	11	12 or older
		⊖	⊖	⊖	⊖	⊖	⊖	⊖	⊖
6.	Which school did you take the test at?								
7.	How many years have you been learning English?								
8.	How often do you use computers?	Every day	Once or twice a week	Only at weekends					
		⊖	⊖	⊖					
9.	Where do you use computers?	In school	In English classes	At home					
		⊖	⊖	⊖					
10.	What do you use computers for?	English homework	Email or chat with friends in English	Anything else?					
		⊖	⊖	⊖⊖				

(continued)

11.	Which type of computer do you use most at home?	Desktop (PC/Mac) ⊖	Tablet ⊖	Laptop ⊖
12.	Do you prefer taking tests on paper or on computer?	On paper ⊖	No difference ⊖	On computer ⊖
				

Listening and Reading and Writing tests on COMPUTER

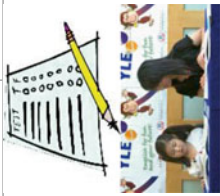

				
		Yes	Not sure	No
1.	I knew how to change the volume in the listening test.	⊖	⊖	⊖
2.	It was easy to write my answers on the computer.	⊖	⊖	⊖
3.	It was easy to select my answers on the computer.	⊖	⊖	⊖
4.	It was easy to colour my answers on the computer.	⊖	⊖	⊖
5.	It was easy to move between questions.	⊖	⊖	⊖
6.	It was easy to move between tasks.	⊖	⊖	⊖
7.	I had enough time to answer all the questions on the computer.	⊖	⊖	⊖
8.	I looked at the timer how much time I had left in the test.	⊖	⊖	⊖
9.	I was worried when I saw the timer counting down on the computer.	⊖	⊖	⊖

10.	I liked the pictures in the test on the computer.	☺	☺	☺
11.	The examples et the start of each task helped me understand what to do in the test on the computer.	☺	☺	☺
12.	I liked taking the Listening and Reading & Writing tests on the COMPUTER.	☺	☺	☺
Speaking test on COMPUTER				
				
		Yes	Not sure	No
13.	I understood how to check the microphone in the speaking test.	☺	☺	☺
14.	I knew when to <u>start speaking</u> in the speaking test on the computer.	☺	☺	☺
15.	I knew when to <u>stop speaking</u> in the speaking test on the computer.	☺	☺	☺
16.	I used the clock to help me know <u>how much time I had to speak</u> in the speaking test on the computer.	☺	☺	☺
17.	I had enough time to <u>think about my answers</u> in the speaking test on the computer.	☺	☺	☺
18.	I had enough time to <u>give my answers</u> in the speaking test on the computer.	☺	☺	☺
19.	I felt nervous taking the speaking test on the computer.	☺	☺	☺
20.	I liked speaking to a computer.	☺	☺	☺

(continued)

Anything else you would like to WRITE or DRAW about the test on the COMPUTER?

Your preferences about the YLE test on PAPER or COMPUTER

	On paper/ face-to-face with examiner	No difference	On computer
			
21. I found the test easier ... Why?	⊖	⊖	⊖
22. I prefer listening ... Why?	⊖	⊖	⊖
23. I prefer speaking ... Why?	⊖	⊖	⊖
24. I can read more easily and more quickly ... Why?	⊖	⊖	⊖
25. I can write more easily and more quickly ... Why?	⊖	⊖	⊖
26. I preferred taking the test ... Why?	⊖	⊖	⊖
Anything else you would like to WRITE or DRAW about the test on PAPER or COMPUTER?			
<i>Thank you for your time.</i>			

Appendix B: CB YLE Observer Checklist

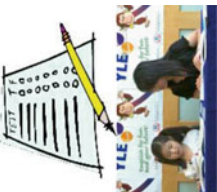

CB YLE OBSERVER CHECKLIST						
Dear YLE Examiner, Test administrator, Usher, or Teacher,						
We would like to ask you a few questions about the CB YLE tests you have observed. Your responses will allow us to make sure the test is working properly. Your answers will be strictly confidential. Please tick <input type="checkbox"/> the boxes or WRITE your answers as appropriate.						
1.	What is your name?					
2.	What is your examiner ID (if relevant)?					
3.	What is your mother tongue?					
4.	Are you male or female?	Male	Female			
5.	How old are you?	18-20	21-25	26-30	31-35	36-40
		☐	☐	☐	☐	☐
6.	In which centre/location did you observe the test?					
7.	How many years have you been examining/administering/preparing learners for YLE?	Examining	Administering	Preparing learners		
		☐	☐	☐		
	 years years years		
8.	How often do your students use computers?	Every day	Once or twice during the week	Only at weekends for homework assignments		
		☐	☐	☐		
9.	Where do your students use computers?	In school	In English classes	At home for homework		
		☐	☐	☐		
10.	What do your students use computers for?	English homework	Email or chat with friends in English	Anything else?		
		☐	☐	☐		
					
					☐	

(continued)

11.	Which type of computer do your students use most at school?	Desktop (PC/Mac)	Tablet	Laptop
		⊖	⊖	⊖
12.	Do your students prefer taking tests on paper or on computer?	On paper	Not sure	On computer
		⊖	⊖	⊖
13.	Which level of CB YLE tests have you observed?	Starters	Movers	Flyers
		⊖	⊖	⊖
14.	What type of computer were the candidates using during the test you have observed?	Desktop (PC/Mac)	Tablet	Laptop
		⊖	⊖	⊖
Your observations on the CB YLE Speaking test				
		Yes	Not sure	No
27.	The candidates checked the microphone in the speaking test.	⊖	⊖	⊖
28.	The candidates understood clearly what they had to do in the speaking test on the computer.	⊖	⊖	⊖
29.	The candidates knew when to <u>start talking</u> in the speaking test on the computer.	⊖	⊖	⊖
30.	The candidates knew when to <u>stop talking</u> in the speaking test on the computer.	⊖	⊖	⊖
31.	The animations were helpful for the candidates to know how and when to <u>start</u> talking.	⊖	⊖	⊖
32.	The animations were helpful for the candidates to know how and when to <u>finish</u> talking.	⊖	⊖	⊖
33.	The candidates checked the timer to see how much time they had to speak.	⊖	⊖	⊖
34.	I noticed some candidates rushing their answer in response to the timer.	⊖	⊖	⊖
35.	The candidates had enough time to <u>think about their answers</u> in the speaking test on the computer.	⊖	⊖	⊖
36.	The candidates had enough time to <u>give their answers</u> in the speaking test on the computer.	⊖	⊖	⊖

37.	I noticed candidates were nervous while taking the speaking test on the computer, e.g. they hesitated, looked confused or distracted.	☺	☺	☺
38.	The candidates seemed to like speaking to a computer.	☺	☺	☺
39.	Lack of human examiner support did not prevent candidates from providing responses.	☺	☺	☺
Your observations on the CB YLE Listening and Reading & Writing tests				
		Yes	Not sure	No
40.	The candidates changed the volume in the listening test.	☺	☺	☺
41.	The candidates understood what they needed to do in the Listening test on the computer.	☺	☺	☺
42.	The candidates understood what they needed to do in the Reading and Writing test on the computer.	☺	☺	☺
43.	The candidates were able to click/tap to write their answers on the computer.	☺	☺	☺
44.	The candidates were able to select their multiple choice answers on the computer.	☺	☺	☺
45.	The candidates were able to colour their answers on the computer.	☺	☺	☺
46.	The candidates were able to move easily between questions.	☺	☺	☺
47.	The candidates were able to move easily between tasks.	☺	☺	☺
48.	The candidates had enough time to answer all the questions in the Listening test on the computer.	☺	☺	☺
49.	The candidates had enough time to answer all the questions in the Reading and Writing test on the computer.	☺	☺	☺
50.	The onscreen timer in the Listening and Reading/Writing tests made candidates anxious.	☺	☺	☺
51.	The examples/model answers helped candidates answer the questions in the test on the computer.	☺	☺	☺

(continued)

	The candidates liked taking the Listening and Reading & Writing tests on the computer.	On paper / face-to-face with examiner	No difference	On computer
52.	The candidates liked taking the Listening and Reading & Writing tests on the computer.			
Candidate preferences about the YLE test on PAPER or COMPUTER				
53.	Candidates find the YLE test easier ...			
	Why?			
54.	Candidates prefer listening ...			
	Why?			
55.	Candidates prefer speaking ...			
	Why?			
56.	Candidates can read more easily and more quickly ...			
	Why?			
57.	Candidates can write more easily and more quickly ...			
	Why?			
58.	Candidates prefer taking the YLE test ...			
	Why?			
Any other comments about what you observed in the CB YLE Speaking test?				
Any other comments about what you observed in the CB YLE Listening and Reading & Writing tests?				
Any other observation about candidate preferences on PAPER-based or COMPUTER-based YLE tests?				
<i>Thank you for your time.</i>				

Appendix C: Effect Plots from Regression Analyses (Figs. 7, 8, 9, 10, and 11)

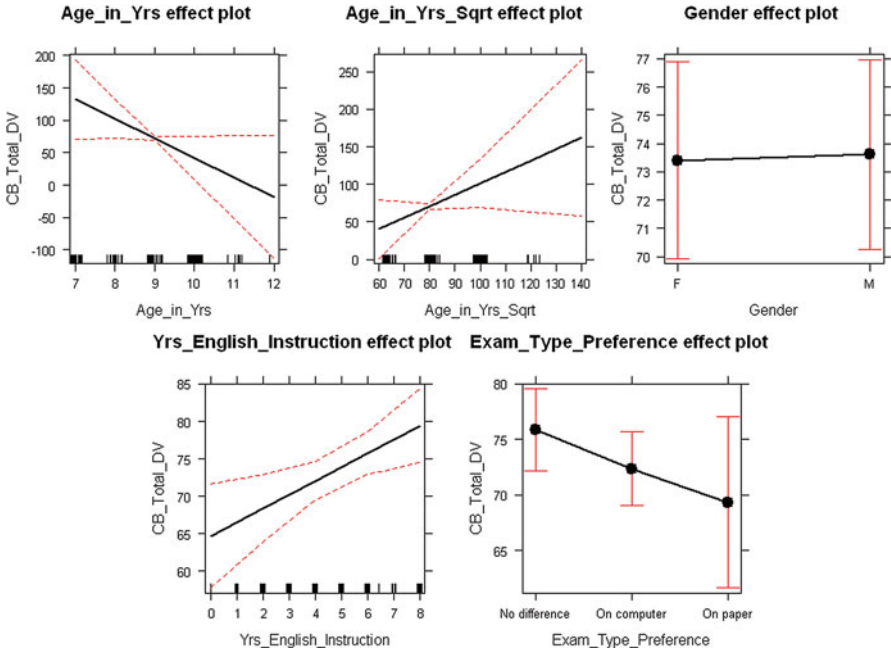


Fig. 7 STARTERS – Model 1 effect plot

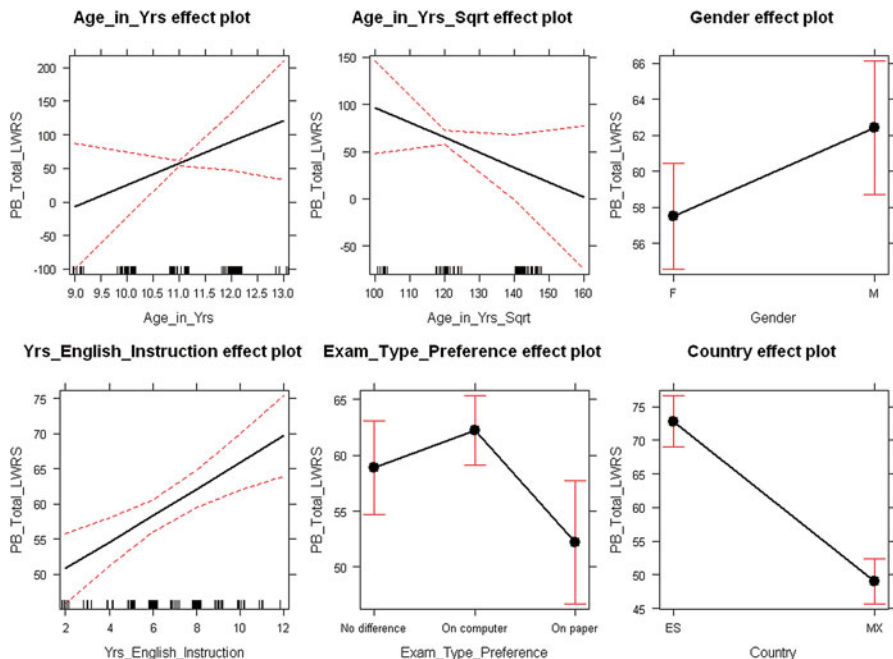


Fig. 8 FLYERS – Model 2 effect plot

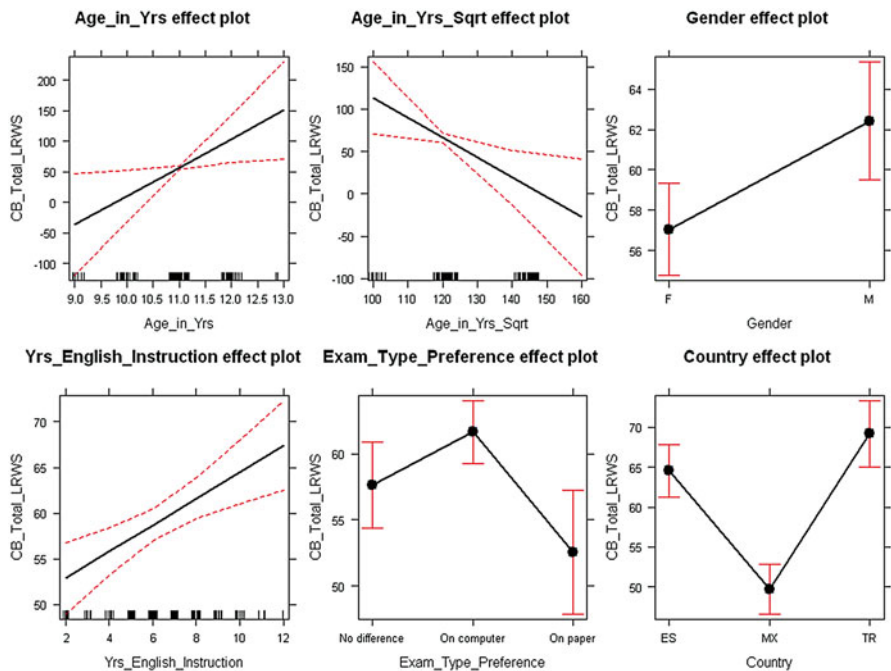


Fig. 9 FLYERS – Model 3 effect plot

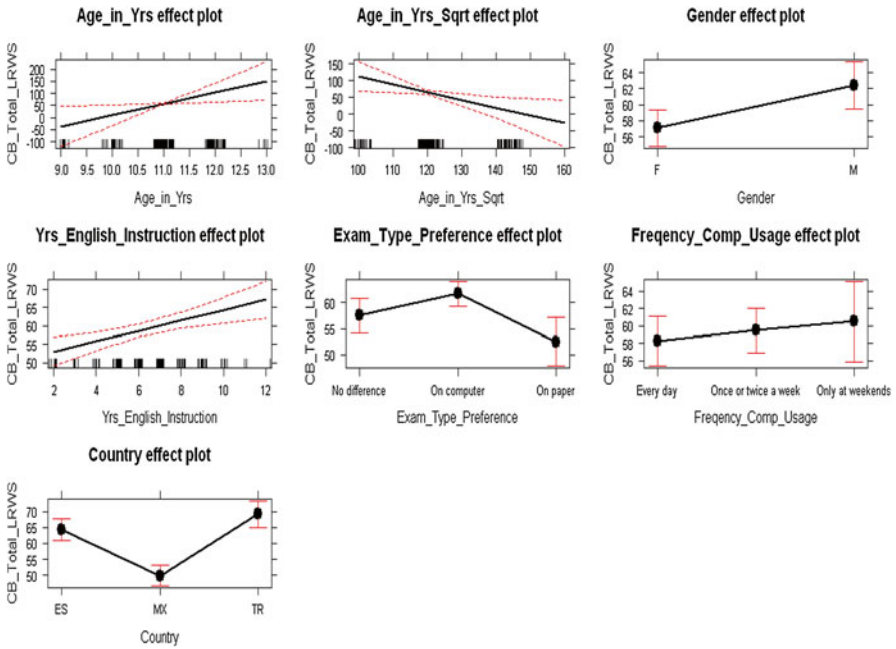


Fig. 10 FLYERS – Model 4 effect plot

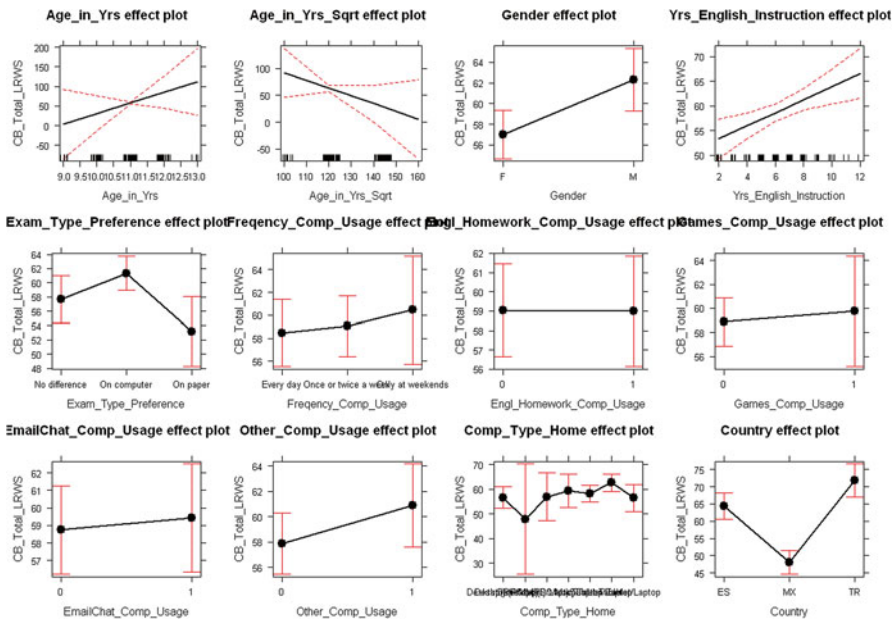


Fig. 11 FLYERS – Model 5 effect plot

References

- Barnes, S. K. (2010a). Using computer-based testing with young children. In *Proceedings of the NERA conference 2010: Paper 22*. Retrieved from http://digitalcommons.uconn.edu/nera_2010/22
- Barnes, S. K. (2010b). *Using computer-based testing with young children*. PhD dissertation, Number: AAT 3403029, ProQuest Dissertations and Theses database.
- Becker, H. J. (2000). Who's wired and who's not: Children's access to and use of computer technology. *The Future of Children: Children and Computer Technology*, 10, 3–31.
- Bennett, R. E. (1998). *Reinventing assessment: Speculations on the future of large-scale educational testing*. Princeton, NJ: Policy Information Center, Educational Testing Service.
- British Educational Research Association. (2011). *Ethical guidelines for educational research*. London: BERA. Retrieved from <http://content.yudu.com/Library/A2xnp5/Bera/resources/index.htm?referrerUrl=http://free.yudu.com/item/details/2023387/Bera>
- British Psychological Society. (2009). *Code of ethics and conduct*. Leicester, UK: BPS. Retrieved from http://www.bps.org.uk/sites/default/files/documents/code_of_ethics_and_conduct.pdf
- Brown, J. D., & McNamara, T. (2004). The devil is in the detail: Researching gender issues in language assessment. *TESOL Quarterly*, 38(3), 524–538.
- Chapelle, C., & Douglas, D. (2006). Assessing languages through computer technology. In C. J. Alderson & L. F. Bachman (Eds.), *Cambridge language assessment*. Cambridge, UK: Cambridge University Press.
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with test mode effect. *British Journal of Educational Technology*, 33(5), 593–602.
- Council of Europe. (2001). *The common European framework of reference for languages*. Cambridge, UK: Cambridge University Press.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage.
- Economic and Social Research Council. (2012). *ESRC Framework for Research Ethics (FRE) 2010*. Swindon, UK: ESRC. Retrieved from http://www.esrc.ac.uk/_images/framework-for-research-ethics-09-12_tcm8-4586.pdf
- Endres, H. (2012). A comparability study of computer-based and paper-based writing tests. *Research Notes*, 49, 26–33, Cambridge, UK: Cambridge ESOL.
- European Commission. (n.d.). *The RESPECT project code of practice*. The European Commission's Information Society Technologies (IST) Programme. Retrieved from <http://www.respectproject.org/code/index.php>
- Fox, J. (2002). *An R and S-PLUS companion to applied regression*. Thousand Oaks, CA: Sage.
- Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks, CA: Sage.
- Hackett, E. (2005). The development of a computer-based version of PET. *Research Notes*, 22, 9–13, Cambridge, UK: Cambridge ESOL.
- Hong Kong Special Administrative Region Government Education Bureau. (n.d.). *General studies for primary curriculum*. Retrieved from <http://www.edb.gov.hk/en/curriculum-development/kla/general-studies-for-primary/index.html>
- Hong Kong Special Administrative Region Education Bureau Information Services Department. (2014). *The fourth strategy on information technology in education. Realising IT potential, unleashing learning power*. Retrieved from http://www.edb.gov.hk/attachment/en/edu-system/primary-secondary/applicable-to-primary-secondary/it-inedu/Policies/4th_consultation_eng.pdf
- Jones, N. (2000). BULATS: A case study comparing computer-based and paper-and-pencil tests. *Research Notes*, 3, 10–13, Cambridge, UK: Cambridge ESOL.
- Jones, N. (2006). Assessment for learning: The challenge for an examination board. In R. Oldroyd, (Ed.), *Excellence in assessment: Assessment for learning* (pp. x–u). Cambridge, UK: Cambridge Assessment.

- Jones, N., & Maycock, L. (2007). The comparability of computer-based and paper-based tests: Goals, approaches, and a review of research. *Research Notes*, 27, 11–14, Cambridge, UK: Cambridge ESOL.
- Lee, H. K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing*, 9(1), 4–26.
- Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, 11(1), 5–21.
- Maycock, L. & Green, T. (2005). The effects on performance of computer familiarity and attitudes towards CB IELTS. *Research Notes*, 20, 3–8, Cambridge, UK: Cambridge ESOL.
- McDonald, A. S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessment. *Computers & Education*, 39(4), 299–312.
- Merrell, C., & Tymms, P. (2007). What children know and can do when they start school and how this varies between countries. *Journal of Early Childhood Research*, 5(2), 115–134.
- National Association for the Education of Young Children (NAEYC). (2009). *Joint position statement from the National Association for the Education of Young Children and the National Association of Early Childhood Specialists in State Departments of Education: Where we stand on curriculum, assessments and program evaluation*. Retrieved from <http://www.naeyc.org/files/naeyc/file/positions/StandCurrAss.pdf>
- Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement*, 22(1), 71–83.
- O'Sullivan, B., Weir, C., & Yan, J. (2004) Does the computer make a difference? *IELTS Research Project Report*. Cambridge, UK: Cambridge ESOL.
- OECD/CERI (2008). *New millennium learners: Initial findings on the effects of digital technologies on school-age learners*. Retrieved from <http://www.oecd.org/site/educeri21st/40554230.pdf> and <http://www.oecd.org/edu/ceri/centreforeducationalresearchandinnovationceri-newmillenniumlearners.htm>
- Pedró, F. (2006). *The new millennium learners: Challenging our views on ICT and learning*. Paris: OECD/CERI.
- Pedró, F. (2007). The new millennium learners: Challenging our views on digital technologies and learning. *Nordic Journal of Digital Literacy*, 2(4), 244–264.
- Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based tests. *The Journal of Technology, Learning and Assessment*, 2(6), 3–44.
- Pomplun, M., Frey, S., & Becker, D. (2000). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62, 337–353.
- Prensky, M. (2001). Digital natives, digital Immigrants. *On the Horizon*, 9(5), Bradford, UK: MCB University Press.
- Rideout, V. J., Vandewater, E. A., & Wartella, E. A. (2003). *Zero to six: Electronic media in the lives of infants, toddlers and preschoolers*. Menlo Park, CA: The Henry J. Kaiser Family Foundation.
- Russell, M., & Haney, B. (1997). Testing writing on computers: An experiment comparing students' performance on test conducted via computer and via paper-and-pencil. *Education Policy Analysis Archive*, 5(3), 1–19.
- Sim, G., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: Lessons from the literature. *ALT-J*, 12(3), 215–229.
- Sim, G., & Horton, M. (2005). Performance and attitude of children in computer based versus paper based testing. In P. Kommers & G. Richards (Eds.), *Proceedings of ED-MEDIA World conference on educational multimedia, hypermedia & telecommunications*. Seattle, WA: AACE.
- Social Research Association. (2003). *Social Research Association ethical guidelines*. London: Social Research Association. Retrieved from <http://the-sra.org.uk/wp-content/uploads/ethics03.pdf>

- Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). The relationship between computer familiarity and performance on computer based TOEFL test tasks. *TOEFL Research Reports*. Princeton, NJ: ETS.
- Tymms, P., & Merrell, C. (2009). On-entry baseline assessment across cultures. In A. Anning, J. Cullen, & M. Fleer (Eds.), *Early childhood education: Society and culture* (2nd ed., pp. 117–128). London: Sage.
- Tymms, P., Merrell, C., & Hawker, D. (2012). *IPIPS: An international study of children's first year at school*. Paris: OECD.
- Wall, K., Higgins, S., & Tiplady, L. (2009, September). *Pupil views templates: Exploring pupils' perspectives of their thinking about learning*. Paper presented at 1st International Visual Methods Conference Leeds.
- Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in grade k–12 mathematics tests. *Educational and Psychological Measurement*, 67, 219–238.
- Zandvliet, D. (1997). A comparison of computer-administered and written tests. *Journal of Research on Technology in Education*, 29(4), 423–438.

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

Eva Wilden and Raphaela Porsch

Abstract Do primary school children achieve better listening and reading skills when they start learning EFL in year 1 instead of year 3? Addressing this question this chapter sums up an empirical study investigating the EFL achievements of more than 6,500 primary school children in Germany. Data was collected in 2010 and 2012 as part of the interdisciplinary longitudinal research study *Ganz In* allowing for the comparison of two cohorts who differ in the length and quantity of early EFL instruction due to curricular changes: Whereas the 2010 cohort learned EFL for 2 lessons per week over 2 years (beginning at the age of ~8) the 2012 cohort learned EFL for two hours per week over 3.5 years (beginning at the age of ~6). In summary the findings show that children with three and a half years of early EFL education demonstrated higher receptive achievements than children with 2 years of early EFL education. Independent of their mono- or multilingual backgrounds *all* learners seemed to benefit from extending EFL education. The results of a multilevel regression analysis indicate that the language background of young learners cannot explain any variance in their receptive EFL achievements. Instead, their reading skills in German (the language of schooling) in addition to their socio-economic status and gender were identified as factors.

Keywords EFL • Foreign languages • Listening • Reading • Primary education • Empirical study • Multilingualism • Large scale study • Multilevel regression analysis

E. Wilden (✉)
English Department, University of Vechta, Vechta, Germany
e-mail: eva.wilden@uni-vechta.de

R. Porsch
Institute of Educational Science, University of Muenster, Muenster, Germany
e-mail: raphaela.porsch@uni-muenster.de

1 Introduction

Do primary school children achieve better listening and reading skills when they start learning English as a foreign language (EFL) in year 1 instead of year 3? This chapter sets out to present the design and results of an empirical study relating to the receptive EFL achievements of more than 6,500 primary school children in Germany and to find a preliminary answer to this research question. The data that were collected in 2010 and 2012 as part of the interdisciplinary longitudinal research project *Ganz In – All-Day Schools for a Brighter Future* allow us to compare two cohorts that, due to curricular changes, differ in the length and quantity of early EFL instruction. Whereas the 2010 cohort learned EFL over the course of 2 years at two lessons per week (beginning approx. at the age of 8), the 2012 cohort learned EFL over three and a half years at two hours a week (beginning approx. at the age of 6). This chapter seeks to answer a question relevant throughout Europe and beyond: whether earlier EFL education at primary level leads to better learning outcomes.

The chapter is structured as follows: After sketching out the current curricular situation with regards to early foreign language learning in Germany, the theoretical concepts underlying this study, particularly listening and reading competences as well as multilingualism, will be presented. This is followed by a summary of prior research findings on listening and reading competences in early foreign language education with a particular focus on research on young mono- and multilingual learners. In the empirical section, the research questions, the research hypotheses and the research design will be presented before the findings of the study are described and discussed.

2 Political and Theoretical Background

2.1 *The Political Background and Curricula: Germany*

In Germany education is mainly the task of the federal states (*Länder*). As a consequence, each of the 16 states has its own school system and own curriculum. However, all of the different school systems do share most of the following characteristics: In general, children enter primary education at the age of 6. In most states, children enter secondary education after year 4, in two states after year 6. It is compulsory for children to attend at least 10 years of schooling; teenagers aiming at university education attend school for 12 or 13 years in total. Most federal states begin with EFL education at the primary level in year 3, in five states children already start learning EFL in year 1 (Rixon, 2013, pp. 116–117; Treutlein, Landerl & Schöler, 2013, pp. 20–22). As the present study was conducted in the federal state of North-Rhine Westphalia (NRW), the political and curricular situation in this particular state is outlined in greater detail. Compulsory EFL education in year 3 was first introduced in NRW in the 2003/2004 school year. Just 5 years later it was

moved forward to the second term of year 1. These curricular changes caused significant transformations within a relatively short time span for both teachers and school management. Even though early EFL education was embraced by both EFL researchers and many teachers, there was a huge media controversy about these curriculum changes as exemplified in an article by Kerstan (2008) in the German broadsheet *Die Zeit* titled, “No Murks, please. Stoppt den Fremdsprachenunterricht an Grundschulen! [No screw ups, please. Stop foreign language teaching in primary schools!]”.

As a consequence of these curricular changes in NRW, the two cohorts tested in this study differ in two respects: On the one hand, they differ in the length of EFL education with the groups having two years and three and a half years respectively (approximately eighty 45-min lessons per school year). On the other hand, they were taught on the basis of two different curricula: The cohort tested in 2010 was taught on the basis of the 2003 curriculum (see MSWNRW, 2003) which first introduced primary EFL education in NRW. The second cohort tested in 2012 was the first group to be taught in accordance with the 2008 curriculum (see MSWNRW, 2008). A comparative analysis of these curricula (Wilden, Porsch & Ritter, 2013, pp. 173–176) showed that the latter curriculum prescribed a more pronounced integration of written language: teachers were asked to give written input to support EFL learning right from the start. Furthermore, the 2008 curriculum for the first time determined explicit EFL competence levels for the end of primary education in year 4 after 4 years of schooling. Both curricula highlight oral competences as one of the main objectives of early EFL education along with the acquisition of listening and audio-visual skills (also see Benigno & de Jong, 2016; Nikolov, 2016b in this volume).

2.2 Foreign Language Listening and Reading Competences in This Study

In this study, primary school children were tested on their English reading and listening skills. In this context, listening concerns the ability to extract information from spoken English. This is a complex, dynamic, active and two-sided (bottom-up and top-down) process during which learners deduce and attribute meaning and interpret what they heard (see Field, 2008; Nation & Newton, 2009; Vandergrift & Goh, 2012 for a detailed introduction to the listening construct).

The term ‘reading’ or ‘reading comprehension’ describes the ability to extract information from written English texts. This includes various simultaneous processes of understanding in the course of which readers construct meaning with the help of information given in the text (bottom-up), world knowledge gained from experience (top-down) as well as reading strategies (see Grabe & Stoller, 2011; Nation, 2008; Urquhart & Weir, 1998 for a detailed introduction to the reading construct).

2.3 Mono- and Multilingual Backgrounds of Children in the Study

A special focus of this study is on the EFL achievements of children with mono- and multilingual backgrounds in German primary schools (also see Mihaljević Djigunović, 2016 in this volume). The concept of multilingualism is used in various disciplines with different, though overlapping meanings (see Hu, 2010; Roche, 2013a, pp. 189–199). In foreign language education, multilingualism is considered to be both a prerequisite and a goal (Hu, 2004, p. 69). On the one hand, the European Commission set the political goal that every European should have communicative competences in several languages. On the other hand, active use of several languages is already part and parcel of the life of many school children in Germany even though German is the official and predominant language in Germany. This is due to the fact that there is a significant population of immigrants in Germany and according to the most recent 2012 census about 16.3 million people living in Germany (out of a total of about 80.5 million people) have a migration background (Statistisches Bundesamt Deutschland, 2013).

In line with the interdependence hypothesis (Roche, 2013b, p. 136; Rohde, 2013, p. 38) as well as the cultural dimension of multilingualism, knowledge and use of several languages outside of school should therefore be considered as a factor in further school-based language education (Hu, 2003; Roche, 2013b, pp. 193–195; Schädlich, 2013, p. 33).

We consider children to be multilingual if the following aspects apply to their lives: (a) They use German as the language of schooling and it is not their first, but their second or even third, etc. language, and (b) they either grew up with more than one language before starting their formal education or they changed to the German education system from another one to learn German as official language alongside other foreign languages on the basis of their first language (Hu, 2010, pp. 214–215). In this sense, children are categorized as being ‘multilingual’ in this study if they are growing up with more than one language in their lives outside of school and are learning English as a third or fourth language. In contrast, children are categorized as ‘monolingual’ if they are growing up with only German. The English they learn in primary EFL education is their second language.

3 Prior Research Findings on Receptive EFL Competences in Primary Education

In what follows, several empirical studies on listening and reading in early language learning of EFL will be outlined with a particular focus on studies situated in Germany. In order to limit the scope of the overview, studies relating to other aspects of early foreign language education are not considered (however see in this volume Szpotowicz & Campfield, 2016; Papp & Walczak, 2016; Mihaljević Djigunović, 2016).

3.1 Prior Research Findings on Listening Competences in Early Foreign Language Learning

The EVENING study (Paulick & Groot-Wilken, 2009) tested children in Germany (NRW) in 2006 ($N=1748$) and 2007 ($N=1344$) at the end of primary education in year 4 (age 9–10 years) on their listening and reading skills after two years of EFL learning. The tests developed in the study complied with the requirements of the relevant curriculum (MSWNRW, 2003) and even exceeded them considerably in terms of the listening test (Paulick & Groot-Wilken, p. 185). However, there were some differences between the two parts of the listening test (cf., pp. 185–187): In the first part, children heard isolated sentences and scored a mean value of 11.5 out of 17 points, which the authors of the study interpreted as being 'good' or even 'very good' results (p. 185). The second part of the test (in which children answered questions on a story they heard twice) appeared to be more challenging, for they scored a mean value of 5.5 out of 11 points. More than 73 % of the children tested were able to answer more than half of the listening items correctly and 15 % answered correctly more than three quarters of the items. The report by Paulick and Groot-Wilken does not specify whether the data analysis was based on both surveys. The absolute values in the tables on pp. 191–192 seem to indicate, however, that the results of the data analysis presented are solely based on the 2007 survey (N =approx. 1300). These results occurred in spite of the fact that the listening test was far more demanding than required in the curriculum and many teachers had assessed it as too difficult prior to its administration (p. 186).

The KESS 4 study (May, 2006) tested all primary school children in the federal state of Hamburg at the end of year 4 (ages 9–10 years) on their EFL listening achievements with a test developed for the study. The results indicated that most of the children were able to understand individual statements and answer simple questions after 2 years of EFL learning (p. 223). Twenty-five percent of the children belonged to the high-achieving group who were able to understand a coherent text read to them and connect different parts of the text with one another.

The 3-year longitudinal ELLiE study (Enever, 2011) examined among other aspects the listening skills of roughly 1,400 children in seven European countries (Germany did not take part). Beginning in the second year of EFL learning pupils aged 7–8 years were tested in listening at the end of each school year from 2007 to 2010 (The ELLiE team, 2011, pp. 15–16). By repeating four items (at the CEFR level A1; see Szpotowicz & Lindgren, 2011, p. 129) in each testing phase, the study was able to analyse the development of children's listening skills. The results showed, with only a few exceptions and country-specific variations, an improvement of children's listening achievements during the three years (pp. 130–133). The authors identified non-school related factors such as the use of the language in society or the media as factors influencing the development of foreign language listening skills (p. 133).

In a quasi-experimental study with 10 year 3 classes (age 8–9 years), Duscha (2007) researched the influence of reading and writing on various aspects of early

language learning. All ten groups were taught six parallel units, with half of the classes receiving no written language input. The pupils were tested at the end of each teaching unit. The impact of written language input on listening comprehension was tested with a picture-sentence-matching task at the end of a four-lesson unit on prepositions (after a total of 15 lessons). The children who had participated in the lessons with written language input on average scored better on the listening test than the children who had received no written language input (p. 288). These findings could be seen as an indicator that written language input in early language learning could be beneficial for the development of listening skills.

In conclusion, outcomes of these studies on the listening comprehension of primary school children (school years 1–4, aged approx. 6–10 years) can be summed up as follows (also see Bacsá & Csíkó, 2016 in this volume): The majority of children are able to understand individual, spoken sentences after two years of EFL learning and high-achieving children can even understand longer, coherent texts (May, 2006; Paulick & Groot-Wilken, 2009). In a longitudinal European comparative study, the majority of primary school children demonstrated a development of their listening skills over three years (Szpotowicz & Lindgren, 2011). Written language input in the primary EFL classroom was identified as beneficial for the development of listening comprehension (Duscha, 2007).

3.2 Prior Research Findings on Reading Competences in Early Foreign Language Learning

In recent years there has been an increase in studies on the effect of written language input in early foreign language learning in primary schools in Germany. This trend stems from the academic discourse among researchers and teachers on when the best time is to introduce written language into the early foreign language classroom (see Bleyhl, 2000, 2007; Diehr & Rymarczyk, 2010; Doyé, 2008; Treutlein et al., 2013). These studies explore both reading silently for comprehension and reading out loud for focusing on pronunciation. In line with the research focus of this study, only studies on silent reading are overviewed in this section.

On the reading test of the EVENING study, children at the end of year 4 demonstrated good reading skills after two years of EFL education – a result similar to the one found on the listening test. In the first part of the test, the young learners had to read individual sentences and match them with another sentence. On average the children scored 9.1 out of 14 points (Paulick & Groot-Wilken, 2009, pp. 188–190). In the second part of the reading test, they had to reconstruct a narrative text through a sentence-picture matching activity. On average they scored 5.6 out of 8 points. Thus, the authors of the study conceded that this part of the test appeared to be too easy for the target group (p. 189). Moreover, they stated that future studies should also go beyond the sentence level and test reading comprehension at the text level as well (p. 195). Overall, 74.2 % of the children solved more than half of the items

on the reading test and 32.5 % managed to get more than three-quarters right. The authors of the EVENING study had not expected these results (p. 195), as hardly any written language input had been presented in the 88 lessons that were evaluated in the study (Groot-Wilken, 2009, p. 137). Moreover, the teachers interviewed in the study had considered written language use to be a subordinate aspect of primary EFL teaching (p. 132).

In the ELLiE study, reading comprehension was tested with a matching activity in which the children had to fill in speech bubbles in a comic strip (Szpotowicz & Lindgren, 2011, p. 133). This task allowed for a differentiation of reading skills based on the level of difficulty of the different items. While more than 75 % of the children were able to match texts to concrete objects in a picture, only 32 % were able to correctly match a text for which they had to use contextual information and “vocabulary knowledge from the world beyond the cartoon” (p. 135).

Rymarczyk (2011) researched the EFL reading skills of year 1 and year 3 pupils and found that even underachieving learners demonstrated considerable achievements in reading provided that written language input was supplied in the EFL classroom. The author identified differences in silent reading for comprehension and reading out loud. On the one hand, the children relied on the German grapheme phoneme correspondence and thus did less well in reading out loud activities. On the other hand, they achieved much better results in silent reading comprehension activities in which they had to match pictures and words (pp. 61–65). On the basis of these results, the author argues in favour of using written language input from year 1 of primary EFL education (p. 65).

In a study examining two primary school classes who had learned English from the second semester of year 2, Frisch (2011) researched both the participants' reading comprehension and pronunciation in EFL reading. Over a period of 10 months they were taught according to two different methods. Whereas one class was taught following the *whole word approach*, the other one was taught following the *phonics method* (see Thompson & Nicholson, 1998). The study originated in the grapheme-phoneme correspondence of the English language which, compared to the more regular German grapheme-phoneme correspondence, is rather obscure (Frisch, p. 71). While the whole word approach aims at inductive-implicit reading, the phonics method explicitly deals with sound letter relationships. At the end of the project, both groups showed good test results in reading comprehension (p. 82). Moreover, the children's pronunciation appeared to have benefited from learning EFL following the phonics method (p. 84). On the basis of these findings, Frisch argued for using written language input in the early EFL classroom explicitly and systematically.

In conclusion, these empirical studies on the EFL reading comprehension of primary school children (school years 1–4) can be summed up as follows: After 2 or 3 years of early foreign language education, most children are able to understand simple sentences (Paulick & Groot-Wilken, 2009; Szpotowicz & Lindgren, 2011) as well as to reconstruct narratives with the help of pictures (Paulick & Groot-Wilken). The children demonstrate these good reading skills even if the teaching

mainly focused on fostering oral skills (Paulick & Groot-Wilken, 2009; Szpotowicz & Lindgren, 2011). From the first year of FL learning children appear to benefit in their reading comprehension from written language input and the explicit teaching of reading comprehension (Frisch, 2011; Rymarczyk, 2011).

3.3 Prior Research Finding on Receptive Foreign Language Competences of Mono- and Multilingual Children in Primary Education

The following studies explored the receptive foreign language competences of mono- and multilingual children in primary education (also see in this volume Bacsa & Csíkos, 2016; Mihaljević Djigunović, 2016): In the German EVENING study (see Sects. 3.1 and 3.2), the authors conducted a differentiated analysis by comparing the listening and reading achievements of children with different linguistic backgrounds (Paulick & Groot-Wilken, 2009, p. 190–194). They presented test results from children growing up (a) monolingually with German, (b) multilingually with one language being German and (c) multilingually without German in their families (p. 191). The findings show that monolingual children growing up with German (group a) scored slightly better on the tests than their multilingual peers (groups b and c). Children growing up in multilingual families *with* German (group b) achieved better test results than children growing up *without* German in their families (group c). The difference in the listening and reading scores between groups (a) and (b) is 1.73 and 1.26 points, between (b) and (c) 0.66 and 0.27 points and between (a) and (c) 1.99 and 1.53 points respectively. The authors did not give values for statistical significance and effect size.

A study conducted in the Swiss canton of Aargau (Husfeldt & Bader Lehmann, 2009) explored the listening and reading skills of primary school children ($N=748$) after two years of early EFL education at the end of year 4 using the instruments of the German EVENING study as well as additional background questions. The authors conclude that the children showed very good results overall and exceeded the expectations set prior to the study (p. 26). Concerning the achievements of mono- and multilingual children, they found that children growing up in monolingual families with only Swiss-German (62 %) tended to score better on both the reading and listening test (p. 16). However, Husfeldt and Bader Lehmann (p. 16) concede that other factors relating to the participants' family situation may have potentially caused this effect.

The German KESS study found that monolingual children acquired higher EFL listening competences overall than multilingual children (May, 2006), as classified by their migration background in this study. Those pupils whose neither parents were born in Germany scored on average significantly lower on the listening test than all other children (p. 213). In contrast, children with one parent who was born in Germany and the other parent abroad achieved test results similar to those of

monolingual children growing up with German as their only language spoken in their families. The study found that under certain conditions children with one parent born outside of Germany even showed higher listening scores than monolingual children. This was the case if the family spoke the language of the parent born outside of Germany and if this language happened to be a European language. The author concedes that, when interpreting the results, it must be born in mind that the linguistic situation in families often correlates with other socio-economic factors (p. 214).

Elsner (2007) examined the listening achievements of children with two years of EFL education ($N=214$) in Northern Germany at the end of primary education in year 4. She compared the EFL listening achievements of monolingual children with German as their first language to those of multilingual children with Turkish as their first language. Considering other factors included in a follow-up study (language use in families and school, motivation, motives for foreign language learning, parents' attitudes as well as learning strategies were included; pp. 181–236), Elsner found that children with Turkish as their first language scored on average significantly lower than their monolingual German speaking peers (p. 175). On the basis of her findings, Elsner disagrees with the assumption that multilingualism benefits the EFL achievements of primary school children as a matter of principle (p. 176). Furthermore, she identifies school grades in German (the language of schooling) as a relevant factor for the EFL listening achievements for both the mono- as well as the multilingual participants in her study. In this context, she highlights that children with Turkish as their first language in particular demonstrate deficits in German (p. 176), and are thus more likely to achieve lower EFL listening results.

The ELLiE study did not compare the achievements of pupils with mono- and multilingual backgrounds. However, as part of a parent questionnaire the authors collected data on background variables (out-of-school factors; Munoz & Lindgren, 2011) to research their influences on children's reading and listening achievements. They identified the professional use of a foreign language by the parents as one of the factors affecting the receptive EFL achievements of the children participating in the study (pp. 113–114). Even if this factor does not necessarily match the definition of 'multilingualism' as used in the present study, it can still be regarded as another type of out-of-school contact with another language for the children.

Based on these empirical studies on receptive EFL achievements of mono- and multilingual primary school children, the following conclusions can be drawn. The children's mono- or multilingual backgrounds are defined differently in these studies: either by the language(s) spoken in their families (Elsner, 2007; Husfeldt & Bader-Lehmann, 2009; Paulick & Groot-Wilken, 2009) or by the place of birth of their parents (May, 2006). Children growing up in multilingual families tend to show lower receptive skills after two years of English education than children growing up in monolingual families (Elsner, 2007; Husfeldt & Bader-Lehmann, 2009; Paulick & Groot-Wilken, 2009). Children growing up in multilingual families with German tend to show better receptive skills than children growing up in multilingual families without German (May, 2006; Paulick & Groot-Wilken, 2009).

Under certain conditions, children growing up in multilingual families with German achieve higher receptive skills than children growing up in monolingual families with German (May, 2006). German skills (as the language of schooling) appear to be a factor in English listening skills for both mono- and multilingual children (Elsner, 2007). Moreover, the professional use of a foreign language by one's parents was identified as a factor influencing children's receptive skills in that particular foreign language (Munoz & Lindgren, 2011).

4 Research Design

4.1 *Research Questions and Hypotheses*

The aim of this study is to determine the effect of extending the EFL learning time at German primary schools on listening and reading comprehension and to compare test results from young learners after learning EFL for two years with those who have learned EFL for three and a half years. Relating to the discourse on the pros and cons of early foreign language learning for children growing up in mono- or multilingual families, the data on receptive EFL achievements are further analysed to see whether all children benefit from the extended learning time. In other words, the study compares the test results of children with different linguistic backgrounds. The study aims at answering the following research questions:

- (1) Do learners with three and a half years of early EFL learning show higher listening and reading competences than learners with two years of early EFL learning?
- (2) Considering their mono- and multilingual backgrounds, do learners show higher degrees of listening and reading competences after three and a half years of EFL learning than after two years?
- (3) Do the mono- or multilingual backgrounds of EFL learners influence their EFL listening and reading competences at the end of primary education when statistically controlling gender, socio-economic background (SES) and German reading skills?

The following hypotheses were devised on the basis of prior research findings: Children who learned EFL for three and a half years demonstrate higher listening and reading skills than those who learned English for only two years (hypothesis 1). All children demonstrate higher receptive EFL achievements through extending the EFL learning time – independent of their linguistic backgrounds (hypothesis 2). Children growing up in multilingual families with German will demonstrate higher receptive EFL achievements than children growing up in multilingual families without German (hypothesis 3). Regarding research question 3, the existing empirical evidence is currently insufficient to devise a hypothesis.

4.2 Design and Participants

The data for this study were collected as part of the research dimension of the German *Ganz In* project. This project supports 30 secondary schools (*Gymnasien*) in NRW as they restructure their school organizations to become all-day schools. In 2010 (group 1) and 2012 (group 2) two cohorts of year 5 pupils were tested immediately after their transition from primary to secondary school (in the first 6 weeks after the summer holidays). The paper-pencil tests were administered by trained test administrators following standardized test manuals. The children in group 1 ($N_1=3216$) had learned EFL for two years, whereas those in group 2 ($N_2=3279$) had learned EFL for three and a half years. The composition of both groups was compared with regard to various background variables (nominal-scaled responses): gender, first language (German or other) and place of birth (Germany or other) in order to ensure the comparability of the two groups (see Table 1).

For the metric-scaled variables (age, number of books at home (SES) and the grades in German and English) descriptive values and results from *t*-tests are presented in Table 2.

Table 1 Pupils’ background variables (chi-squared test results)

	Group 1 (2010)	Group 2 (2012)
Gender	Girls=51.4 %	Girls=51.1 %
	Boys=48.6 %	Boys=48.9 %
	$p=.787$	
First language German	Yes=74.1 %	Yes=72.7 %
	No=25.9 %	No=27.3 %
	$p=.921$	
Country of birth Germany	Yes=96.1 %	Yes=96.2 %
	No=3.9 %	No=3.8 %
	$p=.196$	

Table 2 Pupils’ background variables (*t*-test results)

	Group 1 (2010)	Group 2 (2012)
	<i>M (SD)</i>	
Age	10.13 (.49)	10.09 (.46)
	$t(6298)=3.942; p<.001$	
Books at home (scale 1–5 from “0 to 10 books” to “more than 200 books”)	3.46 (1.15)	3.38 (1.14)
	$t(6401)=2.931; p<.001$	
Grade German (1–6)	1.87 (.59)	1.88 (.61)
	$t(6303)=.159; p=.552$	
Grade English (1–6)	1.77 (.61)	1.76 (.64)
	$t(6306)=7.476; p=.931$	

NB: *M* mean; *SD* standard deviation

Table 3 Number of pupils grouped by the languages acquired at home (in brackets percentage)

	Group 1 (2010)	Group 1 (2012)	total
(a) Monolingual with German	1647 (63.3)	1749 (66.5)	3431 (64.9)
(b) Multilingual with German	614 (23.6)	557 (21.2)	1180 (22.3)
(c) Multilingual without German	342 (13.1)	325 (12.4)	679 (12.8)

Chi-Square-Test results show that there is no statistically significant difference between the composition of groups 1 and 2 regarding the background variables considered. The *t*-test results show that there are statistically significant differences between groups 1 and 2 regarding the pupils' age and the number of books at home; however, these differences are very small and can thus be neglected. In summary, the composition of both groups appears to be comparable, thus allowing for a comparison of pupils' test results.

Furthermore, the pupils were grouped according to their linguistic backgrounds, that is, whether they are growing up in (a) monolingual families with German, (b) multilingual families with German or (c) multilingual families without German. This grouping is based on parents' responses to the question which language(s) they speak with their child (see Table 3). Apart from German, parents reported speaking the following languages with their children (numbers for both times of measurement): Polish (1.8 %/1.5 %), Russian (4.9 %/3.8 %), and Turkish (7.9 %/9 %). Although several other languages were also reported, they comprised less than 1 % of the parental group.

4.3 Measures

At both times the participants completed the same measures of EFL listening and reading comprehension as well as a socio-demographic background questionnaire. The EFL listening comprehension test that was developed for the EVENING study consisted of two tasks with a total of 28 items ($\alpha = .68$). The EFL reading comprehension test, with a total of 24 items ($\alpha = .69$), was also partially developed in the EVENING-study (Börner, Engel & Groot-Wilken, 2013; Paulick & Groot-Wilken, 2009). On both tests, the items were either multiple-choice or short answer questions designed to test the *Common European Framework of Reference for Languages* (Council of Europe, 2001) levels A1 and A2. Furthermore, proficiency scores from a reading comprehension test in German (Adam-Schwebe, Souvignier & Gold, 2009) were estimated using a Rasch model (18 items, $\alpha = .70$) as well as an index for estimating the SES in addition to the aforementioned background variables. This index is based on Bourdieu's theory (1983) and includes the pupils' and parents' responses to assess their economic, social and cultural capital, thus allowing for the allocation of pupils to four different groups (1–4) that indicate a lower or higher SES.

4.4 Data Analysis

In order to obtain proficiency scores, the students' responses were first coded as being either correct or false (dichotomous variables). Second, the data for both cohorts were scaled in one model using a probabilistic approach (Rasch model; see Rasch, 1960/1980), but for each domain (listening comprehension and reading comprehension) separately in order to get a common mean value for both groups. Analyses were computed with ConQuest 2.0 (Wu, Adams, Wilson, & Haldane, 2007) estimating person parameters (weighted likelihood estimates, WLE; Warm, 1989). The estimates are based on the scale provided by ConQuest and reach from roughly minus three to about three with a mean of zero. Following the conventions of international studies such as PISA, the scores were transformed on a scale with a mean of 500 and a standard deviation of 100. Finally, in order to estimate whether the means for the different groups are statistically significant, *t*-tests were conducted with an adjustment of the probability value (Bonferroni correction; see e.g., Mayers, 2013).

In order to answer the research questions, three multi-level analyses were conducted (*random intercept models*) instead of traditional linear regression models. Multilevel modelling accounts for the variability at different levels, as it bears in mind that the data structure is nested or hierarchical in nature (i.e., children nested within classrooms within schools). Failing to use multi-level analyses would lead to an inaccurate picture of the results, for the assumption of independent samples would be violated regarding the nested data and the standard errors of the parameters would be underestimated. All of the children tested at grade 5 were from the same school type (*Gymnasium*); however, the schools were regionally diverse (urban and rural), which influenced the composition of the cohorts (e.g., SES, the proportion of children from migrant families). All predictors were z-standardized, which has the advantage that the regression coefficients from multilevel models can be interpreted nearly as standardized regression coefficients (Bryk & Raudenbush, 1992). The analyses were conducted using the free software "R" (package: lme4).

5 Results

The results for answering research question 1 are provided in Fig. 1: On average, the children with three and half years of primary EFL education (group 2 in 2012) demonstrated higher receptive achievements than those with two years of EFL education (group 1 in 2010). On the listening comprehension test, the 2010 cohort scored a mean of 492 points and the 2012 cohort a mean of 507 points ($M=500$, $SD=100$). Similarly, on the reading comprehension test the former group scored a mean of 491 points whereas the latter scored a mean of 508 points. The 16-point difference is statistically significant for both domains.

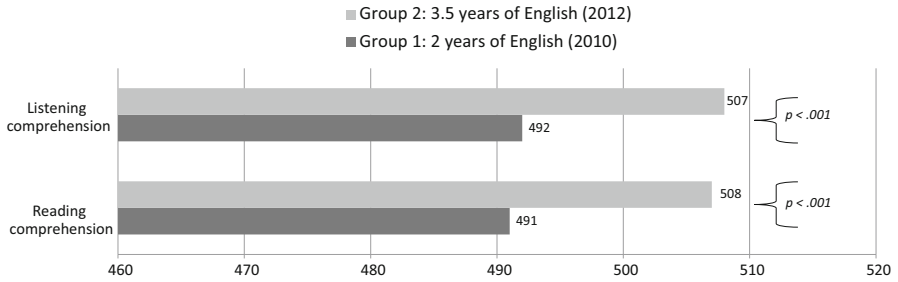


Fig. 1 Results for listening and reading comprehension after 3.5 vs. 2 years of English at primary school

Table 4 Proportion of students in four proficiency groups for listening and reading comprehension (percentage in each group)

		Less than 400	400–499	500–599	600 or more
Listening comprehension	Group 1 (2010)	10.3	46.1	35.6	7.9
	Group 2 (2012)	7.7	38.0	41.6	12.7
Reading comprehension	Group 1 (2010)	15.1	38.1	35.2	11.5
	Group 2 (2012)	12.0	33.9	39.0	15.1

In addition, the distribution of the proficiency scores were analysed in order to see whether there are differences in the distribution due to the length of the learning time. As the test developers did not provide item difficulties that would render it possible to link items to a competence model, it is not possible to interpret the children’s test results in terms of achieved competence levels. Thus, the pupils’ test results were allocated to four groups: less than 400 points, 400–499 points, 500–599 points and 600 or more points (see Table 4). The results show that the 2012 cohort had more overachieving children who scored 600 points or more on both the listening and reading comprehension tests. In contrast, there were more underachieving children in the 2010 cohort who scored 400 points or less on both tests.

Our second research question was whether all of the children, regardless of their language background, benefit from the extended learning time. Differences were made between children growing up in (a) a monolingual family with German, (b) a multilingual family with German, and (c) a multilingual family without German. The results indicate that all three groups appear to have benefited from the longer learning time in that they demonstrated higher receptive achievements (see Figs. 2 and 3). On the one hand, comparing the two multilingual groups with each other shows that, regardless of the length of EFL education, the multilingual children with German scored a little better on the reading comprehension test than the multilingual children without German. The results reversed when it comes to listening comprehension: The multilingual learners without German scored higher than those children growing up with two or more languages but without German.

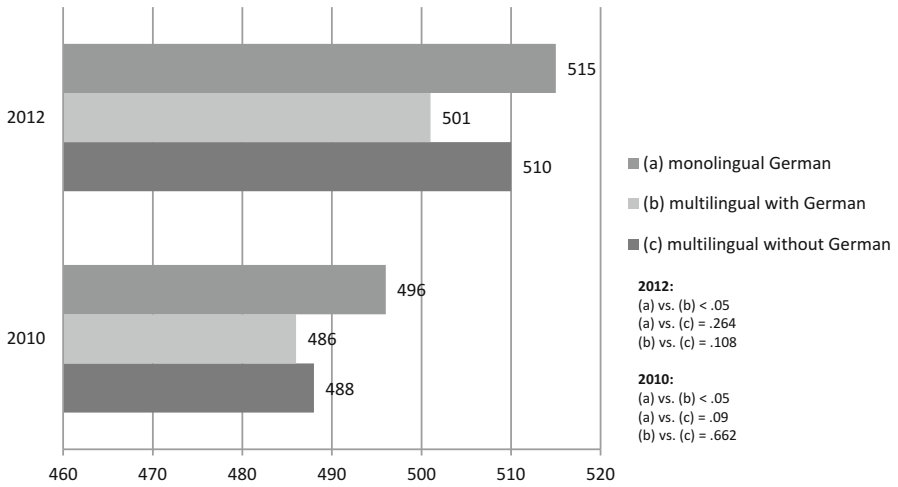


Fig. 2 Results for listening comprehension after 3.5 vs. 2 years of primary EFL education grouped according to language background (mean values and *t*-test results)

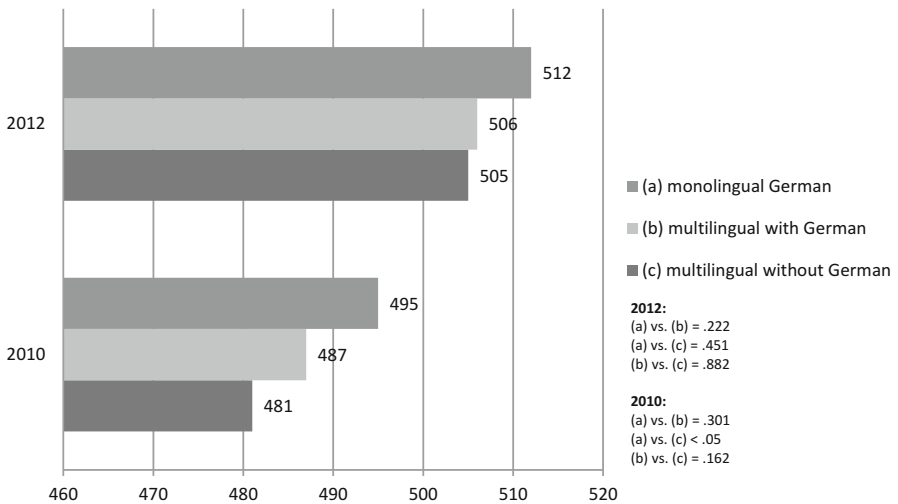


Fig. 3 Results for reading comprehension after 3.5 vs. 2 years of primary EFL education grouped according to language background (mean values and *t*-test results)

However, the differences between the two groups tested in 2010 and in 2012 are low and statistically not significant. However, regardless of the length of EFL education, the children growing up monolingually with German demonstrated the highest receptive achievements in both domains, even if only some of the differences between the groups of the monolingual and multilingual learners were statistically significant (with a maximum difference of 14 points).

Table 5 Results of the multilevel regression analysis (listening comprehension as dependent variable)

	Model 1	Model 2	Model 3
Language background	–	>.01 (.02)	–.02 (.02)
Reading comprehension German	.15* (.02)	–	.15* (.02)
SES	.17* (.02)	–	.17* (.02)
Gender	–.04* (.02)	–	–.05* (.02)
R^2	.09	–	.09

Note. * $p < .001$. In brackets: standard errors. $N = 3279$. R^2 : proportion of explained variance on the individual level (following Snijders & Bosker, 1999)

Table 6 Results of the multilevel regression analysis (reading comprehension as dependent variable)

	Model 1	Model 2	Model 3
Language background	–	.01 (.02)	.03 (.02)
Reading comprehension German	.19* (.02)	–	.20* (.02)
SES	.09* (.02)	–	.09* (.02)
Gender	–.06* (.02)	–	–.06* (.02)
R^2	.04	–	.05

Note. * $p < .001$. In brackets: standard errors. $N = 3279$. R^2 : proportion of explained variance on the individual level (following Snijders & Bosker, 1999)

The third question in this study addresses the influence of the children's language backgrounds on their receptive EFL achievements. The scores from the receptive EFL tests were taken as the dependent variables. Apart from the language background of the children (monolingual German, multilingual with or without German) the following variables were controlled: gender, SES, and reading comprehension skills in German using the proficiency score from the reading test. The analyses are based on the data from the 2012 study. First, the intraclass correlations (ICC) were calculated by applying random intercept models without any predictors (zero model). As a result, the variance proportion of the total variance that can be explained by the different schools is given. For listening comprehension the ICC is .085, meaning that only 9 % of the variance in the performance can be explained by differences across schools; for reading comprehension it was even less at only 3 %. Three models were computed: Model 1 includes the participants' reading comprehension skills in German, SES and gender as predictors for EFL listening and reading comprehension. In model 2 only the language background serves as a predictor. Finally, in model 3 all of the variables were included as predictors (see Tables 5 and 6).

Table 7 Results for German reading comprehension grouped according to pupils' language backgrounds (means and standard deviations in brackets)

	(a) Monolingual (German) ¹	(b) Multilingual with German ²	(c) Multilingual without German ³
Reading comprehension in German	513 (99.57)	492 (94.94)	477 (101.27)

Note. $N^1 = 1634$; $N^2 = 511$; $N^3 = 300$

The results show that the children's language background cannot explain any variance in their performance on the receptive EFL tests. Instead, their reading comprehension skills in German, their SES and their gender were identified as factors that explain some variance. However, the proportion of performance variance (regarding the receptive skills) that can be attributed to the individual level explained by the predictors included in the models is very small. A maximum of 9 % of the listening comprehension skills and 5 % of the reading comprehension skills are explained by model 3. Nevertheless, the findings suggest that instead of the language background of young EFL learners (whether they grow up mono- or multilingually, with or without German in their families) it is actually their German reading skills instead, in addition to their SES and gender, which impacts their receptive EFL achievements. Therefore, the data from the German reading comprehension test were also analysed to differentiate the test results according to the participants' language backgrounds (see Table 7). The proficiency scores from an IRT analysis were transformed and put onto a scale with a mean of 500 and a standard deviation of 100.

The results show considerable differences in the German reading comprehension test scores depending on the children's language background. As expected, children growing up monolingually with German achieved the highest scores. Children growing up in multilingual families with German scored 21 points less, but were still 15 points ahead of children growing up in multilingual families without German. The differences between the three groups were tested using a ANOVA model ($F[2, 2442] = 21.942$, $p < .001$). Interestingly, the large difference in their German reading comprehension skills appears to have only a small effect on their performance on the EFL tests. Comparing the mean differences in receptive EFL skills (see Tables 2 and 3), the largest difference is 14 points between the three mono- and multilingual groups. In contrast, the largest difference between the three groups on the German reading comprehension test is 36 points. The results from the multilevel analyses point to the general importance of the language proficiency in German – the language of schooling – for achievements in the EFL classroom that cannot be explained by the individual language background of these young learners. This indicates that potentially there are underlying competences which help children to understand written and oral texts across languages.

6 Conclusion and Outlook

The findings from this study can be summarized as follows. On average, the children with three and a half years of early EFL education demonstrated higher receptive achievements than children with two years of early EFL education. In the 2012 cohort, which had three and a half years of early EFL learning, there were more overachieving children who demonstrated very high receptive EFL achievements. In contrast, there were more underachieving children with rather low achievements with regards to their receptive EFL skills in the 2010 cohort, which had two years of early EFL learning. The comparison of the receptive EFL achievements of children growing up in (a) monolingual families with German, (b) multilingual families with German and (c) multilingual families without German showed that all learners seemed to benefit from extending the EFL learning time from two to three and a half years, for all three groups demonstrated higher receptive EFL skills after three and half years of EFL learning.

Furthermore, the results of a multilevel regression analysis indicated that the language background of young learners – whether they are mono- or multilingual – cannot explain any variance in their receptive EFL achievements. Instead, their reading skills in German (the language of schooling) in addition to their SES and gender were identified as factors that explain a small proportion of variance in the receptive EFL achievements of these young learners. A comparison of mono- and multilingual learners' German reading skills showed considerable differences between the three groups. While the children growing up in monolingual families with German demonstrated the highest German reading skills, the children growing up in multilingual families with German demonstrated considerably lower German reading achievements, but were still significantly ahead of the children growing up in multilingual families without German. However, the large differences in the German reading skills seemed to have only a small effect on their receptive EFL achievements, as the differences between the EFL proficiency scores of the three groups are much smaller. Nevertheless, these findings indicate a general importance of proficiency in the language of schooling for successful EFL learning on the part of young learners.

One possible explanation for this particular finding in the present study might be that children with good German competences benefit more from what teachers say in German in the EFL classroom (even though teachers should predominantly speak English). The DESI study (Helmke et al., 2007) conducted in Germany in 2003/2004 measuring among other aspects the proportion of English and German spoken in year 9 English classrooms found that 84 % of all teacher utterances were in English. However, correlations of the proportion of German/English in the classroom with students' performance in English were not reported. Unfortunately, the present study did not collect data on the language of primary EFL teacher utterances. It might be worth considering this aspect in future research studies in the field of early EFL education.

The results of the present study should be interpreted cautiously, and it would be ill-advised to hastily conclude 'The earlier, the better'. A few limitations of the study should be considered when discussing these results (Wilden et al., 2013, pp. 194–196): On the one hand, the sample is not representative in spite of its being large and standardized, for only children in one German federal state and who are attending one particular secondary school type (*Gymnasium*) in a multipartite school system were tested. Furthermore, the instruments used in the study cannot be linked to any model of competence levels. On the other hand, the curricula have also changed and there were considerable changes in EFL teacher education in NRW which coincided with the introduction of early EFL education in primary schools. These two aspects were not measured in the study; thus, it is not possible to say whether they had an impact on the findings.

Nevertheless, the findings from this study seem to indicate that – in spite of some of the arguments put forward in the German media controversy – early EFL education from year 1 seems to 'work' as all children appear to benefit from the extended learning time. However, whether the children learn 'enough' in the early EFL classroom cannot be determined on the basis of this study. In any case EFL teachers ought to be concerned with fostering their pupils' skills in the language of schooling (here: German) in order to support their foreign language competences as well. This could be done in accordance with a 'language across the curriculum' policy which many schools pursue in order to develop pupils' literacy skills in all school subjects.

Against this background, further research is planned to complement this study by (1) extending it to other secondary school types and federal states, and (2) conducting a longitudinal study on the medium and long-term developments of young EFL learners based on tasks that are linked to a competence scale.

References

- Adam-Schwebe, S., Souvignier, E., & Gold, A. (2009). Der Frankfurter Leseverständnistest (FLVT 5–6). In W. Lenhard & W. Schneider (Eds.), *Diagnose und Förderung des Leseverständnisses* (pp. 113–130). Göttingen, Germany: Hogrefe.
- 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*. New York: Springer.
- Benigno, V., & de Jong, J. (2016). The "Global Scale of English Learning Objectives for Young Learners": A CEFR-based inventory of descriptors. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.
- Bleyhl, W. (2007). Schrift im fremdsprachlichen Anfangsunterricht – ein zweischneidiges Schwert. *Take off! Zeitschrift für frühes Englischlernen*, 1, 47.
- Bleyhl, W. (2000). Empfehlungen zur Verwendung des Schriftlichen im Fremdsprachenerwerb der Grundschule. In W. Bleyhl (Ed.), *Fremdsprachen in der Grundschule: Grundlagen und Praxisbeispiele* (pp. 84–91). Hannover, Germany: Schroedel.

- Börner, O., Engel, G., & Groot-Wilken, B. (Eds.). (2013). *Hörverstehen. Leseverstehen. Sprechen: Diagnose und Förderung von sprachlichen Kompetenzen im Englischunterricht der Primarstufe*. Münster, Germany: Waxmann.
- Bourdieu, P. (1983). *Die feinen Unterschiede: Kritik der gesellschaftlichen Urteilskraft* (2nd ed.). Frankfurt am Main, Germany: Suhrkamp.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models. Applications and data analysis methods*. Newbury Park, CA: Sage Publications.
- Council of Europe. (2001). *Common European framework of reference for languages*. Strasbourg, France: Council of Europe.
- Diehr, B., & Rymarczyk, J. (Eds.). (2010). *Researching literacy in a foreign language among primary school learners/Forschung zum Schrifterwerb in der Fremdsprache bei Grundschulern*. Frankfurt am Main, Germany: Peter Lang.
- Doyé, P. (2008). Sprechen. Zuhören. Schreiben? Lesen? Gehört die Schrift in den Fremdsprachenunterricht der Grundschule? *Grundschule*, 40(3), 53.
- Duscha, M. (2007). *Der Einfluss der Schrift auf das Fremdsprachenlernen in der Grundschule. Dargestellt am Beispiel des Englischunterrichts in Niedersachsen*. Doctoral dissertation, Technische Universität Braunschweig, Germany. Retrieved from <http://www.digibib.tu-bs.de/?docid=00021088>
- Elsner, D. (2007). *Hörverstehen im Englischunterricht der Grundschule*. Frankfurt am Main, Germany: Peter Lang.
- Enever, J. (Ed.). (2011). *ELLiE. Early language learning in Europe*. London: British Council.
- Field, J. (2008). *Listening in the language classroom*. Cambridge, UK: Cambridge University Press.
- Frisch, S. (2011). Explizites und implizites Lernen beim Einsatz der englischen Schrift in der Grundschule. In M. Kötter & J. Rymarczyk (Eds.), *Fremdsprachenunterricht in der Grundschule: Forschungsergebnisse und Vorschläge zu seiner weiteren Entwicklung* (pp. 69–88). Frankfurt am Main, Germany: Peter Lang.
- Grabe, W., & Stoller, F. L. (2011). *Teaching and researching reading* (2nd ed.). New York: Routledge.
- Groot-Wilken, B. (2009). Design, Struktur und Durchführung der Evaluationsstudie EVENING in Nordrhein-Westfalen. In G. Engel, B. Groot-Wilken, & E. Thürmann (Eds.), *Englisch in der Primarstufe – Chancen und Herausforderungen: Evaluation und Erfahrungen aus der Praxis* (pp. 124–139). Berlin: Cornelsen Scriptor.
- Helmke, A., Helmke, T., Kleinbub, I., Nordheider, I., Schrader, F., & Wagner, W. (2007). Die DESI-Videostudie. *Der Fremdsprachliche Unterricht: Englisch*, 90, 37–45.
- Hu, A. (2003). *Schulischer Fremdsprachenunterricht und migrationsbedingte Mehrsprachigkeit*. Tübingen, Germany: Narr.
- Hu, A. (2004). Mehrsprachigkeit als Voraussetzung und Ziel von Fremdsprachenunterricht: Einige didaktische Implikationen. In K.-R. Bausch, F. G. Königs, & H.-J. Krumm (Eds.), *Mehrsprachigkeit im Fokus – Arbeitspapiere der 24. Frühjahrskonferenz zur Erforschung des Fremdsprachenunterrichts* (pp. 69–76). Tübingen, Germany: Narr.
- Hu, A. (2010). Mehrsprachigkeitsdidaktik. In C. Surkamp (Ed.), *Metzler Lexikon Fremdsprachendidaktik* (pp. 215–217). Stuttgart, Germany: Metzler.
- Husfeldt, V., & Bader-Lehmann, U. (2009). *Englisch an der Primarschule. Lernstandserhebung im Kanton Aargau*. Kanton Aarau, Switzerland: Department für Bildung, Kultur und Sport.
- Kerstan, T. (2008, December 17). No Murks, please. Stoppt den Fremdsprachenunterricht an Grundschulen! *Zeit Online*. Retrieved from <http://pdf.zeit.de/2008/52/C-Seitenhieb-52.pdf>
- May, P. (2006). Englisch-Hörverstehen am Ende der Grundschulzeit. In W. Bos & M. Pietsch (Eds.), *KESS 4 – Kompetenzen und Einstellungen von Schülerinnen und Schülern am Ende der Jahrgangsstufe 4 in Hamburger Grundschulen* (pp. 203–224). Münster, Germany: Waxmann.
- Mayers, A. (2013). *Introduction to statistics and SPSS in psychology*. Harlow, UK: Pearson Education Limited.

- 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*. New York: Springer.
- MSWNRW (Ministerium für Schule und Wissenschaft Nordrhein-Westfalen). (2003). *Richtlinien und Lehrpläne zur Erprobung für die Grundschule in Nordrhein-Westfalen*. Düsseldorf, Germany: MSWNRW.
- MSWNRW (Ministerium für Schule und Wissenschaft Nordrhein-Westfalen). (2008). *Richtlinien und Lehrpläne für die Grundschule in Nordrhein-Westfalen*. Düsseldorf, Germany: MSWNRW.
- Munoz, C., & Lindgren, E. (2011). Out-of-school factors – The home. In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 103–122). London: British Council.
- Nation, I. S. P. (2008). *Teaching ESL/EFL reading and writing*. New York: Routledge.
- Nation, I. S. P., & Newton, J. (2009). *Teaching ESL/EFL listening and speaking*. New York: Routledge.
- Nikolov, M. (2016). 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*. New York: Springer.
- 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*. New York: Springer.
- Paulick, C., & Groot-Wilken, B. (2009). Rezeptive Fähigkeiten und Fertigkeiten am Ende der 4. Klasse unter besonderer Berücksichtigung der sprachlichen Schülerbiografien. In G. Engel, B. Groot-Wilken, & E. Thürmann (Eds.), *Englisch in der Primarstufe – Chancen und Herausforderungen. Evaluation und Erfahrungen aus der Praxis* (pp. 179–196). Berlin: Cornelsen.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research, expanded edition (1980) with foreword and afterword by B. D. Wright. Chicago: The University of Chicago Press.
- Rixon, S. (2013). *British Council survey of policy and practice in primary English language teaching worldwide*. London: British Council.
- Roche, J. (2013a). *Mehrsprachigkeitstheorie. Erwerb – Kognition – Transkulturation – Ökologie*. Tübingen, Germany: Narr.
- Roche, J. (2013b). *Fremdsprachenerwerb – Fremdsprachendidaktik*. Tübingen, Germany: UTB.
- Rohde, A. (2013). Erst Deutsch lernen? Erst-, Zweit- und Drittsprache bei Migrantenkindern. *Grundschulmagazin Englisch*, 2013(1), 37–38.
- Rymarczyk, J. (2011). Lautes Lesen=mangelhaft/Leises Lesen=sehr gut? – Diskrepanzen in den Leseleistungen von Erst- und Drittklässlern im Fremdsprachenunterricht Englisch. In M. Kötter & J. Rymarczyk (Eds.), *Fremdsprachenunterricht in der Grundschule: Forschungsergebnisse und Vorschläge zu seiner weiteren Entwicklung* (pp. 49–67). Frankfurt am Main, Germany: Peter Lang.
- Schädlich, B. (2013). Mehrsprachigkeit und Mehrkulturalität im Unterricht der romanischen Sprachen: begriffliche, empirische und unterrichtspraktische Perspektiven. *Zeitschrift für Fremdsprachenforschung*, 24(1), 29–50.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Statistisches Bundesamt Deutschland. (2013). *Bevölkerung und Erwerbstätigkeit: Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2012 – (Fachserie 1 Reihe 2.2)*. https://www.destatis.de/DE/Publikationen/Thematisch/Bevoelkerung/MigrationIntegration/Migrationshintergrund2010220127004.pdf?__blob=publicationFile. Accessed 3 Apr 2014.
- Szpotowicz, M., & Campfield, D. E. (2016). Developing and piloting proficiency tests for polish young learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.

- Szpotowicz, M., & Lindgren, E. (2011). Language achievements: A longitudinal perspective. In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 125–142). London: British Council.
- The ELLiE Team. (2011). Introduction. In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 9–20). London: British Council.
- Thompson, G. B., & Nicholson, G. (Eds.). (1998). *Learning to read: Beyond phonics and whole language*. New York: Teachers College Press.
- Treutlein, A., Landerl, K., & Schöler, H. (2013). (Frühe) Schrifteinführung im Englischunterricht – Überlegungen zu Zeitpunkt und Methode auf Grundlage von psycholinguistischen Studien. *Zeitschrift für Fremdsprachenforschung*, 24(1), 3–27.
- Urquhart, S., & Weir, C. J. (1998). *Reading in a second language: Process, product and practice*. New York: Routledge.
- Vandergrift, L., & Goh, C. M. (2012). *Teaching and learning second language listening*. New York: Routledge.
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. doi:[10.1007/BF02294627](https://doi.org/10.1007/BF02294627).
- Wilden, E., Porsch, R., & Ritter, M. (2013). Je früher desto besser? – Frühbeginnender Englischunterricht ab Klasse 1 oder 3 und seine Auswirkungen auf das Hör- und Leseverstehen. *Zeitschrift für Fremdsprachenforschung*, 24(2), 171–201.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest 2.0: Generalised item response modelling software*. Camberwell, England: ACER.

A Longitudinal Study of a School's Assessment Project in Chongqing, China

Jing Peng and Shicheng Zheng

Abstract This case study looks at results of students who took English as a foreign language achievement tests in their Years 4–6 (ages 10–12) at Chongqing Nanping Primary School (CNPS) and analyzes them between 2010 and 2013. The students, as they used different course books, were divided into two groups: PEP English and Oxford English. The investigation of the test papers and scores of the students in the two groups has yielded the following findings: (1) As shown in the test component, in lower grades of both groups, CNPS put more emphasis on speaking and listening than comprehensive abilities; (2) For the language areas assessed, the *PEP English Test* prioritized vocabulary and grammar while the *Oxford English Test* devoted many items to assessing communicative skills; (3) Both groups had high achievers; however, students' performances showed moderate decline as the grade went higher; (4) In-depth interviews with teachers revealed that students and teachers were more motivated in the Oxford English group. The test scores also indicate that this group performed better than the PEP English group.

Keywords Achievement test • Performance • CNPS • PEP English • Oxford English

1 Introduction

In China, English has been offered from grade three (age 9) in elementary schools since 2001. *The New English Curriculum Standards* (NECS, 2001b) and *Basic Requirements of English Teaching in Elementary School* (BRETES, 2001a), which were issued by the Ministry of Education of the People's Republic of China, specify

This article was supported by grants from the Fundamental Research Funds for the Central Universities of China in the project: A study on Regional Foreign Language Education Planning. The project number is CQDXWL-2012-074.

J. Peng (✉)

Research Centre of Language, Cognition and Language Application,
Chongqing University, Chongqing, China
e-mail: pengjing@cqu.edu.cn

S. Zheng

College of Foreign Languages and Cultures, Chongqing University, Chongqing, China
e-mail: franciswindy@sina.com

that assessment of children should focus on their comprehensive abilities, including language skills, knowledge, affect, learning strategies and cultural awareness. In line with these, summative assessment, such as final exams and annual exams, should cover oral and written skills integrating the testing of the above five areas. However, after years of practice, the ideal advocated by the country met many failures (Li, 2010). To improve the efficacy of testing children, *Standard of English Curriculum for Basic Education* (revised in 2011, hereafter *Standard*) suggests that when constructing summative tests, discrete-point items should not be used to the exclusion of integrative ones, which are designed to assess different modes (receptive, productive) and different channels (written and oral) at the same time. Therefore, items need to be constructed so that both English knowledge and skills are assessed, meanwhile tending to students' affect, learning strategies and cultural awareness.

Due to growing interest in attaining documented proof of children's achievement in foreign languages, particularly in English, a nation-wide criterion-referenced test, the National English Achievement Test (NEAT) has been administered across the country since 2004. NEAT was developed by The National Basic Foreign Language Teaching Research Centre and University of Cambridge Local Examinations Syndicate. There are altogether eight levels; level one and two are designed for pupils in the present study. The test aims at assessing the English language performance of students in primary (ages 7–12) and middle school (ages 13–15) to improve learning and to inform teaching. Zhan (2007), in her research, investigated its implementation in local practices and wash-back effects on participants. The results over the years showed positive changes on the part of test-takers, approved by school teachers and policy makers. However, an empirical study by Li (2010) found that NEAT echoed traditional English tests, which still put exclusive emphasis on the absolute accuracy of language rules and forms, irrespective of gauging communicative skills. Such tests are believed to contradict global trends in English language assessments moving towards communicative language testing (e.g., Morrow, 2012; Weir, 1990).

However, NEAT is not a must, since a standardized test designed for national use cannot cater to the needs of learners in different schools and areas. In addition, such large-scale, standardized tests may be designed for administrative purposes. Shohamy (2001) points out that the role of these tests has been shifting to enable centralized bodies to control the content of education (see also Nikolov, 2016 in this volume). Pinter (2006) further questions the appropriateness of using standardized tests for young learners who are more likely to be disempowered. Therefore, many primary schools have now turned to school-based assessments, including formal and informal evaluation, in both formative and summative fashion. While findings of Gardner and Rea-Dickins (2001) suggest that it is not always appropriate to use informal tests, and formal tests may still need to be used to examine the language targets that young learners are expected to achieve, some scholars, for example, Butler and Lee (2010) argue that children have shown highly positive results when assessed in a non-traditional manner, such as classroom observations by teachers, interviews, student portfolios, self-assessments, and peer-assessments. Either way, school-based testing provides internally-referenced assessment information

(Hasselgren, 2005) to follow up students' learning throughout the years of schooling against which their progress can be measured and to improve the quality of feedback and interaction that occur between teachers and learners.

Inspired by the prospect of students taking school-based assessments, a number of primary schools have explored and incorporated several assessment tools in their curriculum, among which achievement tests are the choice for many to conduct an assessment of their students' performance. The administration usually takes place at the end of a course or an academic year to measure the amount of learning achieved by test-takers with very specific reference to a particular course (Brown, 1996). Scores of students on the tests are currently widely used in educational accountability systems, in which students' scores are deemed a reasonable measure of educational output (Koretz, 2002). According to Jacobs and Chase (1992), a well-designed achievement test not only improves and motivates students' learning, but also assists teachers in making adjustments to their teaching. Specifically, achievement tests provide teachers with feedback on how students are learning what they are taught during a certain period, whether they have made progress, and what the strengths and weaknesses are in their learning process. Teachers also use them to check the effectiveness of their teaching, as discussed by Gronlund (1993), who argues that test results can be used to evaluate the effectiveness of various aspects of the instructional process.

While the employment of an achievement test can be encouraging, like any other test, it hardly provides an accurate measurement of whatever variable is being measured. But this does not necessarily impede what teachers and other stakeholders can do with the tests, if they are well constructed to minimize inefficacy. In regard to such consideration, many scholars have studied issues like validity and reliability in test development. Fleurquin (2003), for example, explains how his team developed a standardized achievement test with thorough statistical analyses of item facility index and content validity. However, the problem is that the attention is only drawn to the development and analysis of large-scale, high-stakes tests used at the local, national, or international level (e.g., Cambridge Young Learners English (YLE) Tests, see Bailey, 2005; Pearson Test of English Young Learners (PTE), see Chik & Besser, 2011). A dearth of research has analyzed school-based tests, posing the question whether such tests used at school levels deserve to be looked at. Since young language learners are different from adult learners in characteristics defined by many researchers in the young learner literature (e.g., Cameron, 2001; Halliwell, 1992; Vale & Feunteun, 1995), their assessment also needs special attention.

The present study aims to explore how an elementary school in Chongqing, China, assessed students' English knowledge and skills through school-based achievement tests between 2010 and 2013. The students, as they used different course books, were divided into two groups: PEP English and Oxford English. By comparing two sets of test papers and scores of English learners, the chapter aims to answer the following research questions: What did the tests comprise? What skills were measured? What test and item types were used? How did the students perform on the two tests over 3 years? The whole investigation follows the research method of test item analysis; in addition to test data, interviews were also conducted for triangulation purposes.

2 Method

2.1 Participants

Participants were 498 students and seven English teachers at Chongqing Nanping Primary School (CNPS). The students were randomly divided into two groups according to different course books they would use in 2010 when they entered grade 4. The first group consisted of 304 students in six classes. The other 194 students were put in the second group comprising five classes. Table 1 offers the information on students.

In addition to the students, seven English teachers (including the vice principal, Teacher 2) were interviewed. A demographic profile of the teachers is given below in Table 2.

Four teachers from among the seven, Teacher 1, Teacher 2, Teacher 3 and Teacher 4 participated in the construction of respective English test papers taken by the students during 2010–2013. They were qualified as ‘backbone’ teachers at CNPS; this meant that they were particularly trained in the teaching and assessing of young learners.

2.2 Instruments

The instruments applied in the present study included course books students used, test scores of students and teachers’ feedback. The results of test score analysis were rationalized by examining whether or not course books exerted influence, which was then legitimated via feedback provided by teachers.

Table 1 Numbers of students in classes and the course books they used in grades 4, 5 and 6 in years 2010–2013

	PEP English group	Oxford English group
Number of students	304	194
Class numbers	Classes 1–6	Classes 7–11
Course books	<i>PEP English</i> (Gong, 2003)	<i>Oxford English</i> (Shi, 2010)

Table 2 The teachers, the grades they taught, and the textbooks they used

Interviewee	Grade	Textbook used
Teacher 1	6	Both
Teacher 2	1, 2, 3	PEP
Teacher 3	5,6	Both
Teacher 4	2,3,6	Oxford
Teacher 5	5	PEP
Teacher 6	4	Both
Teacher 7	3	Oxford

2.2.1 The Course Books

Students used different course books: the PEP group adopted *PEP English* (Gong, 2003). It is published by People's Education Press and is widely used in most public schools in China. The Oxford group used *Oxford English* (Shi, 2010), published by Shanghai Foreign Language Education Press. It was introduced from Britain and then adapted by members in Committee of Curriculum Reform in Shanghai. As deemed more difficult than *PEP English*, *Oxford English* is less frequently applied in primary schools. The purpose of using two English course books at CNPS is to examine the difference of impacts on students' learning interests and outcomes.

2.2.2 The Tests

The achievement tests had two versions based on the course books. The PEP group took the PEP English test, whereas the Oxford group took the Oxford English test. Both tests comprised an oral and a written component. Students were required to take an oral test, whereas the written test was a traditional paper-and-pencil test. For the 2010–2011 academic year, students of grade 4 were required to take the test to move to grade 5. For simplicity, in the present research, we also refer to academic year 2010–2011 as Year 4, 2011–2012 as Year 5 and 2012–2013 as Year 6.

The test papers were developed by backbone teachers in respective grades. Usually, in late December each year, the vice principal called a meeting to brief them about the requirements of test drafting. Then, after a week or so, the first draft was produced, which then went through several editing phases before administration. The final versions of test papers were administered to students at the end of each academic year (at the beginning of January). In the written test, some 30–40 students were allocated to each examination room, which was invigilated by an external teacher (not knowing the students taking tests). The written test lasted an hour. The oral test took place (before or after the written tests, depending on the testing schedule) in the teachers' offices (N=10) where an examiner (students' class teacher, who also played the role of an interlocutor in the oral dialogue) assumed responsibility for evaluating the performance of their students in pairs or threes (when the number of students was odd, but the procedure was the same). The oral test usually took 15–20 min for each group of test takers. Following the administration of tests, oral test scores were immediately reported back to the head of the English department in each grade whereas written tests were graded (cross-graded by English teachers from different grades) on the day after administration. A final report card registering the numerical grades was sent to students and their parents.

The present study employs test item analysis. As the test was administered annually during 2010–2013 to students in the PEP and Oxford groups, altogether six test papers were meticulously reviewed and analyzed in terms of the number, format, and language areas assessed. Scores of students on the tests were also computed and interpreted. Distribution charts and graphs were produced using Excel.

2.2.3 The Interview

Following the data analysis of test papers and scores, a semi-structured interview was conducted with the teachers (N=6) and the vice principal. The questions concerned their views on test paper construction and the students' performance. We devised two groups of questions (N=9), among which five were for test writers (N=4) only.

2.3 Procedures

To attain the original test papers of both groups in 3 years, a brief meeting was arranged with the vice principal on Jan 12, 2014. During the meeting, she reviewed the research proposal and agreed to be of assistance in gathering the test papers and score reports. She also appointed the head of the English department as the liaison between CNPS and the researchers.

A week later, a dozen of test papers and score reports in JPG format were emailed to the researchers, which were then printed and reorganized. After that, the test papers were thoroughly reviewed and the statistics of the types, formats and numbers of items were collected. The raw data acquired from students' tests was then entered into a spreadsheet to be analyzed.

While examining the data, problems were identified and written down. Concerning these issues, an interview outline was drafted. Next, interview questions were discussed and proposed, with nine open-ended questions established (see [Appendix 1](#)). On May 7, 2014, face-to-face interviews were conducted with all the seven teachers. Each interview lasted approximately half an hour depending on the informants' responses. The interview was carried out in Chinese so that both parties could express their ideas clearly, reducing the chance of causing any unnecessary misunderstanding. The feedback from each interviewee was written down immediately and the interviews were also recorded with the participants' consent. All data from the interviews was stored in a computer, transferred to written text and categorized according to the research questions. The written texts were then read, compared and analyzed repeatedly, and deductions were made. In the present study, some of the words are quoted (translated from Mandarin Chinese by the researchers).

3 Results

3.1 The Test Papers

3.1.1 Components

Both the PEP English test and the Oxford English test consisted of oral and written components, as displayed in [Table 3](#).

Table 3 Marks allocated to oral and written test components in PEP and Oxford tests

Grade	PEP			Oxford		
	Oral	Written		Oral	Written	
		Listening	Comprehensive skills		Listening	Comprehensive skills
4	50	20	30	50	20	30
5	40	20	40	40	20	40
6	0	30	70	0	30	70

The total mark for each test paper was 100. No difference was detected in the component make-up in the same grade across the two tests. The ratio between the oral and written tests was 50–50 % in Grade 4, 40–60 % in Grade 5, and there was no oral test in Grade 6. The written tests comprised two sections: listening and comprehensive skills. The second section took up a larger share in the written test, with 60 % in grade 4, around 66.7 % in grade 5 and 70 % in grade 6. Put into the whole test, this section also comprised a high proportion of items, especially in grade 6.

3.1.2 Item Formats

The kind of test methods or formats used can affect test performances as much as the abilities we want to measure (Brown, 1996). Thus, it is necessary to examine them to see how they function in testing the students. Some common formats were included in both the PEP and Oxford English tests items. In this part, item formats of the oral, listening and comprehensive skills sections are discussed.

3.1.2.1 Oral Section

In the oral section, the tasks ranged from reading a sentence or a passage, answering questions to doing a talent show, like singing an English song or reciting an English poem. Table 4 describes the make-up of oral section in terms of item format.

Most items in the oral section comprised reading aloud: 62.5 % in the PEP test in grade 5. The least frequently used item required speaking on the given topic.

Table 4 Marks allocated to and distribution of each item format in oral section

Formats	Grade			
	4		5	
	PEP	Oxford	PEP	Oxford
Read aloud	30 (60 %)	25 (50 %)	25 (62.5 %)	10 (25 %)
Talent show	10 (20 %)	5 (10 %)	5 (12.5 %)	15 (37.5 %)
Dialogue with the interviewer	10 (20 %)	20 (40 %)	5 (12.5 %)	5 (12.5 %)
Describe pictures				10 (25 %)
Speak on the given topic			5 (12.5 %)	
Total	50	50	40	40

In reading aloud items, students were given a few seconds to glance through an extract (of 10–15 words) or a familiar passage in the textbooks before reading it out. However, the risk is that such items are meant to assess pronunciation as distinct from free speaking. After all, the ability to read aloud does not equal the ability to converse and communicate with another person. Indeed, Heaton (1988) points out that the backwash of this kind of items may be harmful. However, according to the NECS (2001b), reading aloud is necessary for beginners for familiarizing them with the English sounds so that they can learn to read and speak English by osmosis. However, NECS does not specify if reading aloud can or should be included.

Participating in a dialogue was the second most often used item type. A close examination of these items reveals that the so-called dialogue was more of a single question-answer sequence. For instance, many items were similar like this:

Example 1 (taken from Oxford oral test, grade 5)

1. What did you have for breakfast/lunch/dinner yesterday?

Model Response: I had...yesterday.

2. What's your favorite subject?

Model Response: My favorite subject is...

3. What's the weather like today?

Model Response: It's...

The examiners would first ask the question which was to be answered by the students using words or sentences provided in Model Responses. When answering question one, students only needed to produce the names of the food to provide the information needed for scoring. After that, the conversation was terminated without any feedback from the examiner who moved on to the next question immediately. Thus, questions were unrelated and restricted both students and teachers to a drill with no real communication, except for directing students' attention to specific sentence collocations. According to Heaton (1988), these items are strictly controlled, lacking the essential element of constructive interplay with unpredictable stimuli and responses, leaving no room for authentic and genuine interaction. However, for beginners, these questions may successfully elicit vocabulary and formulaic expressions. Once they have passed this phase, the complexity of the questions can be increased and some unpredictability can be added.

The third most used format was talent show, which provided the students with a stage to showcase their language-related skills and talents. When being tested, students were required to perform solo. The time limit was 5 min, as in this example:

Example 2 (taken from PEP oral test, grade 4)

Item 4: Choose one of the favorite songs you have learned in class to perform.

As a traditional item in oral tests, singing or reciting occurred twice in PEP tests and three times in Oxford tests over the 3 years. Students came to be tested knowing what they were expected to do and prepared for it. However, when they recited texts in class in order to do well on the oral tests, they relied on their memory as well as their speaking skills. NECS (2001b) mentioned the importance of children reciting materials in English without specifying whether orally or in writing.

The second least favored format of oral test items was describing pictures. First, the students were given 1 min to study the picture in front of them. Then, they described the picture in response to the examiner's question (for instance, how many people are there in the pictures?) The description in this sense, however, was not creative in that the students were merely answering questions instead of structuring their own perceptions and putting them into words by themselves.

The least often used item was speaking on a given topic. Students were required to give a short talk on a theme they chose. They were allowed a few minutes to prepare, and in some cases, provided with textbooks for references. In the six test papers, only the PEP test in Grade 5 adopted this item format, which listed five available topics, one lifted from the textbook, the other four covering topics related to the ones in the textbook. Although these tasks are useful for stimulating and provoking students' thinking and learning, these items pose great challenges for EFL learners especially at beginning stages (McKay, 2006).

3.1.2.2 Listening Section

In the listening section, the tasks included three task types: phoneme discrimination, choose an answer to a short question, and complete a passage. Table 5 shows that the first type was the most favored format in the listening tests of both the PEP and Oxford tests, except for PEP test in Grade 6. Usually, children heard a word or sentence and had to decide which one of the three or four words or sentences printed in the answer booklet corresponded to what they heard. Hence, these items not only tested the ability to discriminate between the different sounds of a language but also the knowledge of vocabulary. However, they may appear to be of limited use, mostly for diagnostic purposes because the ability to distinguish between phonemes does not in itself imply an ability to understand verbal messages in real life. In contrast, the second type can be more suitable if we want to measure how well students can understand samples of speech by interpreting and analyzing what they have heard. As for the third type, a short written passage was provided with words omitted at regular or irregular intervals; students were asked to listen to the text and to fill in the

Table 5 Marks allocated to and distribution of each item format in listening section

Formats	Grade					
	4		5		6	
	PEP	Oxford	PEP	Oxford	PEP	Oxford
Phoneme discrimination	15 (75 %)	10 (50 %)	15 (75 %)	12 (60 %)	10 (33.3 %)	20 (66.7 %)
Choose an answer to a short question	5 (25 %)	5 (25 %)	5 (25 %)	8 (40 %)	15 (50 %)	10 (33.3 %)
Listen to complete a passage		5 (25 %)			5 (16.7 %)	
Total	20	20	20	20	30	30

missing words. Also referred to as “aural cloze” items, they focus more on students’ ability to detect sounds of the words being used (McKay, 2006). In fact, students who do not possess appropriate literacy levels to understand the whole passage can write the words down as they hear them, which resembles what they do in a dictation.

3.1.2.3 Comprehensive Skills Section

Some common item formats were found in the comprehensive skills section in both the PEP and Oxford English tests: multiple-choice, true-false, matching, fill-in blanks, short answer and essay. Table 6 demonstrates the difference of the weighting of each item format.

We can see that from grade four to six, the most frequently used item format in both the PEP and the Oxford tests was multiple choice, followed by true-false, and matching. Multiple choice items accounted for at least 35.7 % among all the test items. Its number even added up to half of the items in grade 4. However, Kohn (2000) claims multiple choice items are the “most damaging” type which limits assessment to raw data and neglects the most important features of learning, such as initiative, creativity, curiosity, and imagination. Despite the fact that these items run the risks of assessing recall of knowledge as well as guessing, they are an indispensable part in the achievement tests, and if well-designed, they can be applied to challenge students’ higher level of thinking (Berry, 2008).

The essay items pushed the task beyond discrete-point tests that measured small bits and pieces of a language to challenge their higher-level cognitive skills (Brown, 1996). According to NECS (2001b), an appropriate proportion of essay items can be introduced; however, as for the measurement of this proportion, no yardstick is offered. It was found that the least favored item format (especially in the PEP test) was essay. This might result from the discussion that writing should be age-inappropriate for young EFL learners, since it exerts far more cognitive demands on children than they can process (Weigle, 2002).

Table 6 Weighting of item formats in comprehensive skills sections

Formats	Grade					
	4		5		6	
	PEP	Oxford	PEP	Oxford	PEP	Oxford
Multiple choice	15 (50 %)	15 (50 %)	20 (50 %)	15 (37.5 %)	30 (42.9 %)	25 (35.7 %)
True-false	5 (16.7 %)	5 (16.7 %)	5 (12.5 %)	5 (12.5 %)	10 (14.3 %)	10 (14.3 %)
Matching	5 (16.7 %)	5 (16.7 %)	5 (12.5 %)	5 (12.5 %)	10 (14.3 %)	10 (14.3 %)
Fill-in the blanks			5 (12.5 %)	10 (25 %)	10 (14.3 %)	10 (14.3 %)
Sentence ordering	5 (16.7 %)				5 (7.1 %)	10 (14.3 %)
Essay		5 (16.7 %)	5 (12.5 %)	5 (12.5 %)	5 (7.1 %)	5 (7.1 %)
Total	30	30	40	40	70	70

Thus, merely judging from the number of item formats, it is not possible to decide if they are appropriate for the testees without evaluating what is being tested. In fact, all of the above item formats have been applied widely in the tests of young EFL learners and they have been proved useful (e.g., Hasselgren, 2005; McKay, 2006). Then, it is imperative to look at what the expectations are and what knowledge and skills they should possess to perform well on the tests. To answer the question, we have to study what areas of English language are assessed in the tests. This is the focus of the next section.

3.1.3 Language Areas Tested

According to the *Standard* (revised, 2011), language knowledge covered in the teaching of young EFL learners includes phonology, vocabulary, grammar, and function-notion. The *Standard* specifies requirements of EFL learners in a way that systematically integrate knowledge and skills. In order to analyze what was assessed in the two tests, we referred to the *Standard* and categorized the test items into the above four areas. However, they did not necessarily indicate a clear-cut separation from one language area to another. For instance, in listening comprehension, some items assessed both phonology and vocabulary. In this case, we consulted the item writers about the focal language area the item attempted to assess, so that we could subsume the item to the most suitable category. The examples are taken from the PEP test of Grade 5.

Phonology (items concerned with pronunciation and intonation)

Decide whether the underlined part sounds the same:

play say

Vocabulary (items concerned with word meanings, word formation and collocations)

Decide which word does not belong to the word category

A: winter

B: cool

C: spring

D: summer

Grammar (items concerned with appropriate grammatical forms and rules)

Multiple choice

I' m _____ a letter.

A: write

B: writing

C: writeing

Function-notion (items concerned with appropriate use of language for different purposes in various contexts, e.g., introduce oneself, express apology).

When something terrible happens to your friend, what would you say to him/her?

A: Not at all.

B: I'm sorry to hear that.

C: You're welcome.

A thorough review of the test papers and the items yielded Fig. 1, which shows the difference between marks the PEP and the Oxford tests allocated to items assessing different language areas.

Over the three grades, the four types of language areas assessed by one or more items varied in both the PEP and the Oxford tests. Phonology items were focused in both tests in the first 2 years, whereas vocabulary became highlighted in Grade 6. Grammar also secured its place in the test paper for both groups, with PEP taking up a higher proportion. As for function-notion items, the Oxford tests devoted more items to assessing language use than the PEP tests in all three grades.

The last example (provided above) represented one of these items where students were required to choose the most appropriate answer in a context. The item went beyond language knowledge to assess students' communicative ability. In this context, students needed to understand how to report attitudes properly to the speaker who was in trouble. All the three options were grammatically acceptable but only one of them was appropriate in the context where the dialogue took place. The appropriate response could only be chosen if students understood how to perform the expressive function and to express regret in western culture. Even if they have mastered a number of language elements (the meaning of each option, for example), it is likely that they chose a wrong answer. An item like this offered the students authentic language, though more demanding than retrieval or rote memorization of factual information, and provided them with an opportunity to use the language. Such item is acclaimed by Heaton (1988, p. 10) as "the best test of mastery of a language". Hymes (1972) also points out that learners not only have to use qualified sentences according to the grammar rules, but they should also have the ability to use them in different contexts. Therefore, in an English test paper, it is necessary to develop items with authentic materials in authentic contexts to serve a purpose, which Ao (2002, p. 31) described as "observing if the learners have the competence of using language to achieve the aims of communication."

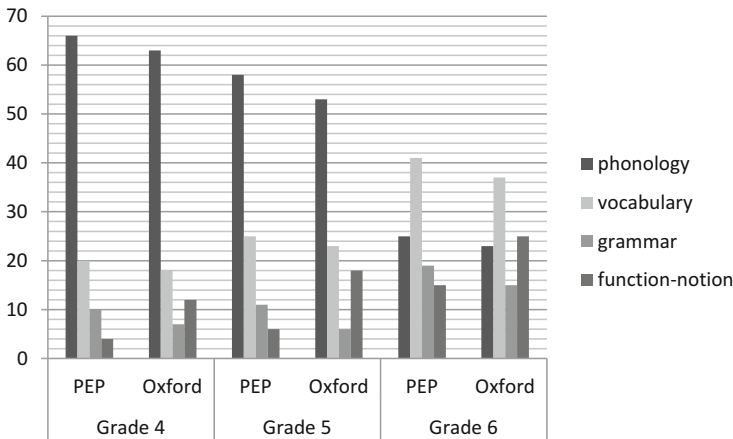


Fig. 1 Marks allocated to items assessing different language areas

3.2 Students' Performance

The students' performance on the tests was described by their scores. Before we discuss comparisons of the two test papers, it is necessary to take a look at the level of difficulty.

3.2.1 Difficulty of Test Papers

The data we collected allowed us to estimate the mean level of difficulty (P) using the formula $P=M/T$ (Yao & Duan, 2004). M represents the mean score of the students while T means the total score of the test paper (100 marks). The higher the value, the easier the test paper is. The value of P ranges from 0 to 1. The M and P values of 3 years on both the PEP and the Oxford tests are given in Table 7. The mean level of difficulty for both tests was relatively low; it increased over the years. Although two different test papers were used, the level of difficulty was comparatively close, with the PEP test paper showing a slightly (almost negligible) higher P than that of the Oxford test. The highest level of difficulty was found for the PEP test paper of Year 6, and the P value reached 0.82.

3.2.2 Score Distribution in Different Bands

As for the score distribution, four bands are applied to see how students performed on the test, which is 90–100; 80–89; 60–79; below 60. Teachers at CNPS generally viewed students who scored in the first band as outstanding performers, those in the second band were considered good performers, in the third band poor performers and students in the last band failed to achieve the required level.

The vertical axis in Fig. 2 shows the number of students who score in each band. In both the PEP and the Oxford groups, while the outstanding performers comprised the largest ratio throughout 3 years, their number declined over the years. As for good performers, both groups showed a steady growth of students, but the PEP group outnumbered the Oxford group. Poor performers could be observed throughout 3 years, with the lowest number appearing in the Oxford Group in year 4, when only ten students were counted. For students scoring below 60 (failed), the number increased gently every year. In year 4, no students failed in any of the groups whereas at the end of primary school education (Year 6), 25 students (8.2 %) in the PEP group failed; this number constituted the largest ratio. As for the Oxford English group, seven students (3 %) failed.

Table 7 M and P values (accurate to the second decimal place)

	Year 4		Year 5		Year 6	
	M	P	M	P	M	P
PEP	91.49	0.91	86.13	0.86	82.58	0.82
Oxford	94.43	0.94	87.21	0.87	83.01	0.83

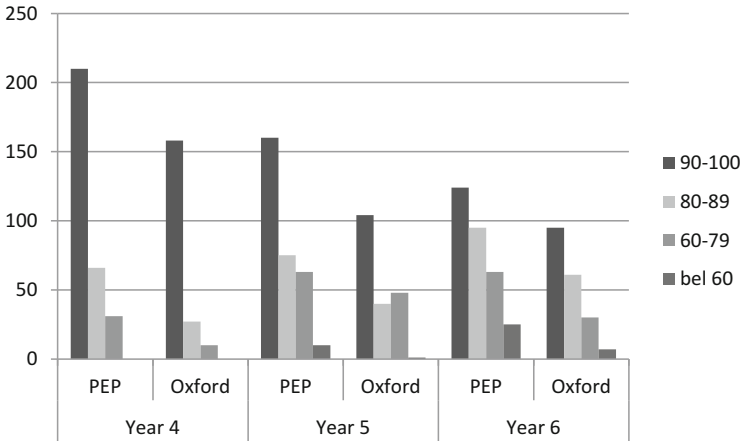


Fig. 2 Number of students in the four score bands

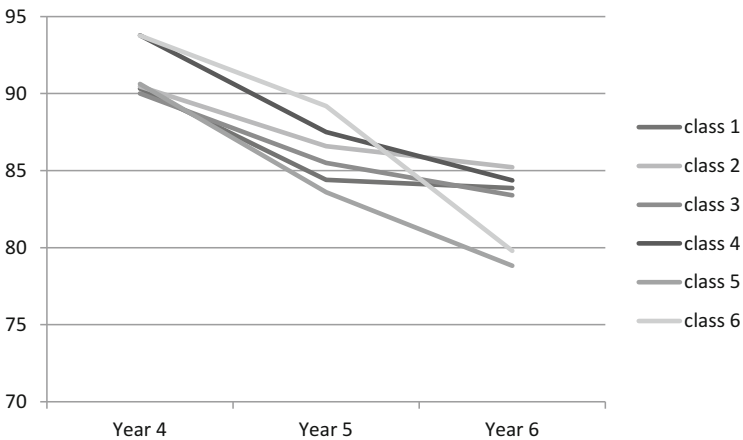


Fig. 3 Mean score attained for six classes in the PEP group

3.2.3 Mean Scores in Different Classes

We computed the mean score for each class in the PEP and the Oxford groups, as depicted in Figs. 3 and 4. The vertical axis denotes the mean score attained by the different classes. There was a general trend of decline in the mean scores in all 11 classes as they entered higher grades. In the Oxford group, however, the situation changed in Year 6: the mean scores in classes 8 and 10 increased slightly. Over 3 years, in the PEP group, the mean score ranged between 75 and 95; whereas it was

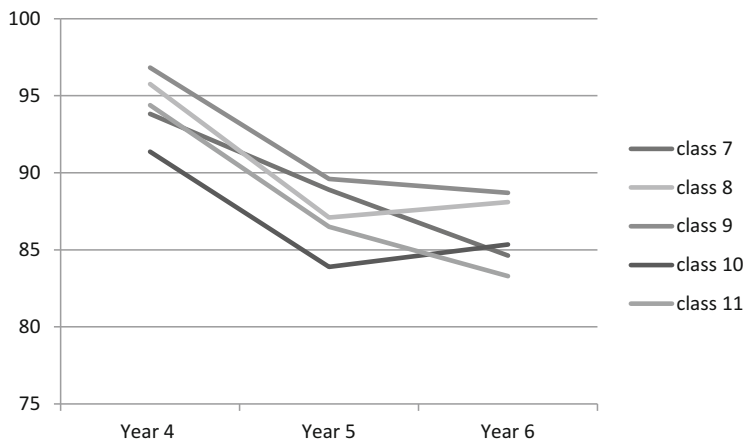


Fig. 4 Mean score attained for five classes in the Oxford group

between 80 and 100 in the Oxford Group. By this measure, it is safe to say that most students in the 11 classes performed well on the tests; those in the Oxford group, overall, performed better than their peers taking the PEP tests.

3.3 Teachers' Views on the Tests and Test Results

We have described and delineated above the English tests of the PEP and the Oxford groups at CNPS. An attempt was made to answer what the tests comprised, what test formats were used to assess what language areas, and how the students performed on the tests over 3 years. This section explores what the teachers and test developers have to say about the tests, how they scored the tests, and we intend to probe into some of the issues in test quality. Another focus is on students' performance, how they performed and why.

3.3.1 The Use of Oral Test

As was shown in Table 3, the oral test took up 50 % in grade 4, 40 % in grade 5, and no oral component was used in grade 6. When asked why she included oral tests, Teacher 1, an item writer, explained her belief as follows:

We (and I) believe... learning to "speak" English at critical ages would exert great influence on children's EFL study. Thus, it's necessary to develop oral tests to signal that oral abilities are important.

(Teacher 1, interview extract, 05/07/2014)

This view concurs with the literature on children's language learning indicating that oral abilities play a critical role. According to Hadley (2001), spoken language

is the main channel of communication and we need to convince learners that communicative language use is the major goal of English instruction once they begin to learn English. Secondly, pupils may be more motivated to learn spoken rather than written language and therefore can achieve higher proficiency (Wilkins, 1974). McKay (2006) also argues that young learners are able to try out their hypotheses about language, receive feedback and form new hypotheses through oral language interactions with the teacher and with each other. She suggests putting the assessment of oral language at the center of assessment in a young learner program because oral language “provides the foundation for language learning” (p. 214).

While most teachers agreed on the importance of teaching students speaking, some doubted the necessity of assessing it. They voiced disappointment about the efficacy of the oral tests. For one thing, most of the items (about 91 % in the PEP test of grade 4, for example) were directly lifted from the textbook with strong familiarity and predictability for the students who were informed about the test content. For another, the scoring of such tests depended on the “feel” of the scorers. Teacher 3 reported how he and his colleagues usually graded the oral tests:

We just follow the gut. But you know sometimes it's hard to differentiate students' performance with scores ... After all, an oral test is all about having fun, why do we have to ruin the mood?

(Teacher 3, interview extract, 05/07/2014)

By “following the gut”, the scorers did not refer to any guidelines or rubrics in the grading process, which may compromise the reliability of the scores. However, in Teacher 3's understanding, this sacrifice was necessary to accommodate for the needs of young language learners. He further added:

Learners of English at this age are very unlikely to speak English unless they are asked to. So we should give them a break when assessing them, otherwise they will be discouraged.

(Teacher 3, interview extract, 05/07/2014)

Teacher 2 resonated with this view:

It's all about making them feel happy about their scores. So normally, we assign to 90 % of the students the full score, and for the other 10 % who can't perform what we have taught or do not utter a word at all, we give them 80 % of the score for this oral part.

(Teacher 2, interview extract, 05/07/2014)

In this vein, the oral tests served to please children rather than to assess them. This, fueled by the huge time-consumption in administration, some teachers proposed a modification of the present oral tests, while others suggested its cancellation.

Another interesting observation is that Table 3 clearly depicts a general decrease of items designed for oral tests, which according to Teacher 2, stood in line with how English teachers at CNPS prioritized their teaching goals.

The makeup of test items doesn't come from nowhere. For example, in low grade, we believe speaking should be given priority. In response, we develop a high proportion of these items in grade 4. As students enter higher grades, we shift the focus to vocabulary and grammar. Hence we design no oral test in grade 6.

(Teacher 2, interview extract, 05/07/2014)

3.3.2 The Use of Listening Tests

The listening section occupied a large portion in the written tests of both the PEP and the Oxford groups, accounting for 40 % in grade 4, around 33 % in grade 5 and 30 % in grade 6. The consideration of devising so many listening items, according to Teacher 4, is to:

...emphasize the input on the part of children so that ...the likelihood of them producing increased language output may not be a fantasy.

(Teacher 4, interview extract, 05/07/2014)

This understanding may find its root in theories of second language acquisition. With insights gained from studies of child language acquisition, Byrnes (1984) highlights the key role listening plays in the development of a learner's second language, particularly at the beginning stages of language development. Without the input provided by listening at the right level, learning cannot begin (Nunan, 1999). McKay (2006) also argues that "listening needs its own profile in assessment" (p. 207) in that it plays an important role, not only in language learning, but also in learning in general.

Despite their huge number, most listening items (as shown in Table 5) were constructed to target students' ability to discriminate between phonemes, with very little emphasis on processing at the semantic level to understand the meaning of an utterance. As Chastain (1979) put it, these items may be valid for testing conscious knowledge of the language, but they are not realistic indications of the ability to comprehend a spoken message. In real life situations, even when occasional confusions over selected pairs of phonemes are made, listeners can still use contextual clues to interpret what they hear. By this measure, the listening test was of a traditional kind, which Teacher 2 justified:

Listening poses much challenge to children...we didn't use too many items to assess "how well they understand a message", not least because children are still limited in the ability to use vocal keys to unlock the meaning of the communication.

(Teacher 2, interview extract, 05/07/2014)

It seems that the skill of "understanding a message" has given way to "recognizing and discriminating sounds". But again, is "understanding a message" something we should expect from English learners at the beginning stage? Teacher 6 gave no to this question:

Should we not be more concerned with children understanding how English "sounds" than what it means?

(Teacher 6, interview extract, 05/07/2014)

3.3.3 The Use of Comprehensive Skills Tests

It was found that most items (35.7–50 %, as shown in Table 6) in this section of the PEP and Oxford tests were multiple choice items. Why use these items? Teacher 1 offered her explanation:

We have a lot of content to cover in a test paper and multiple choice items can do that for us. They can assess more topics than what can be squeezed into other forms of questions, and also they are highly reliable and objective.

(Teacher 1, interview extract, 05/07/2014)

However, McKay (2006) cautions about the danger of some multiple choice items eliciting only selected or limited response, hence they are to be used with more care with young learners. In the tests of the PEP and the Oxford groups, we found that up to 91 % and 83 %, respectively, of the items assessing grammar and vocabulary were designed as multiple choices. While such items assessing individual grammatical forms (e.g., third person singular) focus on accuracy, they do not involve children in purposeful, creative and spontaneous language use in a particular situation (McKay, 2006) because they lack contextual support and authenticity (Zhan, 2007). Likewise, Purpura (2004) commented that they are “old-fashioned and out-of-touch with students’ language learning goals” (p. 253).

Williams (1984) pointed out that language use tasks, similar to those used in the classroom can be reused for assessment of young learners (doing puzzles, solving problems, listening to and retelling stories, etc.). However, using these tasks for assessment means more than handing students a piece of test paper. The administration may be more complex and impractical for teachers at CNPS, each of whom was responsible for more than 40 or even 50 students. Besides, the scoring may be more subjective than using multiple choice items. Considering both sides of the coin, Teacher 3, when she was asked to make a choice, said:

I would still stick to multiple choices because they are more objective items. They make it easier for us to ensure fairness in scoring children.

(Teacher 3, interview extract, 05/07/2014)

Her view is corroborated by Brown (1996, p. 29) who phrases this awareness as “a tendency to seek objectivity” in assessment. But he also points out that many of the elements of language courses may not be testable in the most objective test types. For this reason, among others, CNPS devised a number of essay writing tasks in both groups to assess how well students can use the English language to communicate meaning. These items often provided cue words/sentence pattern guidance in the target or the source language to help students compose a short passage on a topic. However, testing writing in primary school has been the subject of much controversy. Teacher 5 voiced her doubts about constructing the essay items:

I understand the importance of writing. But we seem to follow the logic that since we have listening and reading (input), there must be writing (output). And students might find it perplexing to put into so much effort expected to write a passage, yet attaining at most five marks.

(Teacher 5, interview extract, 05/07/2014)

Teacher 4 reported how some of students came to her complaining about their low scores on the writing item:

Some students were so discouraged that they asked me why they were given a low score, but, you know, actually, 80 % of the students get below three marks...we have so much to take into consideration in the grading of writing, such as spelling, grammar, etc. Once we spotted a mistake, 0.5 mark would be taken away.

(Teacher 4, interview extract, 05/07/2014)

In light of the possible adverse effects of assessing children's writing, Heaton (1988) contends that it is ludicrous to expect skills in a foreign language which the students do not possess in their own language. Therefore, it might be understandable that writing should play a secondary role for students at the primary stage (Zhan, 2007).

3.3.4 Language Areas Assessed in the Tests

As far as the assessed language areas are concerned (shown in Fig. 1), it could be seen that while both the PEP and the Oxford tests concerned four language areas, the PEP tests focused more on testing the first three (i.e. phonology, vocabulary, grammar) than the Oxford test using many items to assess notion-function of language use. It could also be inferred that item writers for the PEP group took a structural approach to language testing, whereas those in the Oxford group adopted a more communicative approach (Heaton, 1988). In grade six, for example, items assessing notion-function were assigned as many as 25 marks in the Oxford English tests. So why did the PEP and Oxford test differ in the assessed language areas? Teacher 4 interpreted this as a result of different textbooks and teaching.

We have to test what we teach and how we teach. Oxford English is structured in a way that emphasizes the use of real-life and practical language while PEP highlights the importance of flowing from words to sentences, then paragraphs.

(Teacher 4, interview extract, 05/07/2014)

In this sense, the differentiation reflected the respective textbooks and the methodologies they followed. Therefore, the items were aligned with desired outcomes defined in the textbooks. If so, then item writers in both groups did a good job. As stated by Heaton (1988), when a more traditional, structural approach to language learning has been adopted, the test should closely reflect such a structural approach. The same goes for the communicative approach. A study by Li (2010) also reported that many local English tests in China at the primary stage assessed individual language performance depending on the curriculum to which pupils were exposed, thus the selection of the test contents and materials was fully combined with teaching objectives and teaching materials.

It is reasonable to state that test writers followed the guidance of teaching materials to develop what they believed to be a good test, which acted as an obedient servant since it followed and aped the teaching (Davies, 1968). However, Hughes (1989) proposed that we cannot expect testing to follow teaching only. Instead, testing should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching. According to communicative language testing theories, "bad teaching" only makes clear what learners know about the language and not how they use the target language in the appropriate context, irrespective of assisting them to use language knowledge in meaningful communicative situations (Canale & Swain, 1980). To change that, using more items assessing the notion-function of language may facilitate good teaching and induce preferable learning outcomes on the part of children.

3.3.5 Test Quality

More and more researchers (e.g., Bachman, 1990; Berry, 2008; Shu, 2001) agree that scientific testing entails the integration of validity and reliability to ensure its quality.

3.3.5.1 Validity

For a test to be valid, it has to credibly measure what it is designed to measure. Therefore, Phelan and Wren (2005) suggest that while constructing a test, teachers need to define and operationalize the learning outcomes (expectations) for students and align each item with a specific goal. In other words, one needs to compare what is required with what is being assessed. As for whether CNPS has put validity on its agenda of tests development, Teacher 2 claimed this:

Although many teachers are not aware of the term “validity”, they actually have been doing things to serve this purpose, such as analyzing textbooks and syllabus, and using highly-credited test papers for reference. However, some teachers think it’s time-consuming and unnecessary. After all, everyone has different methods of developing an item in the way they deemed ok.

(Teacher 2, interview extract, 05/07/2014)

Nonetheless, a threat lurking in newly-arrived teachers’ understanding of tests validity warrants caution. Teacher 4 recalled what a novice teacher once told her:

You guys are overreacting...writing items should be easy. I don’t know why you take it so seriously. We have reference books from which we can see clearly what we are going to do. We don’t need to do such a thing as validity check, don’t you think?

(Teacher 4, interview extract, 05/07/2014)

3.3.5.2 Reliability

However, simply taking good care of validity does not qualify a good test. According to Heaton (1988), for a test to be valid at all, it must be reliable as a measuring instrument. Reliability has to do with the consistency of an examinee’s performance on the test, i.e. the extent to which the results can be considered consistent or stable (Brown, 1996). Hughes (1989) points out that there are two components of test reliability: the performance of candidates from occasion to occasion, and the reliability of the scoring. The first reliability can be estimated with a strategy called the test-retest method, which administers the test in question two times to the same group of students. Once completed, the pairs of scores for each student are lined up in two columns, and a Person product-moment correlation coefficient can be calculated between the two sets of scores. The test-retest method has never been used at CNPS due to the skepticism about the necessity and feasibility of conducting such an

analysis for small-scale, school-based test papers in primary settings. Teacher 2 explained it as follows:

Some teachers have not come to appreciate the value in using statistical theory to ensure test paper quality. Also, budgets and teaching schedules do not allow these types of analysis to take place even if the teachers want to.

(Teacher 2, interview extract, 05/07/2014)

Indeed, though researchers have substantiated with many publicly used tests the significance of catering to issues of reliability (e.g., Ao, 2002; Choi, 2008), few studies have been carried out to probe into small tests, which, however, does not provide an excuse for item writers for not bearing in mind some factors affecting reliability.

When we asked what the teachers had done to keep reliability at a desirable level, the inquiry met with a detailed explanation of the test construction process. We summarized three item writers' words as follows in a way that attempted to describe the process as clear and brief as possible:

We (item writers) would gather together several times to discuss details as to what to incorporate in the tests and how to distribute the weighting. Each one will assume responsibility for one section of items. Following the completion of the first draft, the test paper will be subject to critical scrutiny by another item writer. Then, it is sent to the Jiaoyanyuan (a leading figure in subject teaching in the district. The candidate is appointed by the local educational institution to supervise and evaluate the teaching at school levels) from Teachers' Training Institution in Nan'an District, who reviews the paper and offers suggestions in regard to the paper quality. The final version will then be printed and prepared for administration.

We also asked whether CNPS had given any thought to the second reliability, the scorer reliability. In response, Teacher 2 said:

Because scores of students largely depend on the quality of their response against the criteria set by the scorers, at the beginning of grading, I (or the head of the English department) gather teachers from the same grade to discuss the scoring criterion, especially in the case of subjective items. On some occasions, a detailed scoring key specifying acceptable answers and assigning points will be given out to them.

(Teacher 2, interview extract, 05/07/2014)

Grading began only after scorers agreed upon the criterion. The test papers were randomly distributed to each scorer. In the process of scoring, a leader (usually a backbone teacher) assumed responsibility for clearing doubts in terms of the scoring standards. After the completion of grading, teachers were involved in producing "score reports" with basic analysis of data.

Nevertheless, it was found that not every scorer was willing to toil through such a rigorous grading procedure. For example, in light of the strong subjectivity to personal judgment, oral tests and writing tasks entailed huge demands on scorers. However, oral tests were graded in a more causal way to encourage students' speaking. As for writing items, Teacher 3 filed his complaint:

It's not like those high-stakes tests where the scores decide someone's fate. Personally, I don't favor the idea of making a fuss in grading (writing) even if we are told to.

(Teacher 3, interview extract, 05/07/2014)

Despite some teachers' resistance to changing their grading behaviors, many other scorers demonstrated attentiveness and patience in grading. Teacher 1 talked about why she would endure the painstaking job:

You can only imagine how much scores matter to our children. One can never be too careful while grading.

(Teacher 1, interview extract, 05/07/2014)

This view is in line with teachers' willingness to improve test quality through teacher training in quantitative analysis. Teacher 2 envisaged that:

I hope professionals and experts in English language assessment will come to our rescue. Even though we know little about some testing theories and statistical analysis, we are never afraid of embracing the challenges when it means we can improve teaching and learning.

(Teacher 2, interview extract, 05/07/2014)

3.3.6 Factors Influencing the Performance of Students

It was demonstrated in Table 7 that the mean level of difficulty for both tests was relatively close. Then why did the Oxford group perform better than the PEP group when they took the test of approximately the same level of difficulty? In this section, we asked the teachers and item writers what they thought and report three main reasons:

3.3.6.1 Textbooks

A surprising finding is related to the textbooks. According to Teacher 2:

Textbooks play a critical role in affecting what item writers put in the test papers. In PEP, we cover a larger sum of language points (e.g., grammar structures like preposition of place) than Oxford, which may be overwhelming for children in PEP.

(Teacher 2, interview extract, 05/07/2014)

It appears that teachers' assessment and students' performance in the PEP group were constrained by the textbook. When asked if the textbook really meant the problem, Teacher 4 said this:

We should not be shackled by textbooks. Actually, it is how we use them that determine our teaching outcomes. I think we should induce change in our teaching...any adjustment can be made possible if you embrace it. We all want the same things; don't shackle yourself just because the textbook says so.

(Teacher 4, interview extract, 05/07/2014)

3.3.6.2 Teachers and Activities

The destination is the same, the route of arriving there makes a difference. While English proficiency was what teachers intended for students' learning outcomes, the Oxford group approached it through language-focused activities and games.

Teachers used diversified and dynamic teaching methods to help students enjoy the language-embedded activities. In contrast, in the PEP classrooms, few activities were introduced, some of which were non-language related. Teacher 1 commented on how activities varied:

In PEP, we have to take much time to deal with words, sentence patterns as such. Sometimes, we design activities just because students are tired from learning and we want to cheer them up. But in Oxford, we integrate games in learning, and give children opportunities to practice language, which slides into their heads without them knowing.

(Teacher 1, interview extract, 05/07/2014)

As Chou (2012) pointed out, using games or other forms of play without a clear objective related to language learning is likely to result in ineffective learning, in the sense that the pupils will be unable to demonstrate what they have learned in class through games. However, language-oriented and learner-centered games in the language classroom can yield desirable results. McKay (2006) reports that language-rich activities or games involving doing, thinking and moving can be used to provide children with opportunities to listen and guess from the context, to risk using the target language, and to engage in interactions. Therefore, it might be argued that students in the Oxford group benefited more from carefully designed and language-related activities than their peers in the PEP group and this is why they demonstrated a higher level of English proficiency.

However, students in the Oxford group progressed at a slower pace. The good performances of the Oxford students emerged only after a period of time when students in the PEP group were already making strides ahead. According to Teacher 2:

At the beginning of learning, students find it a headache to keep up with the pace of learning in Oxford textbooks because we have so many activities and things to learn. But as time went by, they have displayed much higher English proficiency.

(Teacher 2, interview extract, 05/07/2014)

3.3.6.3 Motivation

If we consider students' overall performance on the test over 3 years, it could be inferred that a trend emerged: students in both groups, once they became seniors, were not performing as well as in lower grades. The mean scores dropped and the number of those failed also increased as the year went higher. One reason may be that the level of difficulty of the tests increased slightly over the years. Still, the P values of 3 years were so close that one might question whether it was the major force bringing about the decline in the students' scores. As to what the reason may be, we asked all the seven interviewees, all of whom mentioned a common theme: de-motivation.

As a complex psychological construct, motivation is regarded as one of the determinant factors in successful foreign language learning (Lasagabaster, 2011). Studies carried out with young language learners in many different contexts have demonstrated that there is a clear positive correlation between motivation and language achievement (e.g., Jia, 1996; Soto, 1988). Therefore, since the students at CNPS experienced a decrease of achievement, we asked teachers about this phenomenon.

Teacher 1 observed such a demotivation but said it manifested much difference between “good” students and “bad” students. She explained that:

For some students with poor foundation of English, they become more and more disinterested in English learning. Because, you know, as students enter higher grades, the learning materials become more demanding on skills and knowledge. So, these students can't keep up with the learning schedule and lag behind. But for those good students, who always study hard and achieve good grades, they keep it that way and even grow fond of English learning.

(Teacher 1, interview extract, 05/07/2014)

From her words, we see that increasing content complexity was one of the internal factors demotivating students. This view is also supported by Teacher 2:

Knowledge covered in textbooks rolls like a snowball over the years. We encounter more boring grammatical structures and vocabularies. Some students are afraid that ... they are unable to tackle the “hard” part, trying to run away from English and saying it is demon.

(Teacher 2, interview extract, 05/07/2014)

Facing the complex content crisis, teachers tended to cut down or omit fun activities and introduced more serious but boring tasks in higher grades so that they could focus on dealing with the “hard” parts in a step by step fashion. However, this may be the very reason why students became discouraged. A longitudinal study by Nikolov (1999) looked into how Hungarian children’s motivation changed over their 8 years of learning English. She found that for children (ages 6–14), intrinsically motivating activities, tasks and materials meant one of the most important motivating factors.

Deprived of the time spent on learning by doing and playing, students have manifested negative attitudes towards learning English. Apart from this, learning a subject for such a long a time emerged as the second reason accounting for the abatement of motivation, as pointed out by Teacher 2:

Students become more impatient in classes. Some of their parents come to us, reporting that their children have complained that they have studied more and longer than they could handle.

(Teacher 2, interview extract, 05/07/2014)

This is verified by a study of Davies and Brember (2001), who measured attitudes of second and sixth grade students using a Smiley-face Likert scale. They found that all participants harbored significantly less positive attitudes in the higher grade, and concluded that the more years students spent studying a subject, the more disenchanting they became with it.

The third factor has something to do with the abolishment of the general graduation examination in elementary school: since 2011, after graduation from primary school, children automatically enter a neighborhood middle school without taking any form of exams or tests. Since then, less pressure has been endured by the students to fight for better grades. As Teacher 5 observed:

Without struggling through a formal examination to win a ticket to a middle school, some students are slacking off in school, paying less attention during the class session and skipping their homework.

(Teacher 5, interview extract, 05/07/2014)

4 Conclusions

This case study analyzed the achievement test results of the students tested in their Years 4, 5 and 6 at CNPS between 2010 and 2013. Through examining the test papers in the PEP group and the Oxford English group, we have answered questions concerning the component, item types and language areas measured of the two sets of tests. Then by looking into the students' scores, we have attempted to understand how the students performed on the two tests over 3 years and investigate the differences between two groups. A follow-up interview brought us closer to what the teachers and test writers at CNPS built their teaching and testing beliefs upon.

The results document the commitment of teachers and administrators to catering to children's needs by developing well-scrutinized achievement tests. However, we found that not all the seven interviewees interpreted the test scores in a way that provides feedback on how students learn, how they perceive the learning process, and then inform teaching in the best interests of their students. In addition, endeavors have been made to look at how children had performed and analyze the reasons contributing to their performance.

4.1 Implications for Practice

The study bears implications for using achievement tests to assess young EFL learners in elementary schools. The findings contribute to the body of evidence of how primary schools apply language assessment and what can be done to refine test papers and improve teaching, which entails teamwork where teachers, school administrators as well as students themselves all play a part.

For teachers, as indicated in the difference of students' performance and motivation between two groups, they are advised to reflect on how they use textbooks, how they teach and develop good quality tests (see also Hsieh, 2016 in this book). To accommodate the young language learners' age and personality, both teaching and assessment need to be engaging and flexible, without intimidating children and causing boredom. In addition, testing should not be limited to measuring the learning results, but also serve to provide feedback for teaching and support for learning. As Berry (2008) points out, the paradigm of assessment should not only be of learning, but more importantly, for and as learning, which "places special emphasis on the

role of the learner and highlights the use of assessment to increase learners' ability to control their own learning" (p. 9).

When teachers prepare themselves for the changes, the question ensuing would be whether school administrators would support the reform in teaching and assessment and welcome the new ideas that might seem to undermine and even contradict what is prescribed by the education authorities and what is expected by those parents who care only about higher grades. Whereas the general graduation examination in elementary schools has been cancelled, the mindset of some school leaders and parents are still score-oriented. This in a way poses a threat to teachers exploring better ways to serve pedagogy.

It should also be noted that students can participate in the assessment process by providing feedback to teachers on how they feel towards the test, what they think is difficult or easy. With information of this kind, it would give teachers some perspectives on what the students have learned, whether the test has achieved the goals set in their mind, what to do in the next phase of teaching. However, it cannot be substituted for the analysis of test papers and test scores.

4.2 Limitations and Future Directions

The limitations of the present research are manifold. First, we were not able to conduct statistical analyses of test items, because teachers and school administrators failed to store and allow us to process raw scores. It raises an imminent question whether schools like CNPS should at all evaluate the scores of small-scale, school-based, non-public used test papers using quantitative methods. However, from the in-depth interviews with teachers, we find that, despite their impoverished sensitivity to checking test paper quality, they expressed willingness to use what they called "high-above theory" to guide their test development. Some teachers have already begun to consider issues like reliability and validity, and they are looking forward to receiving training in assessment. It is hoped that teachers will use the expertise they gained to create and administer tests, and to interpret the evidence they generate in a scientific way, and eventually, they will be able to reflect on the findings in order to change their practice.

Second, although we have managed to probe into the motivational factors exerting influence on students' performance through interviews, the participants were only seven teachers. Some of their opinions could be personal and biased without cross-checking with students what actually happened to them and what they truly thought. Also, to what extent the motivational factors have contributed to their difference in performance remains to be investigated.

Another drawback is that what has been explored at CNPS cannot be generalized to the whole picture of English language tests in China at the primary level. Given the specific learning context and the relatively small sample size, future research in other contexts and with a wider population of children of the same age group is much warranted.

Appendix 1

Interview Questions

For Test Writers Only

1. Briefly illustrate the procedure in which you design the test paper.
2. Explain how you divide the test paper into three components. What is the rationale in developing each component?
3. How do you decide which test format or task types to use? Why is multiple choice item the most frequently used?
4. Why does the PEP test cover more grammar and vocabulary than the Oxford test which has many items assessing function-notion?
5. How do you ensure the quality of the test paper?

For All the Interviewees

6. Explain how the test is administered and scored.
7. As users, what do you think of the tests? What do you like or dislike about the test?
8. Explain why students' performance decline over the years. What are some of the factors?
9. Why, from your understanding and observation, do students in the Oxford English group perform better than those in the PEP English group?

References

- Ao, S. (2002). *English language testing in China: A survey of problems and suggestions for reform*. Unpublished master's thesis. Ghent University, Brussels, Belgium.
- Bachman, L. F. (1990). *Fundamental concepts in language testing*. Oxford, UK: Oxford University Press.
- Bailey, A. L. (2005). Test review: Cambridge young learners English (YLE) tests. *Language Testing*, 22(2), 242–252.
- Berry, R. (2008). *Assessment for learning*. Hong Kong, China: Hong Kong University Press.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27(1), 5–31.
- Byrnes, H. (1984). The role of listening comprehension: A theoretical base. *Foreign Language Annals*, 17(4), 317–329.
- Cameron, L. (2001). *Teaching languages to young learners*. Cambridge, NY: Cambridge University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1–47.
- Chastain, K. D. (1979). Testing listening comprehension tests. *TESOL Quarterly*, 13(1), 81–88.
- Chik, A., & Besser, S. (2011). International language test taking among young learners: A Hong Kong case study. *Language Assessment Quarterly*, 8(1), 73–91.

- Choi, I. C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing*, 25(1), 39–62.
- Chou, M. H. (2012). Assessing English vocabulary and enhancing young English as a foreign language (EFL) learners' motivation through games, songs, and stories. *Education*, 3(13), 1–14.
- Davies, A. (Ed.). (1968). *Language testing symposium: A psycholinguistic perspective*. London, UK: Oxford University Press.
- Davies, J., & Brember, I. (2001). The closing gap in attitudes between boys and girls: A 5-year longitudinal study. *Educational Psychology*, 21(1), 103–114.
- Fleurquin, F. (2003). Development of a standardized test for young EFL learners. In *Spain Fellow Working Papers in Second or Foreign Language Assessment*, 1(1), 1–23.
- Gardner, S., & Rea-Dickins, P. (2001). Conglomeration or chameleon? Teachers' representations of language in the assessment of learners with English as an additional language. *Language Awareness*, 10(3), 161–177.
- Gong, Y. F. (2003). *PEP primary English students' book*. Beijing, China: People's Education Press.
- Gronlund, N. E. (1993). *How to make achievement tests and assessments*. Needham Heights, MA: Allyn & Bacon.
- Hadley, O. (2001). *Teaching language in context*. Boston, MA: Heinle & Heinle.
- Halliwell, S. (1992). *Teaching English in the primary classroom*. London: Longman.
- Hasselgren, A. (2005). Assessing the language of young learners. *Language Testing*, 22(3), 337–354.
- Heaton, J. B. (1988). *Writing English language tests*. New York: Longman.
- Hsieh, C.-N. (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*. New York: Springer.
- Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.
- Hymes, D. H. (1972). On communicative competence. In *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin.
- Jacobs, L. C., & Chase, C. I. (1992). *Developing and using tests effectively: A guide for faculty*. San Francisco: Jossey-Bass.
- Jia, G. J. (1996). *Psychology of foreign language education*. Nanning, China: Guangxi Education Publishing House.
- Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.
- Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of Human Resources*, 37(4), 752–777.
- Lasagabaster, D. (2011). English achievement and student motivation in CLIL and EFL settings. *Innovation in Language Learning and Teaching*, 5(1), 3–18.
- Li, S. L. (2010). *The communicative English testing framework for students at primary stage*. Unpublished master's thesis, Gannan Normal University, Jiangxi, China.
- McKay, P. (2006). *Assessing young language learners*. Cambridge, UK: Cambridge University Press.
- Ministry of Education. (2001a). *Basic requirements of English teaching in elementary school*. Beijing, China: Beijing Normal University Publishing Group.
- Ministry of Education. (2001b). *The new English curriculum standards*. Beijing, China: Beijing Normal University Publishing Group.
- Ministry of Education. (2011). *Standard of English curriculum for basic education*. Beijing, China: Beijing Normal University Publishing Group.
- Morrow, K. (2012). Communicative language testing. In C. Coombe & B. O'Sullivan (Eds.), *The Cambridge guide to second language assessment* (pp. 140–146). Cambridge, NY: Cambridge University Press.

- 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. (2016). Trends, issues and challenges in assessing young language learners. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives*. New York: Springer.
- Nunan, D. (1999). *Second language teaching & learning*. Oxford, UK: Heinle & Heinle Publishers.
- Phelan, C., & Wren, J. (2005). *Exploring reliability in academic assessment*. Retrieved January 15, 2014, from <http://www.uni.edu/chfasoa/reliabilityandvalidity.htm>
- Pinter, A. (2006). *Teaching young language learners*. Oxford, UK: Oxford University Press.
- Purpura, J. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.
- Shi, J. P. (2010). *Oxford English*. Shanghai, China: Shanghai Foreign Language Education Press.
- Shohamy, E. (2001). *The power of tests: Critical perspectives on the uses of language tests*. Harlow, UK: Longman.
- Shu, Y. X. (2001). *The theories and methods of foreign language testing*. Beijing, China: World Book Publishing Company.
- Soto, L. D. (1988). The motivational orientation of higher- and lower-achieving Puerto Rican children. *Journal of Psychoeducational Assessment*, 6(3), 199–206.
- Vale, D., & Feunteun, A. (1995). *Teaching children English, a training course for teachers of English to children*. Cambridge, NY: Cambridge University Press.
- Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.
- Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.
- Wilkins, D. A. (1974). *Second-language learning and teaching*. London: Edward Arnold.
- Williams, M. (1984). A framework for teaching English to young learners. In C. Brumfit, J. Moon, & R. Tongue (Eds.), *Teaching English to children*. Harlow, UK: Longman.
- Yao, J., & Duan, H. C. (2004). Research on the difficulty distribution of test papers based on the Monte Carlo method. *Computer Applications and Software*, 21(9), 66–67.
- Zhan, C. F. (2007). *The multiple-level English testing framework (MLETF) for English teaching at primary stage*. Unpublished master's thesis, Guangxi Normal University, Guangxi, China.

Jing Peng is a professor in Research Centre of Language, Cognition and Language Application, Chongqing University, China. She teaches English Curriculum and Instruction to MA students at College of Foreign Languages And Cultures. Her areas of interest include: teacher education, teacher development, and ELT methodology. In recent years she has done school, regional and national projects for primary and secondary schools and published extensively in the areas of ELT curriculum innovations, methodology, and teacher education. She has also been involved in teacher education program sponsored by the Tin Ka Pin Foundation, Hong Kong, and projects supported by British Council.

Shicheng Zheng is a graduate student major in English curriculum and instruction, Chongqing University. He has been working as a teaching assistant at College of Foreign Languages And Cultures. His areas of interest include: curriculum, teaching methodology, and teacher talk. He has done several studies of the curriculum development and methodologies of teaching English as a foreign language.

Individual Learner Differences and Young Learners' Performance on L2 Speaking Tests

Jelena Mihaljević Djigunović

Abstract This chapter focuses on motivation and self-concept and their role in oral production in early learning of English as a foreign language. A review of major research findings considering the relationship of these individual learner differences and oral performance by young foreign language learners is followed by presentation and discussion of the study the author carried out with Croatian learners of English as a foreign language. The participants, aged 11 at the start and 14 at the end of the study, were followed for 4 years. Each year their motivation and self-concept were measured by means of smiley questionnaires and oral interviews, while their oral production was elicited each year through picture description tasks and personal oral interviews. The study offers interesting evidence of the dynamics of young learners' motivation and self-concept and their relationship with their developing oral performance. Implications of the findings are considered as well.

Keywords Young learners • Individual learner differences • Motivation • Self-concept • Oral production • English as a foreign language

1 Introduction

Although children are still commonly thought to be highly similar to each other when language learning is concerned, recently research into individual learner differences has extended to young L2 learners as well. Thus, major publications in the early L2 learning field increasingly include sections on how young language learners differ in their approach to L2 learning as well as in various aspects of the language learning process and learning outcomes (e.g., Enever, 2011; Muñoz, 2006; Murphy, 2014; Nikolov, 2009a, 2009b). Attitudes and motivation of young L2 learners have perhaps been investigated the most extensively leading to whole volumes devoted to the topic (e.g., Heinzmann, 2013). Some attention has been paid to young learners' language aptitude (e.g., Alexiou, 2009; Kiss, 2009; Kiss & Nikolov, 2005), learning strategies (e.g., Kubanek-German, 2003; Lan & Oxford, 2003;

J.M. Djigunović (✉)

Department of English, University of Zagreb, Zagreb, Croatia

e-mail: jdjigunovic@gmail.com

Mihaljević Djigunović, 2002; Šamo, 2009; Tragant & Victori, 2006), attributions (e.g., Julkunen, 1994), language anxiety (e.g., Low, Brown, Johnstone & Pirrie, 1995; Seebauer, 1996) and self-concept (Julkunen, 1994; Mihaljević Djigunović, 2014). In some studies interactions between different individual learner characteristics as well as with some contextual factors were also investigated.

In the present study we focus on young learners' oral performance in English as L2 and two individual differences: motivation and self-concept. While the relationship of L2 achievement with the first learner factor has been the focus of interest for some time now, self-concept has only recently caught the attention of young learner researchers.

Most empirical studies suggest that there is a significant relationship between motivation and language learning achievement. Thus, Harris and Conway (2002) report on more motivated Irish young learners of French, German and Italian being more successful at these languages than their less motivated peers. Such a positive relationship has been found in other studies, and has been shown to be evident with learners as young as four (e.g., Bernaus, Cenoz, Espi & Lindsay, 1994) as well as with 14-year-olds (e.g., Bagarić, 2007; Dörnyei, Csizér & Németh, 2006). However, the relationship seems to be quite complex once we take into account different types of measures of learning outcomes, or age and learning experience of young L2 learners as well as types of motivation. Studies have, thus, shown that motivation is less strongly correlated with objective measures of language achievement than with teacher-assigned grades or with learner self-assessment (Margoret, Bernaus & Gardner, 2001). Tragant and Muñoz (2006) have found motivation to be more significantly related to performance on integrative than discrete-point measures. Quite a few studies (e.g., Graham, 2004; Masgoret & Gardner, 2003; Tragant & Muñoz, 2000) have indicated that correlations of motivation with language achievement tend to decrease with increasing age of learner.

Mercer (2011) defines the L2 learner's self-concept as 'an individual's self-description of competence and evaluative feelings about themselves as a Foreign Language (FL) learner' (p. 14). Highlighting the importance of L2 self-concept, Arnold (2007) says that '(l)earners must both *be* competent and *feel* competent.' (p. 18). Due to the common belief that young learners have a positive self-perception as if by default, until recently this affective learner variable was not considered a relevant topic in the early L2 learning field. However, with increasing interest in researching young learners the young L2 learner's self-concept has become a potentially important variable which could offer deeper insight into early L2 learning processes. Harter (2006) claims that children tend to develop too positive self-perceptions because it is difficult for them to distinguish between their real and ideal selves. Based on self-rating of their abilities, Pinter (2011) calls young L2 learners 'learning optimists'. Damon and Hart (1988) suggest that young learners' self-knowledge becomes more complex as they mature. Kolb (2007), however, claims that children possess quite high awareness of their L2 learning process and entertain complex language learning beliefs: they base these on their learning experiences and personal knowledge. Studies by Wenden (1999) and Mihaljević Djigunović and

Lopriore (2011) also suggest that young learners are capable of participating in reflective activities and providing relevant and important data on their L2 learning process. Mihaljević Djigunović and Lopriore have found that young L2 learners display both inter- and intra-learner variability in their L2 self-concept. In her comparative study of children who started L2 learning earlier (at age 6) and those who started later (at age 9), Mihaljević Djigunović (2016) has found that the development of L2 self-concept of earlier and later starters follows different trajectories.

In the past two decades or so assessment of early L2 learning outcomes has focused on different aspects of language achievement. Among these a number of studies have been dedicated to reseaching the mastery of some or all of the four language skills (e.g., García Mayo & García Lecumberi, 2003; Harris & Conway, 2002; Low, Duffield, Brown & Johnstone, 1993; Mihaljević Djigunović & Vilke, 2000; Nikolov & Józsa, 2006). Assessment of the speaking skill is not easy to carry out on larger samples, hence many studies do not include it. Different tasks have been used to test the speaking skills in different studies (see also Nikolov, 2016; Hung, Samuelson & Chen, 2016 in this volume). Thus, Low et al. (1993) used paired interviews, and found that Scottish young learners of French and German showed different rates of progress in speaking.

Harris and Conway (2002) tested the speaking skills of Irish young learners of French, German, Italian and Spanish by means of a complex task which tested both listening and speaking. The speaking part involved responding to the examiner's questions about the pupils themselves and to questions based on a picture of a family having a birthday party at a restaurant. The findings indicated that achievement was connected to the young learners' attitudes and motivation.

Studying the speaking skills of learners of Irish Harris, Forde, Archer, Fhearailé and O'Gorman (2006) designed a complex speaking test which was meant to measure communication, fluency of oral description, vocabulary, control of the morphology of verbs, prepositions, qualifiers and nouns, and syntax of statements in speaking. The communication component consisted of question and reply sequences which resulted in the pupil's telling the examiner about their life, and of role-plays carried out by pairs of pupils.

Medved Krajnović (2007) tested the speaking skills of Croatian year 8 (age 13–14 years) and year 12 (age 17–18 years) learners of English as L2 using a set of tests developed in Hungary (Fekete, Major & Nikolov, 1999). In case of the year 8 participants these included first answering a set of personal questions, then describing a picture and relating it to a personal experience, followed by role-playing (with the examiner) three different age-appropriate life situations. In case of the year 12 participants, the third task was replaced by a different one: the participants were presented with five statements on which people had different opinions, then, they had to choose one and offer four reasons why they thought people agreed or disagreed with the statement. All oral performances were assessed along four criteria: task achievement, vocabulary, accuracy and fluency. Both subsamples scored lower on accuracy than on the other dimensions. Positive attitudes and motivation were found to correlate with the oral performance of all the participants.

Hoti, Heinzmann and Müller (2009) designed a similar speaking test for their 3rd grade learners of English as L2 in Switzerland: the first part included personal questions to the pupil and picture description, while the second part involved role-playing as a speaking task performed by two pupils. The authors analyzed the young participants' oral production taking into account task fulfillment, the participants' interaction strategies, complexity of the utterances produced and vocabulary range. Their findings indicated that whereas the third graders' attitudes proved to be a significant explanatory factor of their speaking skills, motivation and self-concept emerged as unimportant in this context.

2 Context of the Present Study

The study described in this chapter was carried out with Croatian young learners of English as L2. A long tradition of early learning of foreign languages is characteristic of the Croatian context. The foreign language has been the compulsory part of the Croatian primary curriculum for more than seven decades now (Vilke, 2007). For years the starting point was grade 5 (age 10–11 years), then grade 4, and since 2003 it has been the beginning of primary education, that is grade 1 (age 6–7 years). English, French, German and Italian have traditionally been offered. Recently the most popular choice has, like in many other contexts, been English. Thus, estimations indicate that over 85 % of first graders learn English, over 10 % German, while French and Italian are present in very small numbers (Medved Krajnović & Letica Krevelj, 2009). Those young learners who start with a language other than English are required to take it from grade 4, so no learner exits primary school without having had English classes (*National Framework Curriculum*, 2001).

Exposure to English is currently extensive, especially through the media (e.g., undubbed TV programmes with subtitles). Croatian users of English have a lot of opportunity to use it with foreign visitors (e.g., business people or tourists) and can often hear or see English expressions (e.g., advertisements in shopping malls).

3 A Study on the Relationship of Young L2 Learners' Motivation and Self-concept with Performance on Speaking Tests

The study is part of the Croatian national research project entitled *Acquiring English from an early age: Analysis of learner language* (2007–2013) (for more details see Mihaljević Djigunović & Medved Krajnović, 2016). The project was sponsored by the Croatian Ministry of Science, Education and Sport. Motivational factors were investigated in a number of earlier projects carried out with Croatian young learners of English (Mihaljević Djigunović, 1993, 1995; Mihaljević Djigunović & Bagarić, 2007; Vilke, 1976, 1982), each time their relevance for language learning achievement being underscored. The Croatian young learners' self-concept was looked into

in a longitudinal study carried out as part of the ELLiE (Early Language Learning in Europe) project (for details see Enever, 2011; www.ellieresearch.eu). The study (Mihaljević Djigunović, 2014) suggested that young learners' self-concept is a complex and dynamic learner characteristic which interacts with other relevant individual as well as contextual factors.

3.1 Aims

In this study we wanted to find answers to the following research questions:

How does motivation of young learners of English change over time?

What trajectory does it follow?

How does self-concept of young learners of English as L2 change over time?

What trajectory does its development follow?

How does young learners' oral production develop over time?

How do motivation and self-concept interact with oral production over time?

3.2 Sample

There were 24 participants included in the study: 12 boys and 12 girls. They were drawn from four primary schools. In terms of their language learning ability they included four high ability, four average and four low ability boys and girls, respectively. The level of ability was estimated by their respective teacher of English. We followed them for 4 years: from grade 5 (age: 11 years) to grade 8 (age: 14 years). They had all started learning English a year before, when they were in grade 4, which means that we studied their motivation, self-concept and oral production from their second to their fifth year of learning English as L2. Their L1 was Croatian.

3.3 Methodology

The instruments used to elicit data on the young learners' motivation and self-concept were taken over from the ELLiE study. The participants' motivation was measured by means of smiley questionnaires and oral interviews. Towards the end of each year they were asked to indicate in the smiley questionnaire how much they liked learning English and how much they liked learning new English words. The latter item was introduced because it had been shown that learning new words can be an important source of motivation in early L2 learning (Szpotowicz, Mihaljević Djigunović & Enever, 2009). In the annual interviews the participants were asked which school subject was their favourite. In earlier projects on early L2 learning in Croatia (Mihaljević Djigunović, 1993, 1995; Vilke, 1982) it was found that such a

question elicited valuable information about young L2 learners' motivation. Scores on these three items were aggregated to compute a single motivational variable.

Information about the participants' self-concept was elicited by the following items asked each year in the oral interview: 'Compare yourself to your classmates. Do you think you are just as good at English as your classmates, or worse, or better than they are?' Some participants decided on the answer right away, but some found it difficult to decide: they claimed to be better at some aspects of learning English (e.g., learning vocabulary) but not at others (e.g., learning grammar).

The speaking tests consisted of two parts: a picture description and a personalized interview. We were interested in seeing whether young learners' oral production based on visual stimuli differs from their production during free conversation. The grade 5 picture description task comprised two pictures. One presented a family house, its different rooms with furniture and various objects such as a computer, a TV set, a bath tub, a kitchen table, and toys. In one of the rooms a boy was playing a computer game, and in another a woman was reading a book. In the other picture a park was shown where children and adults were walking, eating ice-cream and looking at animals such as a lion, a giraffe, and a bear. After the participants described the first picture they were asked about their own home, who they lived with and what their place looked like. After the second picture description they were asked about a park near their home, whether and when they would go to the park, as well as what the park looked like.

In grade 6 the picture description task was based on four pictures depicting the same settings as the two pictures in grade 5, but with more details in terms of the number of objects and people depicted in them. The first two pictures showed the living room and the dining room of a house: family members could be seen eating dinner, watching TV, taking a nap, and studying. The other two pictures were intended to introduce the topic of free time. One showed people around a lake in the countryside engaged in fishing, walking, and sitting on a bench. The other depicted a scene at a beach where people were sunbathing, swimming and enjoying their drinks. The description of first two pictures was followed by an interview about where the participants lived, with whom, what their place looked like and about their eating habits. After the participants finished describing the pictures showing the countryside and the beach, they were prompted to talk about a park near where they lived, if and when they went to the park, what the park looked like, where and how they spent their summer holidays.

In grade 7 the participants were first required to look at a four-part picture of a house depicting four rooms: a bathroom, a living room, a bedroom, and a hall. In each room they could see one or two members of the family doing something. They were instructed to describe everything they could see in the picture. Then, they were asked whether they would like to live in such a house, and what they liked or disliked about it. Following this, the participants were asked to describe their favourite room at home, and to talk about their meals: where they had their meals, who in the family cooked meals, and what they themselves were able to cook.

The grade 8 test required, first, describing a picture of a messy kitchen, where the father was doing the dishes, children were running around and playing, and the

mother could be seen through the kitchen window hanging the washing. The participants were then asked to compare the kitchen in the picture to their own kitchen at home, to say whether they would like to have the kitchen like the one in the picture, as well as to describe their ideal kitchen.

The two parts of each speaking test were assessed separately by two independent raters. Each part of the participant's oral production was assessed along the following four criteria: task achievement, vocabulary, accuracy and fluency. A maximum of five points could be assigned per criterion. The points were determined on the basis of the extent to which the participant met the national curricular targets for each grade.

3.4 Results and Discussion

Below we first present results concerning the individual learner differences we measured. This is followed by presentation of the participants' performance on the oral tests. Finally, we will display the interactions between individual differences and achievements on the tests.

3.4.1 Motivation and L2 Self-concept

As can be seen in Fig. 1, the young participants' motivation displayed variability during the 4 years. It showed a downward trend from grade 5 to grade 7, with a particularly noticeable drop after grade 6, and then increased again in grade 8.

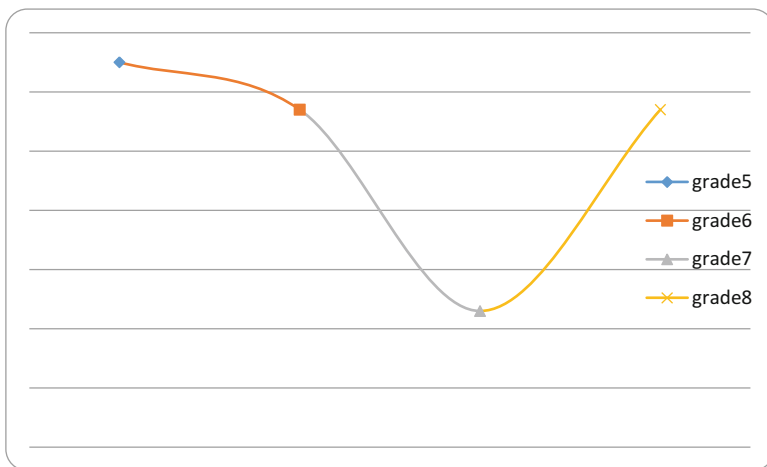


Fig. 1 Developmental trajectory of young learners' motivation over the 4 years

Most studies on motivation suggest that motivation is not a stable variable. With young learners it is usually intrinsic at the start and connected with motivating classroom activities (Nikolov, 2002, 2009b) and the teacher (Nikolov, 1999; Vilke, 1995). Low levels of motivation are usually associated with uninspiring teaching or unfavourable conditions in which L2 is taught (Mihaljević Djigunović, 2009). With increasing length of learning the classroom seems to turn less inspiring for young learners. There may be multiple reasons for this. In contexts where there is high exposure to L2, learners may find it hard to connect the L2 they are learning in school with what they are exposed to outside school. Unfortunately, many teachers fail to integrate the L2 knowledge which their learners bring to the L2 classroom. It is possible as well that learners' interest switches to the new subjects which are introduced in later grades, as is the case with the Croatian curriculum for grades 6 and 7. Also, during the early teens young learners enter puberty and have to deal with new challenges. The rise in motivation from grade 7 in our sample may be the result of the young learners getting more mature and realising the value of knowing English. Their motivation may be getting more instrumentally oriented and may reflect awareness that all school marks are important for their entry into secondary education (which, in Croatia, takes place after grade 8).

As far as the participants' self-concept is concerned, its developmental trajectory was different from that of motivation. Their self-concept peaked in grade 6, and then steadily decreased (see Fig. 2).

If we take into consideration Pinter's (2011) observation that young learners can be considered 'learning optimists', it seems that the young learners in this study became more realistic after grade 6. Teacher feedback, marks in English as well as comparison with classmates probably influenced their self-perception during the fourth year of learning English. It is interesting that, although self-concept is generally thought to be a good predictor of motivation, in this study the trajectories of these two learner characteristics are different.

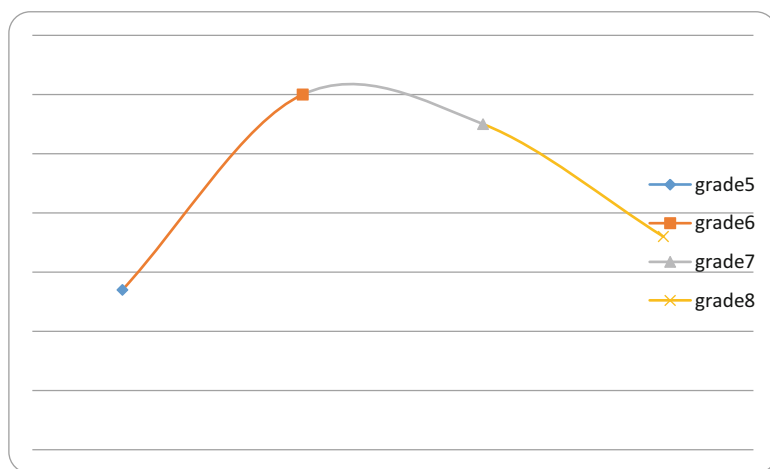


Fig. 2 Developmental trajectory of young learners' self-concept over the 4 years

3.4.2 Oral Performance in Grades 5–8

As Fig. 3 indicates, the participants' overall oral performance was lowest in grade 5 and highest in grade 6. After grade 6 it slowly decreased during grade 7 and remained at more or less the same level in grade 8. It is interesting to observe that changes in oral performance seem to follow the self-concept developmental pattern, which suggests that their self-concept was realistic.

The lowest performance in grade 5 can perhaps be assigned to less experience in describing pictures the participants were presented with for the first time. It is very likely that they had practised describing pictures in class only after they were gradually familiarized with the relevant structures and vocabulary through guided classroom activities.

Besides the overall scores on the speaking tests in each grade, we looked into how the young participants scored on each of the four criteria (task achievement, vocabulary, accuracy and fluency) in the two subtasks (picture description and personal conversation) taken together and taken separately each year.

3.4.2.1 Task Achievement

As can be seen in Figs. 4 and 5, task achievement was quite high over the 4 years. In fact, there was only one participant who did not manage to complete the two subtasks (continually for 3 out of the 4 years). These results confirm that the young learners in this study were generally able to engage in communication in English at the level set out in the national curriculum. It is interesting to note that in

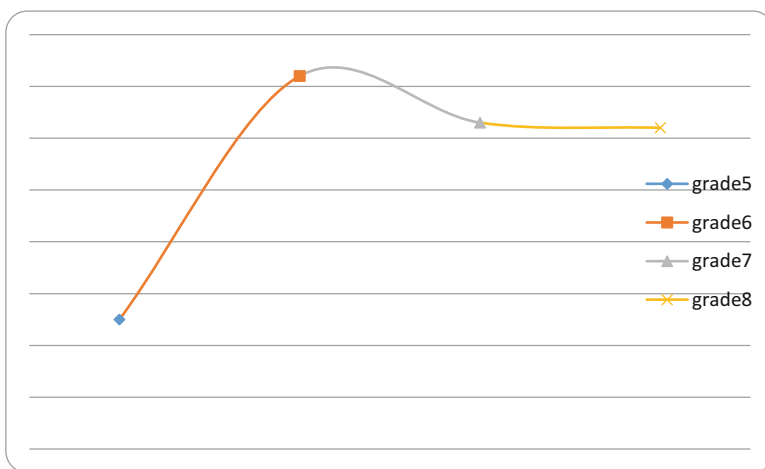


Fig. 3 Young learners' overall oral performance over the 4 years

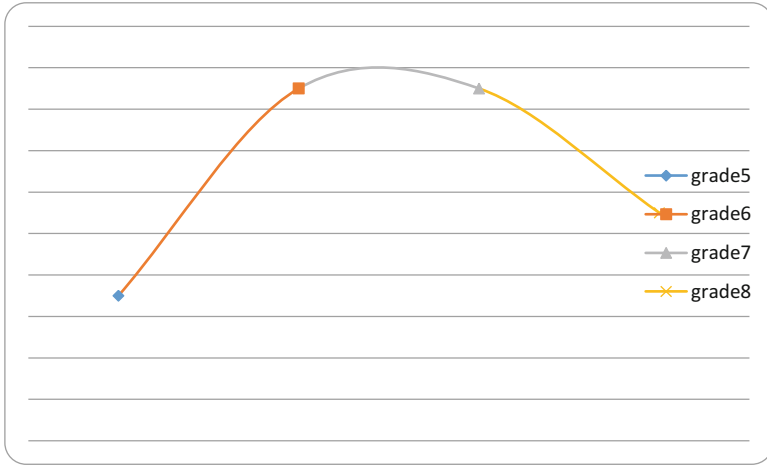


Fig. 4 Scores on task achievement over the 4 years

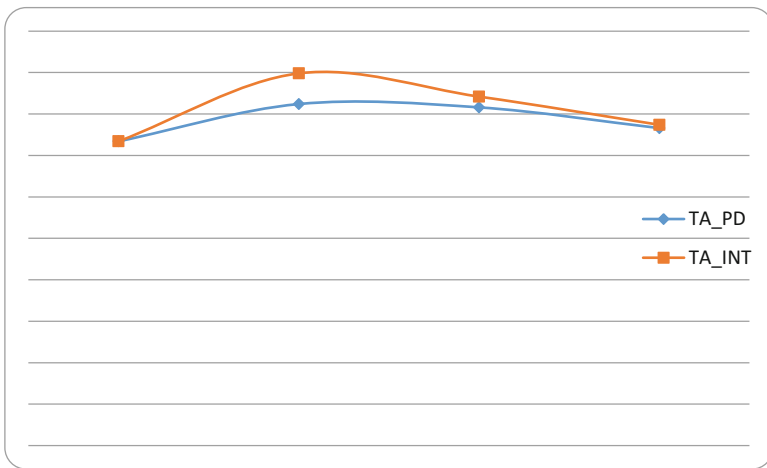


Fig. 5 Scores on task achievement separately for the two subtasks. TA task achievement, PD picture description, INT interview

grades 6 and 7 task achievement was higher in the personalized interview than in the picture description task. It is possible that the participants were more eager to talk about themselves in those grades. It could also be assumed that free conversation in the interview subtask was facilitated by the preceding practice in oral production in picture description.

3.4.2.2 Vocabulary

The overall vocabulary range (Fig. 6) was also good. Interestingly, in grades 5 and 8 it was higher in the personalized interview part than in the picture description task. A possible explanation may be that the questions asked in these grades were more stimulating in terms of vocabulary range (Fig. 7).

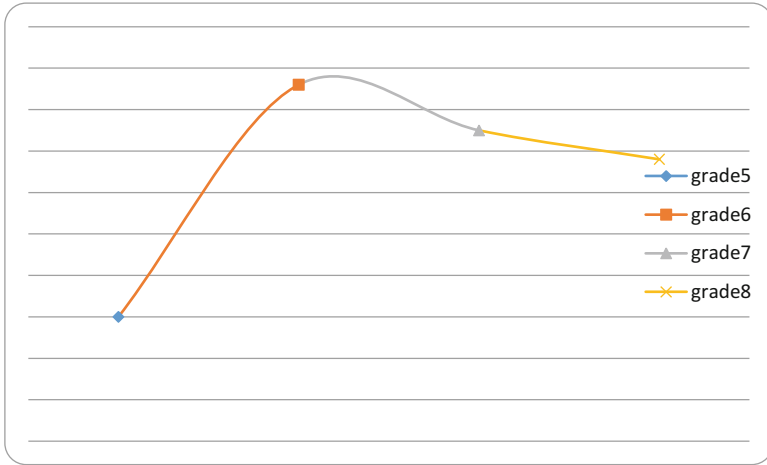


Fig. 6 Scores on vocabulary over the 4 years

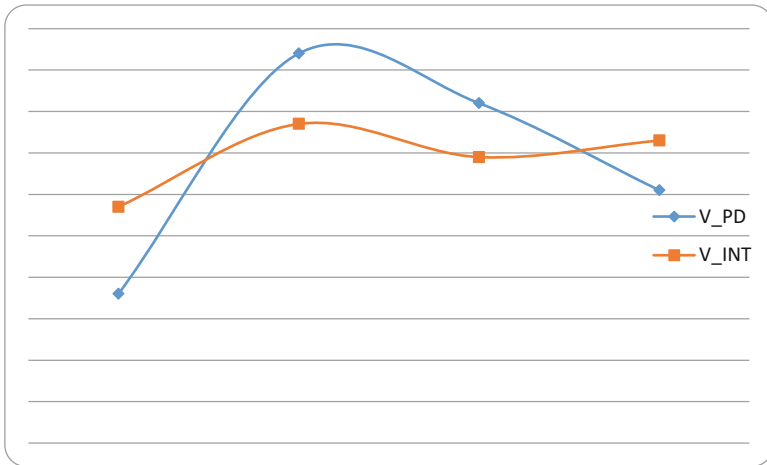


Fig. 7 Separate scores on vocabulary for the two tasks. V_PD vocabulary in picture description, V_INT vocabulary in personalized interview

3.4.2.3 Accuracy

As can be seen in Fig. 8, progression is evident in overall accuracy in all years except in grade 7, when it was slightly lower than in the previous year. Again, contrary to expectations, it was not consistently lower in the interviews; in fact the scores on accuracy were lower in the picture description tasks in grades 6 and 7 (Fig. 9). We tend to think that in these cases picture description served as a kind of speaking practice or warm up activity which perhaps led to lower anxiety and resulted in the participants' more accurate production in the interviews.

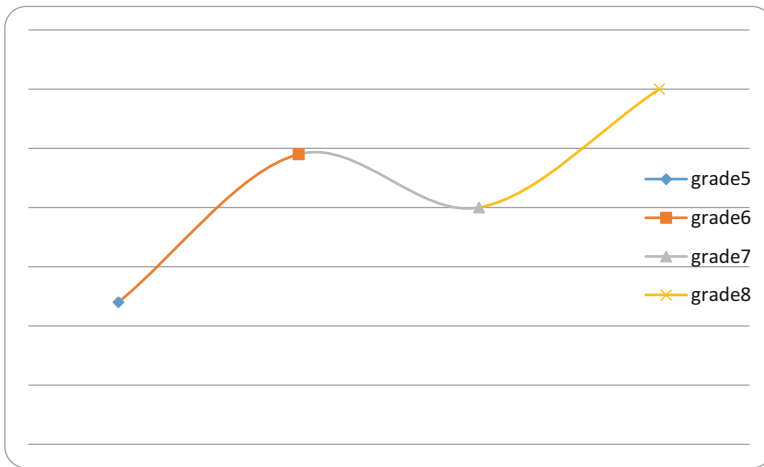


Fig. 8 Scores on accuracy over the 4 years

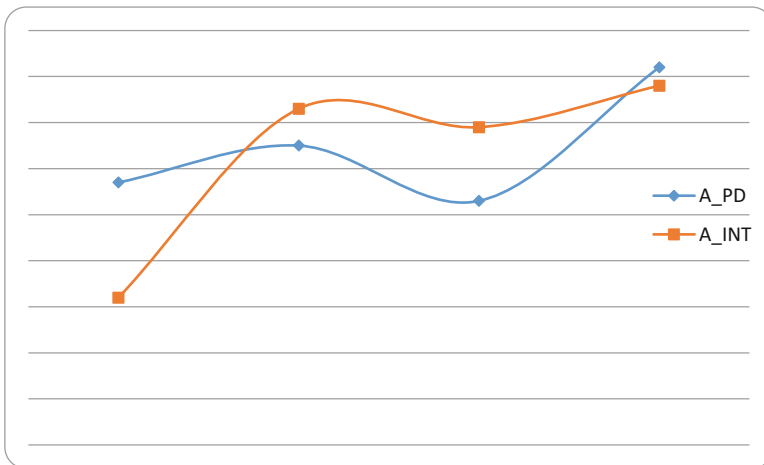


Fig. 9 Scores on vocabulary separately for the two tasks. A_PD accuracy in picture description, A_INT accuracy in personalized interview

3.4.2.4 Fluency

Overall fluency is the dimension of the speaking skills that showed the most consistent development over the 4 years (Fig. 10). It progressed in parallel in the two subtasks. It may be assumed that fluency increases with speaking practice (Fig. 11).

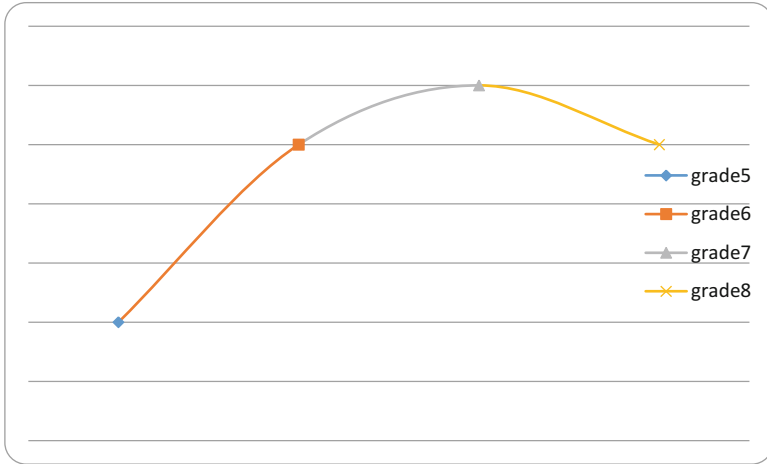


Fig. 10 Scores on fluency over the 4 years

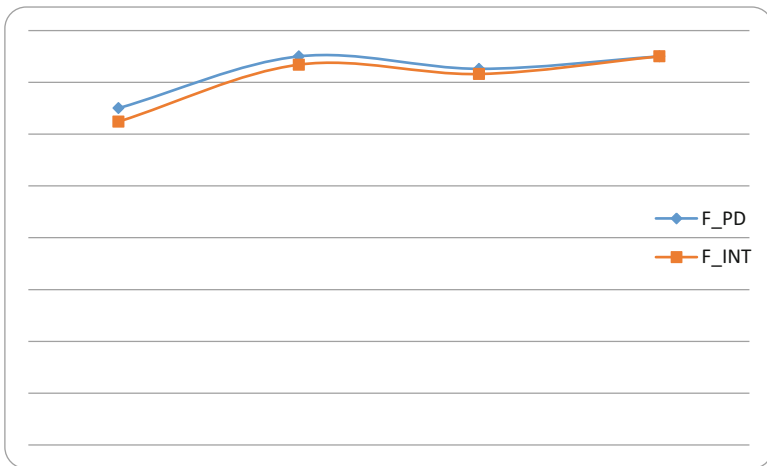


Fig. 11 Scores on fluency separately for the two subtasks. *F_PD* fluency in picture description, *A_INT* fluency in personalized interview

3.4.3 Interaction of Motivation and L2 Self-concept with Oral Performance

Similarly to Hoti et al. (2009), motivation in this study did not emerge as a significant factor in explaining the development of speaking skills of the young learners. There were no significant correlations between motivation and oral performance on the speaking tests.

However, contrary to the Swiss study, which found that self-concept was not an important factor in terms of explaining 3rd graders' speaking skills, in our study the young learners' L2 self-concept proved to be important. Many of the correlations we computed were found to be statistically significant. The strongest correlation between overall oral performance and self-concept was found in grade 5 ($r = .693$, $p = .001$), and it was also significant in grade 6 ($r = .450$, $p = .046$) and grade 7 ($r = .498$, $p = .038$). In grade 8, however, the correlation was not significant ($r = .254$, $p = .293$). This shows that the relationships weakened over the years.

In Tables 1, 2, 3, and 4 below we show the correlations of the participants' self-concept in each grade with the four criteria along which we assessed the oral performances in the respective years.

Table 1 Correlations of self-concept and four assessment criteria for grade 5

		Task achievement grade 5	Vocabulary grade 5	Accuracy grade 5	Fluency grade 5
Self-concept grade 5	Pearson correlation	.665**	.674**	.627**	.711**
	Significance	.003	.002	.005	.001

**Correlation is significant at the 0.01 level (2-tailed)

Table 2 Correlations of self-concept and four assessment criteria for grade 6

		Task achievement grade 6	Vocabulary grade 6	Accuracy grade 5	Fluency grade 6
Self-concept grade 6	Pearson correlation	.589**	.465*	.376	.518*
	Significance	.006	.039	.102	.019

**Correlation is significant at the 0.01 level (2-tailed)

*Correlation is significant at the 0.05 level (2-tailed)

Table 3 Correlations of self-concept and four assessment criteria for grade 7

		Task achievement grade 7	Vocabulary grade 7	Accuracy grade 7	Fluency grade 7
Self-concept grade 7	Pearson correlation	.586**	.472*	.394	.472*
	Significance	.007	.036	.085	.036

**Correlation is significant at the 0.01 level (2-tailed)

*Correlation is significant at the 0.05 level (2-tailed)

Table 4 Correlations of self-concept and four assessment criteria for grade 8

		Task achievement grade 8	Vocabulary grade 8	Accuracy grade 8	Fluency grade 8
Self-concept grade 8	Pearson correlation	.040	.307	.415	.307
	Significance	.870	.200	.077	.200

As can be seen in Table 1, all the correlations were statistically significant, with L2 self-concept being the most strongly associated with fluency. In grade 6 (Table 2) all correlations were significant except the one with accuracy. It is interesting to observe that these significant coefficients were lower than those in grade 5. The following year the pattern was similar: only the correlation with accuracy was non-significant (Table 3). In the final year (Table 4) no significant correlations were established with any of the four criteria.

The correlational analyses suggest that self-concept is more important in earlier than in later years, and learners seem to associate their self-concept more with task-achievement and fluency than with the other two criteria. We assume other individual learner factors (e.g., willingness to communicate, anxiety) emerge in later years as more relevant, and cancel out the linear relationship between L2 self-concept and oral performance.

4 Conclusions

The findings of the study described above offer, first of all, further evidence that young learners' motivation and self-concept are unstable affective learner variables, and that their oral production is also characterised by inter- as well as intra-variability as they progress from year to year. The interaction these variables enter are dynamic, too. Contrary to most previous research we found that L2 achievement as reflected in learners' oral production need not be related to motivation as conceptualised in this study. It seems that it might be useful to define motivation of young L2 learners at more specific levels than is usually done. Perhaps it would be more revealing if task motivation was used as a measure when looking into interaction of motivation with speaking skills of L2 learners aged between 9 and 14 years. The relevance of L2 self-concept comes as no surprise, but it seems worth noting that its interaction with speaking skills is not linear, but more complex and dynamic than so far assumed. Our findings suggest that L2 self-concept is more strongly associated with the quality of oral performance during earlier years than later, and that the accuracy dimension of oral production is the first to show non-significant relationships between the two variables.

5 Limitations of the Study and Future Directions

The findings of this study are based on a rather small sample and do not allow us to venture more definite conclusions. Before making generalisations about the relationship of motivation and self-concept with oral performance of young L2 learners our findings should be verified on a larger sample. In future research it would be useful to examine how the relationships we found are impacted by classroom practices; how teachers value, for example, fluency over accuracy in their feedback, or how peers react to one another. Including other measures of motivation might prove useful too. Perhaps it might be good to also include other individual learner factors which may be relevant for the development of speaking skills, such as willingness to communicate or language anxiety. Comparing young learners' achievement in the other language skills may also be revealing and could be a fruitful focus in future research.

6 Implications for Practice

Classroom teachers can benefit from the insights presented in this chapter in a number of ways. The evidence the study offers of the dynamics of young learners' motivation and self-concept during the primary years can help teachers raise their awareness of how their learners feel and, as a result, understand better their language learning behaviour. The speaking skill is complex and hard to master and requires a lot of time and effort on the part of both teachers and learners. Based on the findings of the current study, teachers may try to design classroom activities that would be more aligned with their learners' affective needs. By doing that teaching may become more inspiring and offer the scaffolding young learners may need at different points during their early years of learning English.

References

- Alexiou, T. (2009). Young learners' cognitive skills and their role in foreign language vocabulary learning. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 46–61). Berlin/New York: Mouton de Gruyter.
- Arnold, J. (2007). Self-concept as part of the affective domain in language learning. In F. Rubio (Ed.), *Self-esteem in foreign language learning* (pp. 13–29). Newcastle, UK: Cambridge Scholars Publishing.
- Bagarić, V. (2007). English and German learners' level of communicative competence in writing and speaking. *Metodika*, 14, 239–257.
- Bernaus, M., Cenoz, J., Espí, M. J., & Lindsay, D. (1994). Evaluación del aprendizaje del inglés en niños de cuatro años: Influencias de las actitudes de los padres, profesores y tutores. [Assessment of EFL learning in four-year-old children. Impact of teacher, parent and guardian attitudes]. *APAC of News*, 20, 6–9.

- Damon, W., & Hart, D. (1988). *Self-understanding in childhood and adolescence*. New York: Cambridge University Press.
- Dörnyei, Z., Csizér, K., & Németh, N. (2006). *Motivation, language attitudes, and globalisation: A Hungarian perspective*. Clevedon, UK: Multilingual Matters.
- Enever, J. (Ed.). (2011). *ELLiE: Early language learning in Europe*. London, UK: The British Council.
- Fekete, H., Major, É., & Nikolov, M. (Eds.). (1999). *English language education in Hungary: A baseline study*. Budapest, Hungary: British Council Hungary.
- García Mayo, M. P., & García Lecumberi, M. L. (Eds.). (2003). *Age and the acquisition of English as a foreign language*. Clevedon, UK: Multilingual Matters.
- Graham, S. (2004). Giving up on modern foreign languages? Students' perceptions of learning French. *The Modern Language Journal*, 88, 171–191.
- 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, NY: Mouton de Gruyter.
- Harris, J., & Conway, M. (2002). *Modern languages in Irish primary schools. An evaluation of the national pilot projects*. Dublin, Ireland: Institiúid Teangeolaíochta Éireann.
- Harris, J., Forde, P., Archer, P., Fhearail, S. N., & O'Gorman, M. (2006). *Irish in primary schools. Long-term national trend in achievement*. Dublin, Ireland: Department of Education and Science.
- Harter, S. (2006). The self. In W. Damon, R. M. Lerner, & N. Eisenberg (Eds.), *Handbook of child psychology. Vol. 3. Social, emotional, and personality development* (pp. 505–570). New York: Wiley.
- Heinzmann, S. (2013). *Young language learners' motivation and attitudes*. London: Bloomsbury Academic.
- Hung, Y.-J., Samuelson, B. L., & Chen, S.-C. (2016). The relationships 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*. New York: Springer.
- Julkunen, K. (1994). Gender differences in students' situation- and task-specific foreign language learning motivation. In S. Tella (Ed.), *Näytön paikka opetuksen kulttuurin arvioiti [Evaluating the culture of teaching]* (pp. 171–180). Helsinki, Finland: University of Helsinki, Department of Teacher Education.
- 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. 253–276). Berlin/New York: Mouton de Gruyter.
- Kiss, C., & Nikolov, M. (2005). Preparing, piloting and validating an instrument to measure young learners' aptitude. *Language Learning*, 55(1), 99–150.
- Kolb, A. (2007). How languages are learnt: Primary children's language learning beliefs. *Innovation in Language Learning*, 2(1), 227–241.
- Kubaneck-German, A. (2003). *Frühes intensiviertes Fremdsprachenlernen. Bericht zur wissenschaftlichen Begleitung eines Modellprojekts des Kultursministeriums des Freistaates Sachsen [An intensified programme for teaching modern languages to children by the Saxon Ministry of Education. Research Report]*. Braunschweig/Dresden.
- Lan, R., & Oxford, R. (2003). Language learning profiles of elementary school students in Taiwan. *IRAL*, 41, 339–379.
- Low, L., Brown, S., Johnstone, R., & Pirrie, A. (1995). *Foreign languages in primary schools: Evaluations of the Scottish pilot projects 1993-1995. Final report*. Stirling, Scotland: Scottish CILT.
- Low, L., Duffield, J., Brown, S., & Johnstone, R. (1993). *Evaluating foreign languages in primary schools*. Stirling, Scotland: Scottish CILT.
- Masgoret, A., Bernaus, M., & Gardner, R. C. (2001). Examining the role of attitudes and motivation outside of the formal classroom: A test of the mini-AMTB for children. In Z. Dörnyei &

- R. Schmidt (Eds.), *Motivation and second language acquisition* (pp. 281–295). Honolulu, HI: The University of Hawaii Second Language Teaching and Curriculum Center.
- Masgoret, A., & Gardner, R. C. (2003). Attitudes, motivation, and second language learning: A meta-analysis of studies conducted by Gardner and associates. *Language Learning*, 53, 123–163.
- Medved Krajnović, M. (2007). Kako hrvatski učenici govore engleski?/How well do Croatian learners speak English? *Metodika*, 14, 173–190.
- Medved Krajnović, M., & Letica Krevelj, S. (2009). Učenje stranih jezika u Hrvatskoj: politika, znanost i javnost [Foreign language learning in Croatia: policy, research and the public]. In J. Granić (Ed.), *Jezična politika i jezična stvarnost* [Language policy and language reality] (pp. 598–607). Zagreb, Croatia: Croatian Association of Applied Linguistics.
- Mercer, S. (2011). *Towards an understanding of language learner self-concept*. Dordrecht/Heidelberg/London/New York: Springer.
- Mihaljević Djigunović, J. (1993). Investigation of attitudes and motivation in early foreign language learning. In M. Vilke & Y. Vrhovac (Eds.), *Children and foreign languages I* (pp. 45–71). Zagreb, Croatia: Faculty of Philosophy.
- Mihaljević Djigunović, J. (1995). Attitudes of young foreign language learners: A follow-up study. In M. Vilke & Y. Vrhovac (Eds.), *Children and foreign languages II* (pp. 16–33). Zagreb, Croatia: Faculty of Philosophy.
- Mihaljević Djigunović, J. (2002). Language learning strategies and young learners. In B. Voss & E. Stahlheber (Eds.), *Fremdsprachen auf dem Prüfstand. Innovation-Qualität-Evaluation* (pp. 121–127). Berlin: Pädagogische Zeitschriftenverlang.
- Mihaljević Djigunović, J. (2009). Individual differences in early language programmes. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 199–225). Berlin/New York: Mouton de Gruyter.
- Mihaljević Djigunović, J. (2014). Developmental and interactional aspects of young EFL learners' self-concept. In J. Horváth & P. Medgyes (Eds.), *Studies in honour of Marianne Nikolov* (pp. 53–72). Pécs, Hungary: Lingua Franca Csoport.
- Mihaljević Djigunović, J. (2015). Individual differences among young EFL learners: Age- or proficiency-related? A look from the affective learner factors perspective. In J. Mihaljević Djigunović & M. Medved Krajnović (Eds.), *Early learning and teaching of English. New dynamics of primary English* (pp. 10–36). Bristol, UK: Multilingual Matters.
- Mihaljević Djigunović, J., & Bagarić, V. (2007). A comparative study of attitudes and motivation of Croatian learners of English and German. *Studia Romanica et Anglica Zagrebiensia*, 52, 259–281.
- 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: The British Council.
- Mihaljević Djigunović, J., & Medved Krajnović, M. (Eds.). (2015). *Early learning and teaching of English: New dynamics of primary English*. Bristol, UK: Multilingual Matters.
- 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. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon, UK: Multilingual Matters.
- Murphy, V. A. (2014). *Second language learning in the early school years: Trends and contexts*. Oxford: Oxford University Press.
- National Framework Curriculum (2001). Zagreb: Ministry of Science, Education and Sport.
- 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. (2002). *Issues in English language education*. Bern, Switzerland: Peter Lang.

- Nikolov, M. (Ed.). (2009a). *Early learning of modern foreign languages. Processes and outcomes*. Bristol, UK: Multilingual Matters.
- Nikolov, M. (Ed.). (2009b). *The age factor and early language learning*. Berlin/New York: Mouton de Gruyter.
- Nikolov, M. (2016). 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*. New York: Springer.
- Nikolov, M., & Józsa, K. (2006). Relationships between language achievements in English and German and classroom-related variables. In M. Nikolov & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 197–224). Pécs, Hungary: Lingua Franca Csoport, PTE.
- Pinter, A. (2011). *Children learning second languages*. Basingstoke, UK: Palgrave Macmillan.
- Šamo, R. (2009). The age factor and L2 reading strategies. In M. Nikolov (Ed.), *Early learning of modern foreign languages: Processes and outcomes* (pp. 121–131). Bristol, UK: Multilingual Matters.
- Seebauer, R. (1996). Fremdsprachliche Kompetenzen und Handlungskompetenzen von Grundschullehrern. Empirische Evidenz und Neuorientierung. [Linguistic competence in the foreign language and work competence of primary school children. Empirical evidence and new directions]. *Praxis des neusprachlichen Unterrichts*, 43(1), 81–89.
- Szpotowicz, M., Mihajevic Djigunovic, J., & Enever, J. (2009). Early language learning in Europe: A multinational, longitudinal study. In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 141–147). Reading, UK: Garnet Publishing Ltd.
- Tragant, E., & Muñoz, C. (2000). La motivación y su relación con la edad en un contexto escolar de aprendizaje de una lengua extranjera. [Motivation and its relationship to age in language learning in the school context]. In C. Muñoz (Ed.), *Segundas lenguas. Adquisición en el Aula* (pp. 81–105). Barcelona, Spain: Ariel.
- Tragant, E., & Victori, M. (2006). Reported strategy use and age. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 208–236). Clevedon, UK: Multilingual Matters.
- Vilke, M. (1976). The age factor in the acquisition of foreign languages. *Rassegna Italiana di Linguistica Applicata*, 3, 179–190.
- Vilke, M. (1982). Why start early? In R. Freudenstein (Ed.), *Teaching foreign languages to the very young* (2nd ed., pp. 12–28). Oxford: Pergamon Press.
- Vilke, M. (1995). Children and foreign languages in Croatian primary schools: Four years of a project. In M. Vilke & Y. Vrhovac (Eds.), *Children and foreign languages II* (pp. 1–16). Zagreb, Hungary: University of Zagreb.
- Vilke, M. (2007). English in Croatia: A glimpse into past, present and future. *Metodika*, 8(14), 17–24.
- Wenden, A. (1999). An introduction to metacognitive knowledge and beliefs in language learning: Beyond the basics [Special Issue]. *System*, 27, 435–441.

The Role of Individual Differences in the Development of Listening Comprehension in the Early Stages of Language Learning

Éva Bacsa and Csaba Csíkos

Abstract This chapter discusses the results of a longitudinal project examining the development of listening comprehension and the role of individual differences in this process in an early language learning context. We aimed at exploring how language learning aptitude, motivation, attitudes, the use of listening strategies, beliefs about language learning and listening anxiety as decisive variables of individual differences (Dörnyei, *AILA Rev* 19:42–68, 2006; *Lang Learn* 59(1):230–248, 2009; Mihaljević Djigunović, Role of affective factors in the development of productive skills. In: Nikolov M, Horváth J (eds) *UPRT 2006: empirical studies in English applied linguistics*. Lingua Franca Csoport, Pécs, pp 9–23, 2006; Individual differences in early language programmes. In: Nikolov M (ed) *The age factor and early language learning*. Mouton de Gruyter, Berlin, pp 198–223, 2009) relate to each other and to the learners' performances on listening measures. The main objective of the present study is to explore and identify the internal structure, roles and relationships of individual variables in the development of early language learners' listening comprehension based on a multi-factor dynamic model of language learning (Gardner & MacIntyre, *Lang Teach* 26:1–11, 1993) and its reinterpretation (Dörnyei, *The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective*. In: Macaro E (ed) *Continuum companion to second language acquisition*. Continuum, London, pp 247–267, 2010).

A total of 150 fifth and sixth graders (11–12-year-olds; 79 boys and 71 girls) of ten school classes in Hungary participated in the research. The findings are in line with the predictions of the theoretical framework: the variables of individual differences are themselves multifactor constructs, the components are in constant

Note: The applied instruments are available at: <http://www.doktori.hu/index.php?menuid=193&vid=13859>

For more information, please contact the first author: evabacsa2010@gmail.com.

É. Bacsa (✉)

Kiss Bálint Reformed School, Szentes, Hungary

e-mail: evabacsa2010@gmail.com

C. Csíkos

Department of Educational Assessment and Planning, University of Szeged, Szeged, Hungary

e-mail: csikoscs@edpsy.u-szeged.hu

interaction with each other and with their environment, thus, changing and creating a complex developmental pattern.

The results of the two phase assessment project clearly indicate that language aptitude defined as one of the main cognitive factors and parents' education are strong predictors of listening performance. The affective factors (e.g., listening anxiety) also contribute to the performance on the listening tests, but their rates change over time and they are sensitive to the context of language learning. Beliefs and emotions are interrelated and they also play a decisive role in the development of listening skills in the early years of language learning. Consequently, what the learners think or believe about language learning and how they feel about it influence the learners' achievement in listening comprehension. In our model, these beliefs are rooted in the students' social background (parents' education) and language aptitude, and this relationship is exactly in contrast with the direction displayed in Gardner and MacIntyre's (Lang Teach 26:1–11, 1993) model.

Keywords EFL • Early language learning • Listening comprehension • Individual differences

1 Introduction

In recent decades, the study of affective factors in second language learning has gained significant ground in addition to the research of cognitive variables, which, according to researchers of the field, could considerably contribute to the understanding and interpretation of individual differences (Dörnyei, 2006, 2009; Gardner, 1985; Gardner & MacIntyre, 1992, 1993; Mihaljević Djigunović, 2006, 2009) The underlying question of the research has been: what might be the main cause of significant variance in the achievement of students from similar backgrounds in similar circumstances. Hence, individual differences became the focus of study in the field originally covering two subfields, language aptitude (e.g., Hasselgren, 2000; Kiss & Nikolov, 2005; Ottó, 2003; Sáfár & Kormos, 2008; Skehan, 1998) and motivation for language learning (e.g., Dörnyei, 1998, 2001; Gardner, 1985; Heitzmann, 2009; Martin, 2009; Nikolov, 2003a). Later on, research on learning styles (Dörnyei & Skehan, 2003) and language learning strategies (e.g., Cohen, 1998; Griffiths, 2003; Mónus, 2004; Nikolov, 2003b; O'Malley & Chamot, 1990; Oxford, 1990; Wenden & Rubin, 1987) also received more attention. Yet, the question remained, what could account for the individual differences where no significant variance is perceived in internal and external circumstances. One possible explanation might be self-perception that fostered the investigation of variables such as attitude to language learning, anxiety, interest and beliefs (e.g., Bacsa, 2012; Brózik-Piniel, 2009; Csíkos & Bacsa, 2011; Csizér, Dörnyei, & Németh, 2004; Dörnyei & Csizér, 2002; Hardy, 2004; Matsuda & Gobel, 2004; Spinath & Spinath, 2005; Tóth, 2008, 2009; Yim, 2014).

It is widely accepted that foreign language proficiency does not solely result from language teaching, but it is the outcome of several factors related to student

achievement. Moreover, the majority of these factors are not static but change dynamically over time. It is also clear that these factors are not independent from one another but they affect learning outcome in interaction with each other (Dörnyei, 2006, 2009, 2010; Gardner & MacIntyre, 1993; Nikolov & Mihaljević Djigunović, 2006, 2011). Research on individual differences in language learning used to study the relationships between single variables and learning outcomes in general. However, recent studies have had a much narrower scope, targeting one skill area. Hence, the subfields of research on motivation, anxiety and learning strategies in reading, writing, listening and speaking skills have been developed (e.g., Goh, 2008; Kormos, 2012; Woodrow, 2006).

In our research we focus on listening comprehension in the early stages of English as a foreign language (EFL) learning. The review of the relevant literature suggests that listening comprehension is a cornerstone of early language learning, since it is based on the processes of first language acquisition, relying primarily on memory, where language input is provided largely through listening (MacWhinney, 2005; Skehan, 1998). The development of listening comprehension is vital to achieving verbal expression and well developed communicative competence, since high level speech production presupposes highly developed listening comprehension (Dunkel, 1986; Mordant & Olson, 2010). In addition, rapidly spreading digital technology redefines language teaching by providing auspicious possibilities in listening to authentic language sources. However, research in the context of the present study found that listening comprehension was one of the most neglected areas of language teaching even though primary school language teaching ought to focus on listening and speaking skills (Bors, Lugossy, & Nikolov, 2001).

The present research is novel in the field of early language learning in that it is the first survey that investigates the development of listening comprehension skills in interaction with the multicomponent construct of individual variables, and applies diagnostic measures of the development of listening comprehension in school context for testing *for* learning purposes in addition to testing *of* learning (Alderson, 2005; McKay, 2006; Sternberg & Grigorenko, 2002).

First, we provide a theoretical background to the survey; then, we describe the methods and the procedure of the research that is followed by the discussion of findings and their theoretical and pedagogical implications.

2 Literature Review

2.1 *Early Foreign Language Learning and Teaching*

Early Language Learning and Young Language Learners appear more and more frequently in the literature of foreign language learning and instruction. Amongst other aspects, research is targeting the specifics of childhood foreign language learning, the optimal time of start and the effective methods of teaching. Having reviewed the relevant literature of the recent years, Nikolov and Mihaljević

Djigunović (2006, 2011) emphasize the importance of further research in the field due to the increased interest in early language learning in Hungary and across the globe. This interest is based on the widespread assumption held not only by researchers that starting language learning early is directly related to its success: “the younger the better”. However, several empirical studies support “the claim that younger learners are more efficient and successful in all respect and at all stages of SLA is hard to sustain in its simple form” (Nikolov, 2000, p. 41; see details in Halle, Hair, Wandner, McNamara, & Chien, 2012; Larson-Hall, 2008; Mihaljević Djigunović, 2010; Moon & Nikolov, 2000; Nikolov, 2009; Nikolov & Curtain, 2000; Nikolov & Mihaljević Djigunović, 2006, 2011).

Researchers agree that young learners’ development significantly differs from that of older children and adults. Krashen (1985) distinguishes language acquisition and language learning. He claims that foreign language acquisition is mainly instinctive, resembling the acquisition of the mother tongue, whereas language learning is a conscious process typical after puberty.

Several models have been constructed to describe language proficiency (e.g., Bachman & Palmer, 1996; Canale & Swain, 1980; CEFR, 2001). In Hungary, the 2003 revision of the Hungarian *National Core Curriculum* (2003) was the first to define the concept of usable language knowledge besides describing the objective of language teaching:

The objective of foreign language learning is to establish communicative linguistic competence. The concept of communicative linguistic competence is identical with usable language knowledge. It means the ability to use adequate language in various communicative situations. Its assessment and evaluation is possible in the four basic language skills (listening comprehension, speaking skills, reading comprehension and writing skills). (p. 38)

Nikolov (2011) outlined the theoretical framework of the assessment and development of English language proficiency for early language learners in grades 1–6, for children between the ages of 6 and 12. She highlighted that the assessment of English language proficiency has to account for language knowledge as a comprehensive and complex construct corresponding to the level of the learners’ knowledge and their age specifics (also see Nikolov, 2016 in this book).

Several studies point out that traditional summative, exam like performance measurements are not appropriate for this age group (Inbar-Lourie & Shohamy, 2009; McKay, 2006). Such tasks are needed that could provide feedback to the teachers and learners about the level of their language development, their strengths and weaknesses, thus outlining the path for successful future development. In other words, assessment *for* learning, conducted by the teachers in the classroom embedded into their daily work of development, is gaining ground in addition to the practice of external evaluation that are mainly targeting accountability, i.e. assessment *of* learning (Lantolf & Poehner, 2011; Nikolov & Szabó, 2011a).

The most important objective of assessment for learning is to positively influence the learning process by scaffolding young learners’ language development in the process of using measurement and feedback. However, assessment must not be restricted to tasks measuring language knowledge, but it has to provide feedback on other domains, like language learning strategies and motivation as they dynamically

influence the process of early language learning (Nikolov & Szabó, 2011a). Assessment can effectively support development only if assessment and development are in a dynamic relationship; these two have to work together a single process for future development (Sternberg & Grigorenko, 2002).

2.2 *Listening Comprehension*

Understanding speech in one's mother tongue seems simple and effortless; however, in a foreign language it involves difficulties, sometimes causing frustration and it is a source of significant stress for many learners (Chang & Read, 2007). Foreign language listening comprehension is an invisible mental process, which is difficult to describe precisely. The learner has to distinguish the sounds, understand vocabulary and grammatical structures, interpret the stress and tone of speech, keep in mind what has been said and interpret what has been heard the socio-cultural context (Vandergrift, 2012). Listening comprehension is rather poorly represented in research on foreign language learning, despite being a crucial skill: it is first acquired in the mother tongue as well as in early language learning.

Research in cognitive psychology revealed that listening comprehension is more than a mere extraction of meaning from the incoming verbal text. It was found to be the process in which the speech is getting linked to the lexical knowledge one already acquired (Vandergrift, 2006, 2012). Hence it is obvious that listening comprehension goes beyond the perception and processing of acoustic signals. This skill has been described in various ways in recent models. The currently most widely accepted cognitive psychological approach perceives it to be a hierarchically structured interactive process. The interactive model of Marslen-Wilson and Tyler (1980) is based on the assumption that the recognition of words involves simultaneously bottom-up processes, where information derives from the uttered word itself and top-down processes, where the information is deducted from the contextual triggers (Eysenck & Kean, 2005). Hence, speech recognition can be described as a two directional process; on the one hand, bottom-up, when learners activate their linguistic knowledge (sounds, grammatical rules etc.) to understand the message, on the other hand, top-down, when learners activate their contextual prior knowledge (topic, direct context, text type, cultural information etc.) to understand the message. At the same time, listening comprehension does not only work top-down or bottom-up, but is composed of the interaction of the two processes, since the listener uses both prior contextual and linguistic knowledge to comprehend the message. The rate of activation between these two processes depends on the linguistic knowledge, familiarity with the topic and the objective of the listening task (Vandergrift, 2012). According to Field (2004), the two processes could not be considered alternative to each other, since their relationship is a much more complex interdependency.

In recent decades, communicative and competence-based language teaching has emphasized listening comprehension and its implications for teaching methodology. All methods prioritize listening comprehension, since it is much more fre-

quently used than the other skills. Learners need to spend a significant amount of time listening to speech in the target language and they need to comprehend what they listen to (Mordaut & Olson, 2010).

Dunkel (1986, p. 100) points out that we need to “put the horse (listening comprehension) before the cart (speech production)” in order to achieve a high level of communicative competence. In other words, high level of speech production presupposes a high level of listening comprehension. Hence, the task of language teachers is to present their learners with a wide variety of listening comprehension tasks (also see Wilden & Porsch, 2016 in this volume).

Foreign language listening comprehension is heavily influenced by the level of listening comprehension in the mother tongue. Simon’s (2001) findings revealed a close relationship between achievements of listening comprehension in L1 and in a foreign language. The development of listening comprehension is not self-serving, since well-developed listening comprehension significantly enhances the development of other skills (Richards, 2005; Rost, 2002).

2.3 *Individual Differences*

The field of psychology has focused on two contradictory objectives: to understand the general principles of human behaviour and intellect and to reveal “uniqueness of the individual mind” (Dörnyei, 2006, p. 42). This latter approach has created an independent subsystem, which came to be known as individual differences (IDs) covering all research targeting these aspects. IDs are “dimensions of enduring personal characteristics that are assumed to apply to everybody and on which people differ by degree” (Dörnyei, 2005, p. 4). According to another description, “they concern stable and systematic deviations from a normative blueprint” (Dörnyei, 2006, p. 42). Hence, the objective is to reveal and identify those specific learner characteristics that are relevant in foreign language acquisition and are present to different degrees among learners (Dörnyei, 2006; Mihaljević Djigunović, 2009).

The literature on foreign language acquisition traditionally separates IDs into *cognitive* and *affective* factors (Gardner, 1985; Gardner & MacIntyre, 1992, 1993). According to Gardner and MacIntyre (1992, p. 211), cognitive factors “involve different aspects of cognition”. Johnson (2001, p. 117) defined them as “the mental makeup of a person” that include age, aptitude, intelligence, SES, learning strategies and learning or cognitive style, whereas affective factors include “those attributes that involve individuals’ reactions to any situation” (Gardner & MacIntyre, 1992, p. 211). In other words, they represent “the emotional side of human behavior” (Brown, 1994, p. 135) and include personality factors such as anxiety, extroversion/introversion, inhibition, risk-taking, empathy, self-perception, attitude and motivation (Mattheoudakis & Alexiou, 2009).

Researchers assembled detailed lists of factors of individual differences (e.g., Gardner, 1985; Gardner & MacIntyre, 1993; Larsen-Freeman & Long, 1991; Skehan, 1998). According to Mihaljević Djigunović (2009, p. 198) “the term individual

differences, although widely used, still represents a rather loose concept and different authors list different learner characteristics as individual differences.” She collected the most frequently listed variables in recent publications: (1) intelligence, (2) aptitude, (3) age, (4) gender, (5) attitude and motivation, (6) language anxiety, (7) learning style, (8) learning strategies and (9) willingness to communicate.

Others highlight some significant domains instead of giving extensive lists of individual differences. Dörnyei (2009) mentions four important variables: (1) Motivation refers to the direction and extension of student behaviour, including the choice of the learner, intensity of learning and endurance. (2) Ability of language acquisition refers to the capacity and quality of learning. (3) Learning style includes the way of learning. (4) Learning strategies are located halfway between learning style and motivation, indicating the proactivity of the learner in selecting the learning path. “Thus the composite of these variables has seen to answer why, how long, how hard, how well, how proactively, and in what way the learner engages in the learning process” (p. 232).

Prior research predominantly investigated the learner’s characteristics in the context of individual differences and they were generally included in research as background variables that modify, personalize the picture of the language acquisition process (Dörnyei, 2009). Today several researchers perceive foreign language learning as the result of interaction between learner characteristics and the learning context, assuming a complex relationship between these two factors. In addition, increased efforts are put into a deeper understanding of connections between the learners and the context of learning (Mihaljević Djigunović, 2009). Some IDs are more stable and less sensitive to the changes of circumstances (e.g., intelligence, aptitude), while others (e.g., motivation, strategies, anxiety) respond quickly to changed context (e.g., in training program). The question can be raised whether an optimal combination of individual variables could be identified that would particularly enhance the effectiveness of language learning. According to Ackerman (2003), individual characteristics can strongly influence learning success separately as well, however, any combination of these characteristics would definitely have a larger impact.

Research on IDs further highlights the fact that different variables influence success and student achievement to different degrees. Hence, the traditional approach identifies primary and secondary variables (Gardner & MacIntyre, 1992, 1993). According to this classification, aptitude and motivation can be considered as primary variables in foreign language research, since these variables have the strongest demonstrable impact on student achievement: aptitude is the primary cognitive factor and motivation is the primary affective factor. Others extended this class of primary variables to include aptitude, attitude and motivation, social background, status of the target language and the quality of language teaching (Csapó, 2001; Ellis, 1994; Józsa & Nikolov, 2003, 2005; Nikolov, 2007). According to Dörnyei (2010), the perceived effect of these variables also depends on the method applied to measure these constructs.

Furthermore, some recent investigations question the modular approach to individual variables. Dörnyei (2009, 2010) approaches the role of individual differences,

especially the two primary variables (aptitude and motivation), from the perspective of a “dynamic system”. He claims that “identifying ‘pure’ individual difference factors has only limited value [...]; instead, a potentially more fruitful approach is to focus on certain higher-order combinations of different attributes that act as integrated wholes” (Dörnyei, 2010, p. 267; Dörnyei, MacIntyre, & Henry, 2015).

It has been revealed that young learners do not resemble each other in every aspects of their learning either, hence it is possible as well as desirable to study their IDs (Mihaljević Djigunović, 2009; Nikolov, 2009). However, adequate methods and instruments for assessment are scarce, since the majority of available measures were developed for older age groups. According to Mihaljević Djigunović (2009), the main line of future research should focus on exploring the relationships between IDs among early language learners, which ultimately presupposes the development of relevant measures and methods.

Findings of prior research draw a varied picture about the relationship between IDs and student achievement (also see Mihaljević Djigunović, 2016 in this volume). There has been a consensus that cognitive, affective and additional background factors all impact the success of language learning, however, the significance attributed to individual factors varies across the studies (Csapó & Nikolov, 2009). Consequently, the study of student achievements should out cover a wide range of interactions between individual variables (Nikolov & Mihaljević Djigunović, 2011).

3 The Study

3.1 *A Model of Individual Differences in Listening Comprehension*

When defining the theoretical framework of our research a language learning model had to be found that would meet the requirements of complexity, interactivity and dynamism (flexibility, versatility) in terms of the context and components of language learning. The Socio-educational model of second language acquisition proposed by Gardner and MacIntyre (1993) is one of the most often cited models. It perceives the learning process embedded in a comprehensive socio-cultural context, and highlights four different aspects, related to each other: (1) antecedent factors: e.g., age, gender, prior learning experience and beliefs; (2) ID variables: e.g., intelligence, language aptitude, strategies, attitudes, motivation, anxiety; (3) language learning contexts: formal and informal learning contexts; and (4) outcomes: linguistic and non-linguistic achievements. The model describes the factors influencing language learning as interrelated, exerting direct and indirect impact on the process of language acquisition which effects achievement. The authors note that the model is extendable, since several additional cognitive and affective factors might be present in language learning influencing learning outcome. This model was the first to place emphasis on the interaction of variables, perceiving language learning as a

dynamic process influenced by several interrelated factors. At the same time it is passive (Kim, 2001), since it defines the amount and direction of interactions excluding the possibility of integrating further interactions of variables into the model.

In his review of individual differences Dörnyei (2010) challenges the dichotomy of cognitive and affective factors, stating that the two domains overlap. He interprets IDs as a multifactor “umbrella term”, including several underlying factors. Instead of investigating the interaction and effect of isolated areas, Dörnyei suggests the identification of existing (viable) constellations in which “the cognitive and motivation (and also emotional) subsystems of human mind cooperate in a constructive manner” (Dörnyei, p. 267).

Therefore, our investigation is based on Gardner and MacIntyre’s (1993) socio-cultural framework and its set of variables with the addition of Dörnyei’s (2010) points. Hence variables of IDs were perceived as multi-factor constructs where “the constituent components continuously interact with each other and the environment, thereby changing and causing change, and subsequently displaying highly complex developmental pattern” (Dörnyei, p. 267).

Based on the above and relying on findings of prior research among early stage language learners, we conducted our research in a classroom context. *Age*, *gender* and *parents’ education* were included in the study from a group of antecedent (background) variables (Csapó, 2001; Csapó & Nikolov, 2009; Józsa & Nikolov, 2005; Mattheoudakis & Alexiou, 2009; Nikolov & Curtain, 2000). Additional IDs were represented by variables of *language aptitude*, *strategies of listening comprehension*, *beliefs* related to language learning, *attitude* towards and *motivation* for language learning and *anxiety to listening comprehension* (Bacsa, 2012; Csizér & Dörnyei, 2002; Dörnyei, 2006, 2009; Kiss, 2009; Kiss & Nikolov, 2005; Mihaljević Djigunović, 2009; Nikolov, 2003a, 2003b, 2007, 2009; Nikolov & Mihaljević Djigunović, 2006, 2011; Yim, 2014). The context of language learning (formal vs. informal) appears in the analysis as a background variable. Aspect of achievement was restricted to the results of listening comprehension tests (Nikolov, 2011; Nikolov & Szabó, 2011a, 2011b; Szabó & Nikolov, 2013) and school marks in English. Following Dörnyei (2010), the research interprets the variables involved in the research as multifactor constructs rather than independent modules and attempts to draw conclusions on changes in student achievement factors affecting the development of listening comprehension by exploring the relationships and constellations of these factors.

3.2 Aim of the Study

We aimed to explore and identify the internal structure, roles and relationships of individual variables in the development of early language learners’ listening comprehension based on a multi-factor dynamic model of language learning (Gardner & MacIntyre, 1993) and its reinterpretation (Dörnyei, 2010). A further objective was

to understand the development of young language learners' listening comprehension and the influencing factors of its individual differences along with exploring how these factors affect each other creating a unique pattern in the early language learning context and contributing to listening comprehension achievements. We expected that the research findings would help us understand the development of listening comprehension and IDs as well as explain young learners' achievements and foster the facilitation of developing listening comprehension effectively.

The study addressed the following research questions:

1. What tendencies could be seen in the development of students' listening comprehension over a semester?
2. How do separate components of individual differences change over the assessment period?
3. What relationship (pattern) can be detected between the components of individual differences and how are they related to the students' results in listening comprehension assessments?
4. To what extent do pretest results of individual differences predict posttest achievements?
5. What causal relationship could be found between components of individual differences and student achievements?
6. What relationship (pattern) can be detected between the components of individual differences and how are they related to students' school marks in ESL?
7. To what extent do pretest results of individual differences predict English marks?
8. What causal relationship could be found between components of individual differences and English marks?

4 Method

4.1 Participants

Participants were elementary school students in grade 5 and grade 6. A total of 150 students of EFL were involved in ten school classes of a mid-sized town in Hungary. In order to get results that can be generalized, the sample was representative with regards to gender, ability levels of the student groups and socio-economic status.

4.2 Measures and Procedure

The research design applied the methodologies and measures used in the field and the characteristics of the sample with a preference of mixed methods (Moschener, Anschuetz, Wernke, & Wagener, 2008; Nikolov, 2009; Nunan & Bailey, 2009). (1) *Diagnostic listening comprehension tasks* (Nikolov & Szabó, 2011a, 2011b) were provided for teachers to measure and monitor their students' development of

listening comprehension during the assessment period. (2) *Pretests and posttests* (Nikolov & Józsa, 2006) were applied to measure listening comprehension achievements. Relevant adapted and newly developed questionnaires were used to capture IDs in the following areas: (3) *language aptitude* (Kiss & Nikolov, 2005), (4) *strategies of listening comprehension* (Vandergrift, 2005, 2006), (5) *beliefs about language learning* (Bacsa, 2012), (6) *attitude and motivation related to language learning* (Nikolov, 2003a, 2003b) and (7) *listening anxiety* (Kim, 2005). All the questionnaires applied a 5 point Likert-scale to assess statements. We used (8) *interviews* and (9) *think-aloud protocols* to gain in-depth insight into the functioning of listening comprehension.

The features of the questionnaires and the tests are presented in Tables 1 and 2. A longitudinal design was used covering the period of a semester, involving two measurement sessions (except for language aptitude which was measured once between the two assessment periods). All students were given a booklet including diagnostic tasks of listening comprehension and questionnaires of individual differences. The instruments were administered with the help of classroom teachers, whereas the aptitude and the placement tests were completed under the supervision of the first author. The collected data was analyzed with the help of SPSS 22 and AMOS 20 software.

The development of listening comprehension over the period of 6 months was analyzed in previous papers (Bacsa, 2014; Bacsa & Csíkos, 2013). The specifics of individual differences were identified by detailed investigations of the individual variables, which provided a picture of how these variables influenced student achievement and how they changed between the two testing sessions.

The present study provides a synthesis of the main findings of the longitudinal research on the role of IDs in the development of young language learners' listening

Table 1 Features of the questionnaires applied in the research

Measures of individual differences	Number of items	Number of factors loaded	Cronbach- α pretest	Cronbach- α posttest
MALQ (Vandergrift, 2005)	18	4	0.83	0.84
Attitude and motivation to language learning (Nikolov, 2003a, 2003b)	20	3	0.71	0.83
FLLAS (Kim, 2005)	33	5	0.88	0.92
Beliefs about language learning (Bacsa, 2012)	40	8	0.87	0.91

Table 2 Features of the tests applied in the research

Tests	Number of items	Cronbach- α	Mean (%)	Std. deviation
Language aptitude test (Kiss & Nikolov, 2005)	45	0.81	60.0	15.3
Pretest	16	0.51	56.8	15.6
Posttest (part 1)	16	0.64	62.3	17.6
Posttest (total)	30	0.79	63.8	14.8

comprehension skills. Six variables of individual differences (*aptitude*, *beliefs* about language learning, *strategies* of listening comprehension, *attitude and motivation* toward language learning, *anxiety* about listening comprehension, *parents' education*) and three variables of student achievement (*pretest*, *posttest*, *school marks* in English) were used and their interactions were analyzed, in line with the theoretical framework (Dörnyei, 2010; Gardner & MacIntyre, 1993).

5 Results

5.1 *Development in Listening Comprehension*

The diagnostic tasks used in this research for the first time were welcomed by most teachers and students and they also received positive reviews as measurement instrument. The results of the series of assessments monitoring the development of listening comprehension show that the majority of the sample continuously developed throughout the assessment period.

As far as the reliability of the measures is concerned, the results show that the pretest reliability figures (Cronbach- $\alpha=0.51$) were lower than expected and lower than what was found in prior research (Cronbach- $\alpha=0.72$ in Nikolov & Józsa, 2006), which might partially be explained by the lower item and sample size (Dörnyei, 2007), as well as the lower number of distractors. Therefore, we decided to add validated tests and the modified tests provided sufficient differentiation in the posttest (Cronbach- $\alpha=0.79$).

A significant increase was found in overall listening comprehension over the semester long assessment period ($t=-4.268$; $p<0.001$). Subsamples divided by age and gender did not show significant variance; although, boys achieved somewhat lower scores than girls, as did the grade 5 subsample compared to grade 6, where insignificant difference reoccurred on the post-test as well. Significant inter-group variance was found on the pretest [$F(9.127)=4.90$]; $p<0.001$] and the posttest [$F(9.128)=13.20$]; $p<0.001$] along with a considerable within-group variance.

5.2 *Components of Individual Differences*

IDs were assessed by applying quantitative and qualitative research methods. The questionnaires (Attitude and motivation, Beliefs about language learning) were either originally constructed for the age of the sample or adapted (MALQ and FLLS) to their age specifics, by reproducing the original factor structure to measure the construct reliably. This statement is supported by several findings of the qualitative investigations. The reliability indices of subscales deriving from the internal factor structures of the questionnaires were found to be lower in some cases than expected in social scientific research, hence, only those factors were included in the components of

individual differences (final analysis) which reliably measured the construct (Cronbach- $\alpha > 0.70$). This condition was fulfilled by the scales shown in Table 3.

The first component (Cronbach- $\alpha = 0.70$) is comprised of the strategies used by listeners when concentrating to the task at hand and focusing on understanding English speech. The second component is the factor of foreign language learning motivation and attitude towards school learning and classroom conditions (Cronbach- $\alpha = 0.70$). The third factor (Cronbach- $\alpha = 0.79$) includes statements on the learners' self-concept. The fourth component (Cronbach- $\alpha = 0.82$) refers to feelings, anxiety about focusing attention and following the text, the fifth (Cronbach- $\alpha = 0.72$) to anxiety about the difficulty of comprehension, the sixth (Cronbach- $\alpha = 0.72$) to anxiety about unknown words that hinder comprehension. Finally, the seventh factor (Cronbach- $\alpha = 0.78$) covers beliefs on the difficulty of language learning.

In the first section of the results, descriptive statistical data of the selected components and results of the two assessments are presented (Table 4).

Table 3 Components included in the research synthesis (Cronbach $\alpha > 0.70$)

Individual differences	Example
Strategy: <i>directed attention</i>	"While listening to the text I pay attention to the key words."
Attitude and motivation: <i>classroom level</i>	"English classes are extremely boring."
Attitude and motivation: <i>learner level (self-concept)</i>	"No matter how I study, I cannot achieve better in English."
Anxiety about listening comprehension: <i>following the text</i>	"When a person speaks English very fast, I worry that I might not understand all of it."
Anxiety about listening comprehension: <i>difficulty of comprehension</i>	"When someone pronounces words differently from the way I pronounce them, I find it difficult to understand."
Anxiety about listening comprehension: <i>unknown words</i>	"I get annoyed when I come across words that I do not understand while listening to English."
Beliefs: <i>difficulty of language learning</i>	"I learn English quite easily."

Table 4 Components of individual differences in the two assessments

Components of individual differences	First assessment	Second assessment	t	p
Strategy: <i>directed attention</i>	3.68	3.66	0.268	n.s.
Attitude and motivation: <i>classroom level</i>	3.79	3.58	2.602	0.010
Attitude and motivation: <i>learner level (self-concept)</i>	3.47	3.33	1.973	n.s.
Anxiety about listening comprehension: <i>following the text</i>	2.82	2.91	-1.059	n.s.
Anxiety about listening comprehension: <i>difficulty of comprehension</i>	2.27	2.66	-4.904	0.000
Anxiety about listening comprehension: <i>unknown words</i>	2.68	2.87	-2.049	0.042
Beliefs: <i>difficulty of language learning</i>	3.38	3.45	-1.191	n.s.

The data presented in Table 4 show that the revealed (obtained from MALQ) strategy use (metacognitive awareness) of focusing on keywords and understanding scored high in both assessments without a significant difference. Observed strategy use (think-aloud protocol) confirmed the primary usage of focusing on keywords in the listening process. It can also be seen that, based on the average scores, students do not think that they would have difficulties in learning EFL, since they scored high in both assessments on the related belief scales without significant differences. The students' motivational self-concept (learner level) did not reflect a significant change by the end of the school year. However, attitude and motivation in classroom learning decreased significantly by the second assessment, which might be explained by end-of-year exhaustion or incidental negative experiences. The three components of anxiety about listening comprehension scored below 3.00 on average in both assessments, which indicate that the participants' anxiety levels are rather low. In addition, the interviews revealed that their anxiety relates mostly to the test situation and pressure for achievement rather than to the listening comprehension activity itself. The second assessment showed a significant increase in anxiety in case of two variables; however, the increased level does still not reach "general" anxiety level.

In addition to the seven components of individual differences this study includes the results of the *language aptitude test* and *parents' education* which proved to be the main predictors of foreign language achievements of young learners (Csapó & Nikolov, 2009; Kiss & Nikolov, 2005). We wanted to find out to what degree the nine ID variables explain the variance found in the two assessments. Previous research suggested that aptitude would prove to be the best predictor of foreign language learning achievements (Ellis, 1994; Kiss & Nikolov, 2005; Robinson, 2001; Skehan, 1991; Sparks, Patton, & Ganschow, 2011) and that cognitive variables would explain more of the variance in case of younger learners than in older age groups (Csapó & Nikolov, 2009). The results presented in Table 5 support all these prior research findings in both assessments.

Table 5 shows that the components included in the analysis explain 30 % of the variance in the listening comprehension scores in the initial and 46 % in the second

Table 5 Variables of individual differences explaining listening test performances

Individual differences	Pretest	Posttest
Parents' education	1.4	4.6*
Language aptitude	24.4**	29.6**
Strategy: <i>directed attention</i>	-0.5	3.2
Attitude and motivation: <i>classroom level</i>	1.8	-0.1
Attitude and motivation: <i>learner level (self-concept)</i>	-2.1	-1.4
Anxiety about listening comprehension: <i>following the text</i>	0.4	-1.2
Anxiety about listening comprehension: <i>difficulty of comprehension</i>	1.8	3.3
Anxiety about listening comprehension: <i>unknown words</i>	0.0	5.1
Beliefs: <i>difficulty of language learning</i>	3.3	2.4
<i>Total variance explained (R²)</i>	30 %	46 %

**p < .01; *p < .05

assessment. It can be seen that aptitude accounts for a significant degree of variance in both cases: in the first assessment it gave 80 % of the total explained variance as the only significant factor, whereas in the second assessment it covered 65 % of the total variance explained. In both assessments cognitive factors in the traditional sense (Gardner & MacIntyre, 1992, 1993) explained a higher percentage of variance than affective factors. In the first assessment aptitude was found to be the only significant predictor of listening comprehension results, whereas in the second assessment parents' education also proved to be a significant indicator of student achievement. These findings support the findings of previous research that suggested the primary status of cognitive factors in predicting student achievement in younger age groups (Csapó & Nikolov, 2009; Kiss & Nikolov, 2005). They further indicate that variables of individual differences (e.g., attitude, motivation, strategies, beliefs) cannot be viewed as stable constructs, but they change with time reacting to changes in context (Mihaljević Djigunović, 2009; Robinson, 2001).

5.3 Relationships Between Variables of First Assessment and Listening Comprehension Achievement

In this section the relationship between the components yielded from the two assessment session are explored by attempting to filter the situational effect of the context, thus allowing us to understand young learners' development in their English listening comprehension skills. In the following analyses we studied how the experiences, opinions and beliefs found in the first assessment predicted the development of listening comprehension with the help of the factors outlined above. First, a *cluster analysis* was conducted to see how the certain variables relate to listening comprehension, i.e. what clusters they form around achievement. The dendrogram of the cluster analysis conducted by the *furthest neighbour* method is presented in Fig. 1.

The dendrogram in Fig. 1 reflects four separate clusters. Variables of aptitude and achievement are grouped in a well separated cluster. The other variables link to this by forming smaller individual clusters. Anxiety variables are grouped together, motivation variables are connected to strategies, linking to the cluster formed by beliefs and parents' education. Following the steps based on the proximity of connections it can be seen that aptitude and parents' education are followed by the anxiety components which in turn are followed by beliefs about language learning. Motivation is the last connection to them, supporting the findings that it is the most weakly interacting component with achievement.

Following the system of relationships between the variables, predicting values of individual differences are considered. *Regression analysis* was conducted to reveal these factors. The question was to what degree the independent variables of individual differences (in the first assessment) included in the analysis predicted listening comprehension achievement as dependent variables in the posttest. Table 6 shows the results of the regression analysis.

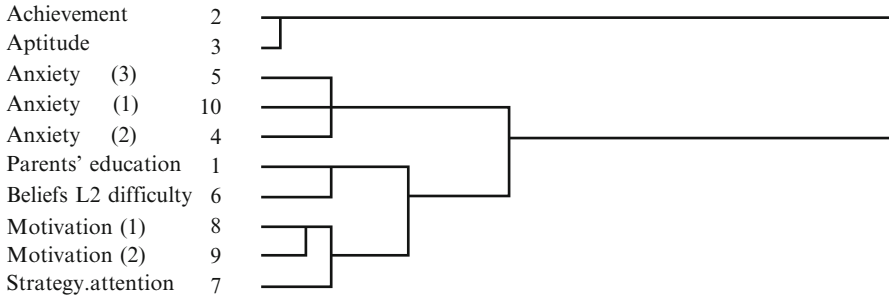


Fig. 1 Dendrogram of variable clusters around listening comprehension achievement. Explanation: *Anxiety (1)*: following the test; *Anxiety (2)*: difficulty of understanding; *Anxiety (3)*: unknown words; *Motivation (1)*: classroom level; *Motivation (2)*: student level (self-concept)

Table 6 Variables of first assessment predicting listening comprehension achievement

Individual differences	β	$r^*\beta$ (%)
Parents' education	0.185**	4.4**
Language aptitude	0.552**	28.3**
Strategy: <i>directed attention</i>	0.041	1.0
Attitude and motivation: <i>classroom level</i>	0.051	0.5
Attitude and motivation: <i>learner level (self-concept)</i>	-0.144	-1.2
Anxiety about listening comprehension: <i>following the text</i>	0.087	0.4
Anxiety about listening comprehension: <i>difficulty of comprehension</i>	-0.264**	5.7**
Anxiety about listening comprehension: <i>unknown words</i>	0.200*	3.2*
Beliefs: <i>difficulty of language learning</i>	0.162*	6.8**
Total variance explained (R^2)		49 %

** $p < .01$; * $p < .05$

Table 6 shows the β values of the regression analysis and the explained variance of variables (R^2). Five out of the nine variables included in the analysis had significant β values. The nine variables in total explained nearly 50 % of the variance found in the posttest. Half of this is explained by aptitude alone. Parents' education representing the learners' socio-economic status was also found to have significant variance, in line with the majority of other studies conducted in this age group in Hungary (e.g., Bukta & Nikolov, 2002; Csapó & Nikolov, 2009; Józsa & Nikolov, 2005). The three additional variables that represent significant explanatory power relate to the thinking and feeling of the students about the difficulties of language learning and listening comprehension.

Finally the paths supposedly leading to listening comprehension achievement were drawn with the help of *path analysis* (Fig. 2). The objective of the path analysis is to reveal the degree and strength of suggested causal relationships (Münnich & Hidegkuti, 2012). The literature (Everitt & Dunn, 1991) suggests drawing the hypothesized path (just-identified/saturated model) prior to conducting the analysis so that the outcome of the analysis may confirm our assumptions. The present anal-

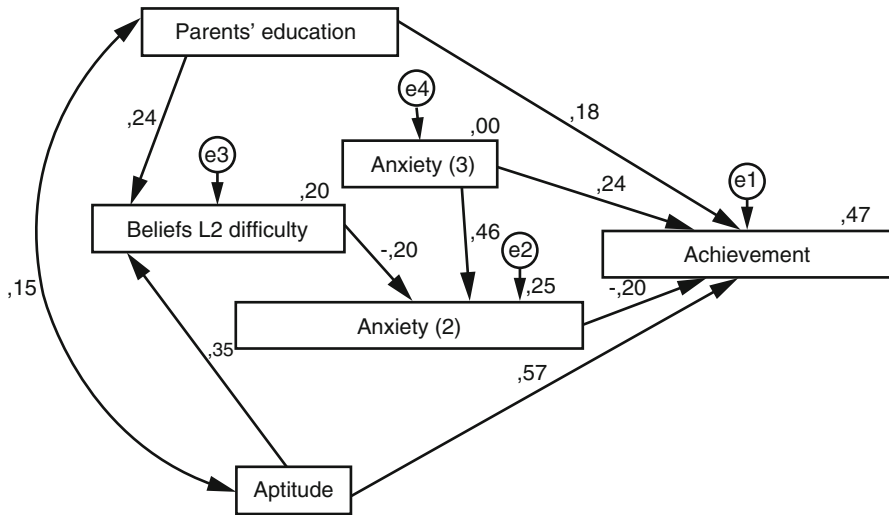


Fig. 2 Variables of individual differences and causal relationships of listening comprehension achievement. *Explanation:* Anxiety (2): difficulty of understanding; Anxiety (3): unknown words

ysis was based on Gardner and MacIntyre’s (1993) model modified by Dörnyei (2010), where individual variables have both direct and indirect effect on test achievement, and as Dörnyei (2010, p. 267) suggests “the cognitive and motivation (and also emotional) subsystems of human mind cooperate in a constructive manner”. Those components of individual variables were included in the path-analysis that resulted in significant β values in the regression analysis. Hence, the final model comprised five variables (exogenous variables) representing the IDs (*aptitude, parents’ education, anxiety about the difficulty of understanding and unknown words and beliefs on the difficulty of language learning*). The interactions and causal relationships of these exogenous variables could explain the development of student achievement (endogenous variable). The path diagram is shown in Fig. 2 below.

The χ^2 -test confirmed our null hypothesis, i.e. the saturated and default models were found to be identical. The parameters were evaluated with the method of *maximum likelihood*, which attempts to maximize the value of the likelihood of the criterion variables.

In this section we describe the indexes of model fit. The saturated model had 27 parameters, the tested model had 21, degrees of freedom (df) was 6 (NPAR). Values of $\chi^2 = 7.95$, $p = 0.242$ indicate that the model fit between the saturated model and the data was not (significantly) worse than between the data and the default model. It can be seen that path coefficients (β values) that are found next to the arrows in the diagram (Fig. 2) are significant in each case. NFI=0.949 and CFI=0.986 values reflect optimal fit, since both indicators exceed the 0.9 (good fit) level. Finally, RMSEA=0.034 value also suggests good model fit: lower than 0.05.

The five variables in the model account for 47 % of the total variance of achievement. The multivariate analysis of individual differences and test achievements

revealed that components of individual variables exert both direct and indirect effect on student achievement. The biggest direct effect on achievement ($\beta=0.57$) was found in case of *language aptitude*, which also directly effected the students' *beliefs on language learning* ($\beta=0.35$). Beliefs, on the other hand, indirectly influence achievement through *feelings related to the difficulty of listening comprehension* (anxiety or the lack of it). *Parents' education* has both a direct ($\beta=0.18$) and an indirect effect on achievement through the related beliefs and feelings. *Anxiety concerning unknown words* was also found to exert a significant impact on achievement directly ($\beta=0.24$) and indirectly through *anxiety about comprehension* ($\beta=0.46$).

It can be stated that students' beliefs act as a mediator of the effects of their aptitude and their parents' education, making their way to achievement through emotional states. In other words, student beliefs, what they think about language learning, and their emotions, how they feel in the learning process, interact in determining children's development. The effect of beliefs on anxiety about listening comprehension ($\beta=-0.20$) and the effect of anxiety about listening comprehension on achievement ($\beta=-0.20$) are both negative, as expected based on the correlations. Those who are less anxious expect English to be easier and have a more positive self-concept as language learners. Consequently, those who are more positively inclined toward language learning achieve better, which is certainly also true the other way around.

School marks were used as additional measures of student achievement that evaluate their work throughout the school year. In the next section we discuss the relationships between IDs and the students' English marks in order to compare the overlaps of the two achievement variable with the variables of individual differences.

5.4 Relationships Between Variables of the First Assessment and English Marks

In Hungarian educational practice, the most significant indicators of school achievement are school marks, due to the lack of standardized methods and instruments of assessment that are the foundation of consistent evaluation of achievement in school subjects in other countries. School marks are traditionally used as indicators of student achievement, although research on school marks (Csapó, 2002a, 2002b) highlighted several controversial phenomena: school marks weakly correlate with the actual knowledge measured by knowledge tests based on the school curriculum and text books (Csapó, 2002b). In this respect, English as a school subject is in a better position compared to other subject, since the highest correlation was found between test results and school marks ($r=0.52$ in grades seven and eight). This finding was explained by traditions in standardized testing in English language assessment in contrast with other school subjects, since language proficiency exams have clearly defined criteria and hence measuring language skills must have improved practice (Csapó).

The discussion of student achievement is complemented by a detailed description of the relationships between English marks and ID variables and we attempt to clarify causal relationships by including this achievement indicator in the path analysis. First, a *cluster analysis* was conducted to explore the system of relationship between the ID variables and to highlight how these variables are grouped in connecting to school achievement. *Furthest neighbour* method was used in the cluster analysis. The dendrogram of the results is shown in Fig. 3.

The variables shown in Fig. 3 are grouped in two larger clusters containing three smaller clusters. Aptitude formed a separate cluster. By reviewing the steps of cluster formation it can be seen how the individual variables relate to one another. English mark is grouped in one cluster with the components of strategy and motivation, whereas aptitude forms a separate cluster with the components of individual differences.

Next, the predictive effect of individual difference variables on English marks was analyzed. Table 7 shows the β and explained variance values of the nine variables.

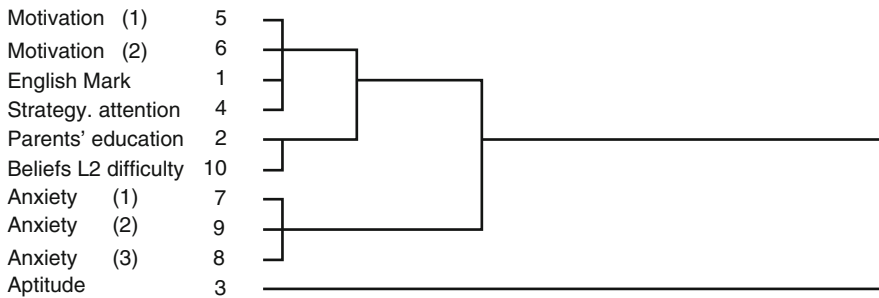


Fig. 3 Dendrogram of clusters around English marks. *Explanation: Motivation (1): classroom level; Motivation (2): student level (self-concept); Anxiety (1): following the text; Anxiety (2): difficulty of understanding; Anxiety (3): unknown words*

Table 7 Variables of first assessment predicting English language marks

Individual differences	β	$r^2\beta$ (%)
Parents' education	0.140	3.9
Language aptitude	0.352**	17.4**
Strategy: <i>directed attention</i>	0.195**	6.3**
Attitude and motivation: <i>classroom level</i>	-0.010	-0.2
Attitude and motivation: <i>learner level (self-concept)</i>	0.260**	11.0**
Anxiety about listening comprehension: <i>following the text</i>	0.034	-0.3
Anxiety about listening comprehension: <i>difficulty of comprehension</i>	0.016	-0.2
Anxiety about listening comprehension: <i>unknown words</i>	-0.084	1.1
Beliefs: <i>difficulty of language learning</i>	-0.009	-0.3
<i>Total variance explained (R²)</i>		38 %

** $p < .01$

According to the data in Table 7, the total variance explained comes close to 40 %. It is apparent that only three variables have significant β values predicting English marks. Aptitude has the highest share in the variance explained, accounting for almost 50 % of the total. The second most significant predictor is the level of student motivation and attitude, i.e. self-concept of the learner, describing how successful or less successful the students perceive themselves. The third significant variable is strategy of directed attention to the keywords; metacognitive awareness about listening comprehension is one of the most important and most frequently applied strategies of listening comprehension, as was confirmed in student interviews. It is also shown that parents' education does not directly predict English marks.

Finally, a *path-analysis* was conducted involving the significant variables resulting from *regression analysis* in order to reveal causal relationships between the variables in relation to the English marks and the paths leading from IDs to student achievement evaluated by school marks. Figure 4 shows the *path-diagram* of assumed causal relationships.

Three of the ID variables had significant β values, meaning that the direct and indirect effects of these three variables explain the variance in English marks. First, the parameters of model fit are reviewed. The saturated model had 14 parameters, the tested model 13, df was 1. Values of $\chi^2=0.088$, $p=0.767$ suggest that the test was not significant, showing that the tested model is a good fit. Path-coefficients (β values) are significant in all relationships. NFI=0.999 and CFI=1.000 values also reflect adequate level, exceeding 0.9 (good fit) level. Finally, RMSEA<0.001 is well below 0.05, indicating good model fit.

There are different paths, however, leading to school marks, the other variable of student achievement. Also, the predictive force of ID variables was considerably lower (35 %) in this case. The most reliable predictor of English language school

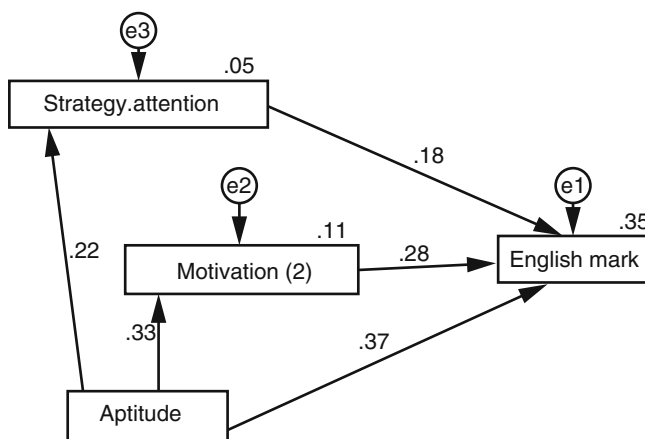


Fig. 4 Variables of IDs and causal relationships of English language marks. *Explanation:* *Motivation (2)*: student level (self-concept)

marks was *language aptitude* both directly ($\beta = 0.37$) and indirectly influencing school marks. In this latter case the effect of language aptitude was mediated by the *attention to keywords strategy* ($\beta = 0.22$) and learners' *self-concept as a motive* ($\beta = 0.33$). These variables had a significant direct effect on marks. As is shown in Fig. 4, mainly cognitive factors in the traditional sense account for the English marks. From the affective factors, only motivational self-concept impacts the marks.

The resulting paths in both analyses seem to support Dörnyei's assumption (2010): we interpret IDs as dynamic interactions of hierarchically organized components and cognitive and affective factors (within and between themselves) as overlapping rather than dichotomic constructs. It became clear that the two achievement variables, listening comprehension test achievement and English school marks, were explained by different variables of individual differences to a different extent.

6 Conclusion

The findings of the research are in line with the predictions of the theoretical framework (Dörnyei, 2006, 2009, 2010; Gardner & MacIntyre, 1992, 1993): the ID variables are multifactor constructs in themselves, the constituents are in constant interaction with each other and their environment, changing and consequently creating a complex pattern of development. Both the components of individual differences and systemic models of the connections in student achievement support Dörnyei's assumption (2010) that the traditional separation of cognitive and affective variables (Gardner & MacIntyre, 1992, 1993) can be problematic.

The findings confirmed that language aptitude and parents' education are significant predictors of young learners' listening comprehension achievements (Csapó & Nikolov, 2009; Józsa & Nikolov, 2005; Kiss & Nikolov, 2005). The other primary factor in the traditional sense (Gardner & MacIntyre, 1993), the motivational component, was excluded from the predictive model of listening comprehension achievement. This seems to contradict previous findings, however, motivation was found to significantly predict school achievement represented by the English marks in this research, in line with others' findings (Dörnyei, 2009, 2010; Mihaljević Djigunović, 2006, 2009, 2014).

It was also revealed that listening comprehension achievement is predicted by the interaction between IDs and the learning context which is constantly changing throughout the learning process (Dörnyei, 2006, 2009, 2010; Mihaljević Djigunović, 2009).

Additionally, the findings shed light on the fact that learners' beliefs, thoughts and feelings related to the difficulty of language learning and students' aptitude have a significant effect both on one another and on achievement (Aragao, 2011; Bacsa, 2012). This means that what young learners think or believe about language learning and how they feel about their learning experience impact their achievement in listening comprehension. According to our model, these beliefs are rooted in the

young learners' social background (indicated by their parents' education) and language aptitude, and the direction of these relationships is the opposite of that displayed in Gardner and MacIntyre's (1993) model.

6.1 Implications and Direction for Further Research

There is a scarcity on the Hungarian research scene of instruments measuring early language learners' individual differences. This research has taken a step closer to developing the methods needed to explore individual differences and to understand the functioning of the already existing instruments for early language learners. Our findings could assist language teachers in identifying the strengths and weaknesses of their learners and discovering the potential in developing listening comprehension of early language learners. Accurate diagnosis could lead to the facilitation of learners' development and training programs.

This research investigated the development of a single skill from the perspective of the multifactor construct of individual differences. Further research involving larger, potentially representative samples would be needed to test the reliability of the instruments we applied and to gather more data from various perspectives on how they could be improved. More measures developed specifically for young learners would also be needed to explore additional hidden aspects of individual differences. Furthermore, it would be important to examine reading, writing and speaking in similar circumstances by using diagnostic measures in order to better understand their development, and to allow teachers to facilitate their young learners better.

References

- Ackerman, P. L. (2003). Aptitude complexes and trait complexes. *Educational Psychologist*, 38, 85–93.
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.
- Aragao, R. (2011). Beliefs and emotions on foreign language learning. *System*, 39(3), 302–313.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.
- Bacsá, É. (2012). Az idegennyelv–tanulással kapcsolatos meggyőződések vizsgálata általános és középiskolás tanulók körében [Beliefs about language learning]. *Magyar Pedagógia*, 112(3), 167–193.
- Bacsá, É. (2014). *The contribution of individual differences to the development of young learners' listening performances*. Doctoral thesis, University of Szeged, Szeged, Hungary. Retrieved from <http://www.doktori.hu/index.php?menuid=193&vid=13859>
- Bacsá, É., & Csíkó, Cs. (2013). *The contribution of individual differences to the development of young learners' listening performances*. 15th European conference for the research on learning and instruction: "Responsible Teaching and Sustainable Learning". Munich, Germany: TUM.

- Bors, L., Lugossy, R., & Nikolov, M. (2001). Az angol nyelv oktatása pécsi általános iskolákban [Teaching English in elementary schools in Pécs]. *Iskolakultúra*, 1(4), 73–88.
- Brózik-Piniel, K. (2009). *The development of foreign language classroom anxiety in secondary school*. Doctoral thesis, ELTE, Budapest, Hungary. Retrieved from <http://www.doktori.hu/index.php?menuid=193&vid=4014>
- Brown, H. D. (1994). *Teaching by Principles: An interactive approach to language pedagogy*. New York: Prentice Hall Regents.
- Bukta, K., & Nikolov, M. (2002). Nyelvtanítás és hasznos nyelvtudás: az angol mint idegen nyelv [Teaching a language for usable language competence: the case of English as foreign language.] In B. Csapó (Ed.), *Az iskolai műveltség* [School literacy] (pp. 169–192). Budapest, Hungary: Osiris.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to language learning and testing. *Applied Linguistics*, 1, 1–47.
- Chang, A. C., & Read, J. (2007). Support for foreign language listeners: Its effectiveness and limitations. *RELC Journal*, 38(3), 375–395.
- Cohen, A. D. (1998). *Strategies for learning and using a second language*. New York: Longman.
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
- Csapó, B. (2001). A nyelvtanulást és a nyelvtudást befolyásoló tényezők [Factors affecting language competence and language learning]. *Iskolakultúra*, 1(8), 25–35.
- Csapó, B. (2002a). Az iskolai tudás felszíni rétegei: Mit tükröznek az osztályzatok? [The superficial stratum of school knowledge: What are school marks for?]. In B. Csapó (Ed.), *Az iskolai tudás* [School knowledge] (pp. 37–63). Budapest, Hungary: Osiris.
- Csapó, B. (2002b). Iskolai osztályzatok, attitűdök, műveltség [School marks, attitudes, and literacy]. In B. Csapó (Ed.), *Az iskolai műveltség* [School literacy] (pp. 37–65). Budapest, Hungary: Osiris.
- Csapó, B., & Nikolov, M. (2009). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences*, 19, 209–218.
- Csíkós Cs., & Bacsa, É. (2011). *Measuring beliefs about language learning*. Paper presented at the meeting of the 14th biennial conference for research on learning and instruction: “Education for a Global Networked Society”. Exeter, UK: University of Exeter.
- Csizér, K., & Dörnyei, Z. (2002). Az általános iskolások idegen nyelv-tanulási attitűdje és motivációja [Language learning attitudes and motivation among primary school children]. *Magyar Pedagógia*, 102(3), 333–353.
- Csizér, K., Dörnyei, Z., & Németh, N. (2004). A nyelvi attitűdök és az idegen nyelvi motiváció változásai 1993 és 2004 között Magyarországon [Changes in language attitudes and motivation to learn foreign languages in Hungary, 1993–2004]. *Magyar Pedagógia*, 104(4), 393–408.
- Dörnyei, Z. (1998). Motivation in second and foreign language learning. *Language Teaching*, 31, 117–135.
- Dörnyei, Z. (2001). New themes and approaches in second language motivation research. *Annual Review of Applied Linguistics*, 21, 43–59.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Laurence Erlbaum Associates.
- Dörnyei, Z. (2006). Individual differences in second language acquisition. *AILA Review*, 19, 42–68.
- Dörnyei, Z. (2007). *Research methods in applied linguistics*. Oxford, NY: Oxford University Press.
- Dörnyei, Z. (2009). Individual differences: Interplay of learner characteristics and learning environment. *Language Learning*, 59(1), 230–248.
- Dörnyei, Z. (2010). The relationship between language aptitude and language learning motivation: Individual differences from a dynamic systems perspective. In E. Macaro (Ed.), *Continuum companion to second language acquisition* (pp. 247–267). London: Continuum.
- Dörnyei, Z., & Csizér, K. (2002). Some dynamics of language attitudes and motivation: Results of a longitudinal nationwide survey. *Applied Linguistics*, 23(4), 421–462.

- Dörnyei, Z., MacIntyre, P. D., & Henry, A. (2015). Introduction: Applying complex dynamic systems principles to empirical research on L2 motivation. In Z. Dörnyei, P. D. MacIntyre, & A. Henry (Eds.), *Motivational dynamics in language learning* (pp. 1–7). Bristol, UK: Multilingual Matters.
- Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Oxford, NY: Blackwell.
- Dunkel, P. (1986). Developing listening fluency in L2: Theoretical principles and pedagogical considerations. *Modern Language Journal*, 70, 99–106.
- Ellis, R. (1994). *The study of second language acquisition*. Oxford, NY: Oxford University Press.
- Everitt, B. S., & Dunn, G. (1991). *Applied multivariate data analysis*. London: Edward Arnold.
- Eysenck, M. W., & Kean, M. T. (2005). *Cognitive psychology: A student's handbook*. Hove, UK: Psychology Press.
- Field, J. (2004). An insight into listeners' problems: Too much bottom-up or too much top down? *System*, 32(3), 363–377.
- Gardner, R. C. (1985). *Social psychology and second language learning: The role of attitudes and motivation*. London: Edward Arnold.
- Gardner, R. C., & MacIntyre, P. D. (1992). A student's contributions to second language learning. Part I: Cognitive variables. *Language Teaching*, 25, 211–220.
- Gardner, R. C., & MacIntyre, P. D. (1993). A student's contributions to second language learning. Part II: Affective variables. *Language Teaching*, 26, 1–11.
- Goh, C. C. M. (2008). Metacognitive instruction for second language listening development: Theory, practice and research implications. *RELC Journal*, 39(2), 188–213.
- Griffiths, C. (2003). Patterns of language learning strategy use. *System*, 31(3), 367–383.
- Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early versus later English language proficiency among language learners. *Early Childhood Research Quarterly*, 27, 1–20.
- Hardy, J. (2004). Általános iskolás tanulók attitűdje és motivációja az angol mint idegen nyelv tanulására [The attitude and motivation of primary school children for learning English as a foreign language]. *Magyar Pedagógia*, 104(2), 225–242.
- Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: an innovative approach. *Language Testing*, 17(2), 261–277.
- Heitzmann, J. (2009). The influence of the classroom climate on students' motivation. In R. Lugossy, J. Horváth, & M. Nikolov (Eds.), *UPRT 2008: Empirical studies in English applied linguistics* (pp. 207–224). Pécs, Hungary: Lingua Franca Csoport.
- Moon, J., & Nikolov, M. (2000). *Research into teaching English to young Learners*. Pécs, Hungary: University Press.
- 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.
- Johnson, K. (2001). *An introduction to foreign language learning and teaching*. London: Longman.
- Józsa, K., & Nikolov, M. (2003). Az idegen nyelvi készségek fejlettsége angol és német nyelvből a 6. és 10. évfolyamon a 2002/2003-as tanévben [Levels of performances in English and German in year 6 and 10]. Budapest, Hungary: OKÉV.
- Józsa, K., & Nikolov, M. (2005). Az angol és német nyelvi készségek fejlettségét befolyásoló tényezők [Factors influencing achievement in English and German as foreign languages]. *Magyar Pedagógia*, 105(3), 307–337.
- Kim, Y. (2001). *Foreign language anxiety as an individual difference variable in performance: An interactionist's perspective* (ERIC Document Reproduction Service No. ED 457 695)
- Kim, J. (2005). The reliability and validity of a foreign language listening anxiety scale. *Korean Journal of English Language and Linguistics*, 5(2), 213–235.
- Kiss, C. (2009). The role of aptitude in young learner's foreign language learning. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 253–276). Berlin, Germany: 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.
- Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing*, 21, 390–403.
- Krashen, S. (1985). *The input hypothesis: Issues and implications*. New York: Longman.
- Lantolf, J. P., & Poehner, M. E. (2011). Dynamic assessment in the classroom: Vygotskian praxis for second language development. *Language Teaching Research*, 15(1), 11–33.
- Larsen-Freeman, D., & Long, M. H. (1991). *An introduction to second language acquisition research*. London: Longman.
- Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research*, 24(1), 35–63.
- MacWhinney, B. (2005). A unified model of language acquisition. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford, UK: Oxford University Press.
- Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, 8, 1–71.
- Martin, A. J. (2009). Motivation and engagement across the academic life span. *Educational and Psychological Measurement*, 69, 794–824.
- Matsuda, S., & Gobel, P. (2004). Anxiety and predictors of performance in the foreign language classroom. *System*, 32(1), 21–36.
- Mattheoudakis, M., & Alexiou, T. (2009). Early language instruction in Greece: Socioeconomic factors and their effect on young learner's language development. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 227–252). Berlin, Germany: Mouton de Gruyter.
- McKay, P. (2006). *Assessing young learners*. Cambridge: Cambridge University Press.
- 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. 9–23). Pécs, Hungary: Lingua Franca Csoport.
- Mihaljević Djigunović, J. (2009). Individual differences in early language programmes. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 198–223). Berlin, Germany: Mouton de Gruyter.
- Mihaljević Djigunović, J. (2010). Starting age and L1 and L2 interaction. *International Journal of Bilingualism*, 14(3), 303–314.
- Mihaljević Djigunović, J. (2014). Developmental and interactional aspects of young EFL learners' self-concept. In J. Horváth & P. Medgyes (Eds.), *Studies in honour of Marianne Nikolov* (pp. 37–50). 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*. New York: Springer.
- Mónus, K. (2004). *Learner strategies of Hungarian secondary grammar school students*. Budapest, Hungary: Akadémiai Kiadó.
- Mordaunt, O. G., & Olson, D. W. (2010). Listen, listen, listen and listen: building a comprehension corpus and making it comprehensible. *Educational Studies*, 36(3), 249–258.
- Moschener, B., Anschuetz, A., Wernke, S., & Wagener, U. (2008). Measurement of epistemological beliefs and learning strategies of elementary school children. In M. S. Khine (Ed.), *Knowing, knowledge and beliefs* (pp. 113–137). New York: Springer.
- Münnich, Á., & Hidegkuti, I. (2012). Strukturális egyenletek modelljei: Oksági viszonyok és komplex elméletek vizsgálata pszichológiai kutatásokban [Models of structural equations: The investigation of causal relations and complex theories in psychological research]. *Alkalmazott Pszichológia*, 12(1), 77–102.
- National Core Curriculum (2003). Budapest, Hungary: Oktatási Minisztérium.
- Nikolov, M. (2000). Issues in research into early FL programmes. In J. Moon & M. Nikolov (Eds.), *Research into teaching English to young learners* (pp. 21–48). Pécs, Hungary: University Press Pécs.

- Nikolov, M. (2003a). Angolul és németül tanuló diákok nyelvtanulási attitűdje és motivációja [Attitudes and motivation of English and German learners]. *Iskolakultúra*, 3(8), 61–73.
- Nikolov, M. (2003b). Hatodikosok stratégiahaználata olvasott szöveg értését és íráskészséget mérő feladatokon angol nyelvből [Sixth-graders' test-taking strategies on reading and writing tasks]. *Magyar Pedagógia*, 103(1), 5–34.
- Nikolov, M. (2007). Variables influencing Hungarian 6th graders' foreign language learning and development. In A. Sheorey & J. Kiss-Gulyas (Eds.), *Studies in applied and theoretical linguistics* (pp. 1–24). Debrecen, Hungary: Kossuth Egyetemi Kiadó.
- Nikolov, M. (Ed.). (2009). *The age factor and early language learning*. Berlin, Germany/New York: Mouton de Gruyter.
- Nikolov, M. (2011). Az angol nyelvtudás fejlesztésének és értékelésének keretei az általános iskola első hat évfolyamán [A framework for developing and assessing English language proficiency in the first six grades of primary school]. *Modern Nyelvtudomány*, 17(1), 9–32.
- Nikolov, M. (2016). 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*. New York: Springer.
- Nikolov, M., & Curtain, H. (Eds.). (2000). *An early start: Young learners and modern languages in Europe and beyond*. Strasbourg, France: Council of Europe.
- Nikolov, M., & Józsa, K. (2006). Relationships between language achievements in English and German and classroom-related variables. In M. Nikolov & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 197–224). Pécs, Hungary: Lingua Franca Csoport, PTE.
- 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 colour: An overview of early teaching and learning of foreign languages. *Annual Review of Applied Linguistics*, 31, 95–119.
- Nikolov, M., & Szabó, G. (2011a). Az angol nyelvtudás diagnosztikus mérésének és fejlesztésének lehetőségei az általános iskola 1–6. évfolyamán [Diagnostic assessment and development of English language knowledge in grades 1 to 6 in the primary school]. In B. Csapó & A. Zsolnai (Eds.), *Akognitív és affektív fejlődés diagnosztikus mérése az iskola kezdő szakaszában* (pp. 13–40). Budapest, Hungary: Nemzeti Tankönyvkiadó.
- Nikolov, M., & Szabó, G. (2011b). Establishing difficulty levels of diagnostic listening comprehension tests for young learners of English. *UPRT* 73–82.
- Nunan, D., & Bailey, K. M. (2009). *Exploring second language classroom research: A comprehensive guide*. Boston, MA: Heinle, Cengage Learning.
- O'Malley, J., & Chamot, U. (1990). *Learning strategies in second language acquisition*. Cambridge, UK: Cambridge University Press.
- Ottó, I. (2003). A nyelvérték és mérése [Foreign language aptitude and its measurement]. *Alkalmazott Pszichológia*, 5(2), 57–64.
- Oxford, R. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House/Harper and Row.
- Richards, J. C. (2005). Second thoughts on teaching listening. *RELC Journal*, 36(1), 85–92.
- Robinson, P. (2001). Individual differences, cognitive, abilities, aptitude complexes and learning conditions in second language acquisition. *Second Language Research*, 17(4), 368–392.
- Rost, M. (2002). *Teaching and researching listening*. London: Longman.
- Sáfár, A., & Kormos, J. (2008). Revisiting problems with foreign language aptitude. *International Review of Applied Linguistics in Language Teaching*, 46(2), 113–136.
- Simon, O. (2001). A magyar és angol beszédészlelési és beszédmegértési teljesítmény összefüggései 11–12 évesek körében [Connections between speech perception and listening comprehension in Hungarian and English languages]. *Alkalmazott Nyelvtudomány*, 1(2), 45–61.
- Skehan, P. (1991). Individual differences in second language learning. *Studies in Second Language Acquisition*, 13, 275–298.
- Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.

- Sparks, R. L., Patton, J., & Ganschow, L. (2011). Subcomponents of second-language aptitude and second-language proficiency. *The Modern Language Journal*, 95, 253–273.
- Spinath, B., & Spinath, F. M. (2005). Longitudinal analysis of the link between learning motivation and competence beliefs among elementary school children. *Learning and Instruction*, 15, 87–102.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, UK: Cambridge University Press.
- Szabó, G., & Nikolov, M. (2013). An analysis of young learners' feedback on diagnostic listening comprehension tests. In M. J. Djigunovic, & M. Medved Krajnovic (Eds.), *UZRT 2012: Empirical studies in English applied linguistics*. Zagreb, Croatia: FF press. http://books.google.hu/books?id=VnR3DZsHG6UC&printsec=frontcover&source=gbbs_ge_summary_r&cad=0#v=onepage&q&f=false
- Tóth, Z. (2008). A foreign language anxiety scale for Hungarian learners of English. *WoPaLP*, 2, 55–78.
- Tóth, Z. (2009). Foreign language anxiety: For beginners only? In R. Lugossy, J. Horváth, & M. Nikolov (Eds.), *UPRT, 2008. Empirical studies in applied linguistics* (pp. 225–246). Pécs, Hungary: Lingua Franca Csoport.
- Vandergrift, L. (2005). Relationship among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, 26(1), 70–89.
- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, 90, 6–18.
- Vandergrift, L. (2012). Listening: Theory and practice in modern foreign language competence. <https://www.llas.ac.uk/resources/gpg/67>
- Wenden, A., & Rubin, J. (1987). *Learner strategies in language learning*. Hemel Hemstead, UK: Prentice Hall.
- 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*. New York: Springer.
- Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal*, 37(3), 308–328.
- Yim, S. Y. (2014). An anxiety model for EFL young learners: A path analysis. *System*, 42, 344–354.

Éva Bacsa is deputy headteacher at Kiss Bálint Reformed School in Szentes, Hungary. She holds PhD in educational science. Her research interest includes individual learner differences in early language learning.

Csaba Csíkos is associate professor of education at the University of Szeged, Institute of Education. His research topics include assessing and improving children's higher level thinking skills, and his previous publications focused primarily on mathematical abilities and reading strategies.

Self-Assessment *of* and *for* Young Learners' Foreign Language Learning

Yuko Goto Butler

Abstract Despite the recent focus on self-assessment (SA) as a tool for enhancing learning, some researchers and practitioners have expressed concerns about its subjectivity and lack of accuracy. Such concerns, however, originated from the traditional, measurement-based notion of assessment (assessment *of* learning) rather than the learning-based notion of assessment (assessment *for* learning). In addition, existing research on SA in second/foreign language education has been concentrated on adult learners, leaving us with limited information on SA among young learners. In this chapter, I address both sets of issues: the confusion between the two orientations for assessment and age-related concerns regarding SA. First, I clarify the two orientations of assessment—*assessment of learning* and *assessment for learning*—and demonstrate that most of the concerns about subjectivity and accuracy apply primarily to the former orientation. Second, I detail the current findings on SA among young learners and identify the most urgent topics for future research in this area. Finally, to help teachers and researchers examine and develop SA items that are most appropriate for their purposes, I propose five dimensions that characterize existing major SAs for young learners: (a) domain setting; (b) scale setting; (c) goal setting; (d) focus of assessment; and (e) method of assessment.

Keywords Self-assessment • Assessment for learning • Assessment of learning • Young learners • Foreign language learning • Summative assessment • Formative assessment • Age

1 Introduction

Coupled with the recent emphasis on learner-centered approaches to language teaching and self-regulated language learning, the use of various forms of SA is on the rise in language programs worldwide. According to Oscarson (1989), SA is a type of assessment where learners engage in “internal or self-directed” activities; as such,

Y.G. Butler (✉)

Graduate School of Education, University of Pennsylvania, Philadelphia, PA, USA
e-mail: ybutler@gse.upenn.edu

SA is “fundamentally different” from assessment conducted from the perspective of external agents such as teachers and test administrators (p. 1). In recent years, SA has gained popularity even among educators of young language learners (typically defined as children up to the end of primary school or sometime around 12 years old). It is no longer uncommon, for example, to see SA items in language textbooks for young learners. Primary school curricula often encourage teachers to use SA as a tool for evaluating students’ performance.

The growing attention paid to SA in early language education may be due to the fact that SA is considered to be a low-stakes form of assessment and so is assumed to be less stressful for young learners. Researchers have developed various types of can-do statements, a form of SA, for young learners and have made the statements readily available for teachers. Major efforts of developing can-do statements include CILT’s *Can-do Speech Bubble*, as part of Language Portfolio in the U.K. (CILT The National Center for Languages, 2006), and *Lingua Folio Junior* (National Council of State Supervisors for Languages, 2014), based on the American Council on the Teaching of Foreign Language (ACTFL) Proficiency Guidelines; the *Lingua Folio Junior* has been implemented on a trial basis in select U.S. states.

Despite the growing interest in SA, it has not had as large a presence in second language (L2) and foreign language (FL) classrooms at the primary school level as policy makers may have expected. The reasons for the slow take-up of SA in practice are presumably varied. But, perhaps most importantly, teachers often do not seem to know how or why they should use SA. Concerns have been expressed about the extent to which SA accurately captures young learners’ actual performance. Subjectivity has also been cited as a serious concern, particularly when SA is implemented in so-called “exam-driven” teaching and learning contexts.

Importantly, concerns regarding the accuracy and subjectivity of SA apply primarily to the traditional, measurement-based notion of assessment (assessment *of* learning) and thus are most relevant when SA is used primarily for summative purposes. When SA is implemented primarily for formative purposes, its accuracy may not be critical. From a process-oriented view of assessment (assessment *for* learning), assessment is considered to be a process of seeking relevant information, interpreting that information so that learners can reflect on their own learning, and making constructive decisions for further learning. As such, when the assessment is for learning, traditional psychometric notions of validity and reliability may not be suitable. Indeed, as Brookhart (2003) suggested, we need to sort out “classroometric” measurement concepts from psychometric measurement concepts (p. 8).

The major motivation for policies to promote SA for primary school teachers came from the theoretical association between SA and learning. Researchers agree that SA is a vital process for facilitating learners’ autonomy and self-regulation (Black & Wiliam, 1998, 2009; Blanche & Merino, 1989; Butler & Lee, 2010; Dickinson, 1987; Oscarson, 1989). The premise that SA can be aligned with self-regulated learning sounds promising. However, we still have only a limited understanding of how SA can best be used to facilitate children’s language learning. What kinds of feedback during and/or after SA would promote young learners’

self-reflection, which, in turn, would lead to further language learning? Researchers have just begun to explore these questions.

In addition, previous research on SA in L2/FL language education has predominantly dealt with adult learners and has paid little attention to the role of age in SA. Age-related concerns—such as the extent to which children can handle self-assessing their performance or abilities in L2/FL in the first place—should be addressed to inform practice.

In this chapter, I clarify the two assessment orientations (namely, *assessment of learning* and *assessment for learning*) while focusing on SA among young learners, and I discuss the possibilities and challenges of implementing SA among young learners from both points of view. I draw on examples from previous studies to illustrate my points. I then characterize major existing SA item types according to five dimensions, and discuss how different types of SA can be used for both assessment of learning and assessment for learning. I conclude by offering suggestions for future research on SA for young learners.

2 Two Approaches to SA for Young Learners

In the following sections, I discuss the two approaches for assessment (*assessment of* and *for* learning) in turn. They originated from different theoretical and epistemological traditions, and the distinctions need to be clarified. That being said, however, these approaches are not necessarily mutually exclusive but can instead be located on a continuum according to the degree of emphasis on learning. In practice, the same SA tool can be used for more evaluation-oriented means (*assessment of learning*) or for more learning-oriented means (*assessment for learning*).

2.1 *Self-Assessment of Learning*

In the assessment of learning orientation, assessment is a means of capturing a learner's true ability. Thus, the assessment is concerned with eliciting meaningful information for making accurate and consistent inferences about a learner's true ability. The learner is a subject being observed and is external to the inferences being made and the actions being taken as the result of the inferences (Brookhart, 2003).

2.1.1 SA as Assessment of Learning Among Adult Learners

Among adult learners, a great deal of research has been conducted with respect to the validity and reliability of SA as well as its use. With a few exceptions (e.g., Matsuno, 2009; Patri, 2002; Pierce, Swain, & Hart, 1993), there is ample evidence

indicating that SA results, at least among adults, are generally correlated with external criteria such as teachers' ratings, final grades in class, and objective tests (Bachman & Palmer, 1989; Blanche, 1990; Brantmeier & Vanderplank, 2008; Brantmeier, Vanderplank, & Strube, 2012; Dickinson, 1987; Hargan, 1994; Leach, 2012; Oscarson, 1997; Stefani, 1994). As a result, SA has been used for relatively high-stakes purposes, such as program placement (Hargan, 1994; LeBlanc & Painchaud, 1985) and choosing the appropriate level of tests (Malabonga, Kenyon, & Carpenter, 2005). However, the degrees of correlations with external criteria varied across studies. Factors that influenced accuracy of SA included the skill domain being assessed, the ways in which items were constructed, and learners' individual characteristics.

With respect to the skill domains being assessed, if we assume that productive skills (i.e., speaking and writing) require higher degrees of meta-awareness, such as pre-planning and self-monitoring, than receptive skills (i.e., listening and reading), we may expect that learners are better at self-assessing their productive skills than their receptive skills. Interestingly, in a meta-analysis of SA, Ross (1998) found the opposite to be the case: adult learners could self-assess their receptive skills (reading in particular) in L2/FL more accurately than their productive skills. It is not clear, however, if receptive skills are inherently easier to self-assess. In speculating about which factors might explain the surprising result, Ross suggested such things as learners' experiences (e.g., adult L2/FL learners at college are more likely to have engaged in reading activities more heavily than the other activities), the reference points that they used (e.g., the adult learners might have judged themselves in relation to the performances of other students in class), and the scales that were used in external measurements (e.g., writing assessments often use nominal or categorical scales that may not be readily applicable to correlational analyses). In general, people tend to more accurately self-assess lower order cognitive skills than they do higher order cognitive skills (Zoller, Tsapalis, Fastow, & Lubezky, 1997).

Second, how the items are worded and constructed influences learners' responses to SA. College students' responses differed based on whether the items were negatively worded (e.g., "I have trouble with..." and "I cannot do...") or positively worded (e.g., "I can do ..."), although the degree of inconsistency varied greatly depending on the items (Heilenman, 1990). Not too surprisingly, learners' SA accuracy improved when the items were provided in their L1 rather than the target language (Oscarson, 1997).

Finally, various factors associated with individual learners are also found to influence their SA accuracy. One of the factors studied most extensively is learners' proficiency levels and experiences with the target language (Blanche & Merino, 1989; Davidson & Henning, 1985; Heilenman, 1990; Orsmond, Merry, & Reiling, 1997; Stefani, 1994; Sullivan & Hall, 1997). These studies generally indicate that students with lower proficiency and/or less experience with the target language tend to overestimate their performance, whereas student with higher proficiency tend to be more accurate or underrate their performance. Other influential factors over the accuracy of SA responses include the ways in which learners understand and respond to scales and items (Heilenman, 1990), the ways in which learners retrieve

relevant memory to self-assess the given skills and performance (Ross, 1998; Shameem, 1998), learners' learning styles (Cassidy, 2007); their anxiety levels (MacIntyre, Noels, & Clément, 1997), and their levels of self-esteem and motivation (AlFallay, 2004; Dörnyei, 2001). Another important factor, which is of particular relevance to the current discussion, is the age of the learners.

2.1.2 SA as Assessment of Learning Among Young Learners

Research on SA as an assessment of L2/FL learning among young learners has been very limited so far. It is largely unclear if the results of studies among older learners described in Sect. 2.1.1 are applicable to young learners of L2/FL.

Responding to SA requires highly complicated mental processing. For example, consider the item "I can ask questions in class," which is included in O'Malley and Pierce's (1996) popular resource book for young learners of English as L2. In order to respond to this item using a 4-point scale (ranging from *not very well*, *okay*, *well*, to *very well*) as instructed, the children need to go through at least the following cognitive processes:

1. Comprehend what the item refers to (what it means to "ask questions");
2. Understand each scale level and differentiate them (what it means to say "I can ask questions *okay*" and how that statement differs from "I can ask questions *well*");
3. Retrieve and synthesize their recent linguistic performance of asking questions in class;
4. Set a reference point to make a judgment (making a judgment in relation to others in class, in relation to the learner's own goal, or based on some other criteria).

While Harris (1997) asserts that "younger learners may be less resistant to the concept of self-assessment" (p. 18), given the complexity of cognitive processing required for answering SA, one may wonder about the extent to which children can accurately assess their own performance and abilities. From the *assessment of learning* point of view, at least two major issues must be examined: (a) how we should interpret children's responses to SA items (interpretation-related issues); and (b) the factors that influence the accuracy of their SA responses (measurement-related issues). I discuss these issues, which are summarized in Fig. 1, in the following sections.

2.1.2.1 Interpretation-Related Issues in Young Learners' SA of Learning

Previous research on children's development of self-appraisal and competence indicates that young learners' self-appraisal has been consistently high regardless of their actual performance. More specifically, children's self-appraisal remains very positive during the pre-school and early primary school years, and it starts declining

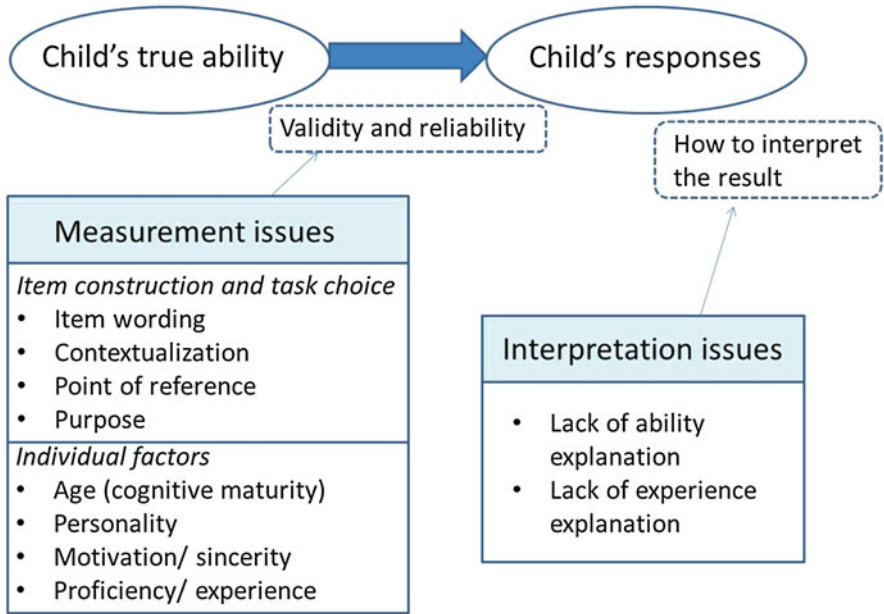


Fig. 1 Two major issues for self-assessment of learning for young learners (Note: SA of learning primarily concerns how best to elicit children’s true abilities. In the process, there are two major issues: measurement issues and interpretation issues)

sometime around the ages of 7–9, with another drop around the ages of 11–13. The accuracy of children’s perceived competence (examined by calculating correlations with external measures such as their teachers’ ratings) increases after the age of 8, when they start using social-comparative information (information indicating that one’s performance or ability is superior or inferior to others). Although social-comparative information begins to influence children’s self-appraisal of *performance* by the time they are around 7 years old, it does not influence self-appraisal of their *abilities* until much later (around 11–12 years old) (R. Butler, 2005).

Researchers’ interpretations of children’s self-appraisal behaviors have been changing in recent years. Traditionally, children’s unrealistically high self-appraisal was mainly attributed to their lack of cognitive maturity for making accurate judgments about their performance and abilities. Piaget’s (1926/1930) well-known stage theory of cognitive development certainly made a tremendous impact on researchers’ interpretation. According to this theory, children at the preoperational stage (ages 2–7) struggle with logical thinking; instead, their thoughts are dominated by concrete reasoning and intuition. This theory also posits that children are egocentric and have a hard time taking other people’s perspectives. The theory goes on to say that children at the concrete operational stage (ages 7–11) gradually begin to operate logical thinking and to differentiate their own thoughts from those of others. However, they still have difficulty handling multiple perspectives systematically

and forming abstract and causal thinking. In line with this theory, Stipek (1984) offered an explanation for why children are not only unrealistic but also excessively positive in their perceived competence by proposing their “wishful thinking” interpretation; namely, children cannot distinguish reality from their wishes, and they tend to make decisions based on the latter.

Similarly, interpretations based on achievement goal theory assumed that children's accuracy in evaluating their own abilities would be partially based on the development of their conception of *ability*. The theory proposed that there are two distinctive goal perspectives when perceiving one's ability: a *task-goal perspective* and an *ego-goal perspective*. The task-goal perspective is based on one's subjective assessment of task achievement and mastery. The ego-goal perspective relies on one's demonstration of superior performance compared to others (Dweck, 1986; Nicholls, 1989). According to this theory, children up to 7 years old cannot distinguish between ability and effort when it comes to determining performance on a task (referred to as *undifferentiated conception of ability*); for them, effort is ability. Thus, for young children, a person with high ability refers to a person who makes effort or obtains a high score in a given task, but they do not understand how to conceptualize a person who makes effort but achieves low in the given task, or vice versa. Researchers believed that young children are relatively invulnerable to failure and that they tend to respond to the failure by increasing effort. They also believed that children do not fully develop the concept of normative difficulty; instead, they tend to judge task difficulty in an egocentric fashion (e.g., this task is difficult because it was hard for me) (Nicholls & Miller, 1983). As they grow, children gradually understand that there is a cause-and-effect relationship between effort and outcome (outcome is a result of effort). But according to this theory, it is only after children reach the ages of 11–12 that they fully understand that one's performance (outcome) is also constrained by one's ability (referred to as *mature conception of ability*). After children reach this level, they can construct perceived competence in relation to other people's performance (Nicholls, 1978; also see Mihaljević Djigunović, 2016 in this volume.)

If children's self-evaluative abilities are mainly constrained by their underdeveloped internal mental structures, it makes sense to hold off on implementing SA of learning until they reach a cognitively mature state. However, in contrast to the results of experimental studies, anyone who spends sufficient time with children may notice that they appear to have more sophisticated self-evaluative knowledge and skills in naturalistic contexts than the cognitive-developmental theories predict. Indeed, neo- or post-Piagetian researchers indicate that children's self-evaluative abilities vary greatly depending on contexts, domains, and tasks at a given age level (see Flavell, 1999, for a review of such studies). Children's self-appraisal becomes more accurate if they can engage in familiar tasks and tasks that require lower levels of cognitive demand to perform. Experiences with different domains (e.g., math, music, language) help them develop distinct, domain-specific, and stable self-evaluative competence. Children who have intensive social contacts with other children can use normative information (information based on social comparison) more appropriately and are less ego-centric than those who don't, as we can see,

for example, in the work of Vygotsky (1934/1978). Children may also be more vulnerable to failure than was previously thought (Carless & Lam, 2014). R. Butler (2005) argued that:

regarding competence assessment, one implication is that self-appraisal may indeed become more accurate, differentiated and responsive to relevant information with age, in large part, however, because of age-related changes in children's typical experiences and contexts, rather than their internal cognitive structures. (p. 208)

In addition, potential problems have been raised with respect to the methodologies of many earlier studies of cognitive development. Children's failure in tasks may not be a sign of their lack of abilities but may be due to their misunderstanding the researchers' questions or intentions. For example, children as young as 4–5 who were once thought to be incapable of rating their performance based on temporal comparison (i.e., comparing their current performance with that in the past) turned out to be able to so as long as the information provided to them for evaluation was meaningful and familiar to them (R. Butler, 2005). These more recent findings on and interpretations of children's assessment competence remind us that we need to pay careful attention to contexts, assessment task choice, and the ways in which SA is constructed and delivered.

2.1.2.2 Measurement-Related Issues in Young Learners' SA of Learning

It is also important to understand measurement-related factors that contribute to children's biases and influence the accuracy of SA responses during the administration of *SA of learning*. As shown in Fig. 1, such measurement-related factors can largely be classified into two types: (a) item construction and task choice issues; and (b) individual factors, such as the child's age, personality, and proficiency.

The factors listed in Fig. 1 are based on previous studies, which were conducted primarily among adult L2 learners. How these factors may influence children's self-assessment responses is largely unknown.

As I examine in detail in later sections, different formats of SA items have been used; some SA tools employ multiple-choice formats while others ask learners dichotomous questions (i.e., requiring either "Yes/Can do" or "No/Cannot do" responses). Many SA items for young learners are short and simple, but some SA items provide the learner with more detailed contextual information. There has been very limited research examining if children have response biases based on different SA formats and item wording when assessing their L2/FL abilities. In a clinical setting, Chambers and Johnston (2002) found that, when asked to rate their own feelings (referred to as a *subjective task*) and other people's feelings (referred to as a *social objective task*) in a Likert scale, younger children (5–6 year olds) tended to show more extreme responses in both tasks than older children (7–9 year olds and 10–12 year olds). However, this response bias was not observed when the same children were asked to rate physical characteristics described in pictures using a Likert scale (referred to as an *objective task*) even among the youngest group that

they examined. Interestingly, the response bias observed in the youngest group was *not* found to be a function of the number of choices in the Likert scales; their responses did not differ between the three-level and five-level Likert scales. We do not know, however, if dichotomous items would have made any difference on the children's responses. Judging from the previous studies conducted in domains other than L2/FL, children do not seem to handle negatively worded items well (e.g., "I am not good at doing math") compared with positively worded items (e.g., "I am good at doing math") (e.g., Marsh, 1986). Considering the possible domain specificity of children's responses, however, we need to examine whether a similar response bias is observed when children self-evaluate their L2/FL.

SA items are often highly decontextualized—see, for example, the item "I can ask questions in class," which I quoted from O'Malley and Pierce (1996) in Sect. 2.1.2. However, depending on the age of children, the degree of contextualization can be a potential threat to the validity of SA of learning. In a study I did with a colleague (Butler & Lee, 2006), we compared children's (9–10 year olds and 11–12 year olds) responses to two formats of SA, an *off-task* SA and an *on-task* SA, concerning their oral performance in an FL. The off-task SA was a type of SA that asked learners to self-evaluate their general performance in a decontextualized fashion, as exemplified by the example item I quote above. The on-task SA was a contextualized SA in which learners were asked to self-evaluate their performance in a specific task immediately after the task was completed. We compared the children's responses to these two types of SA items with an objective proficiency measurement and an assessment of the children based on their teachers' classroom observations. We found that the children could self-assess their performance more accurately in the contextualized format than the decontextualized format. Not too surprisingly, the younger group (9–10 years) had a harder time with the decontextualized format than the older group. We also found that the children's responses to the contextualized format, compared with the decontextualized format, were less influenced by their attitudes and personality factors.

Considering the potential age- and experience-related challenges children may face when making temporal and/or normative comparisons while self-evaluating their abilities (see Sect. 2.1.2.1), it seems safe to assume that how researchers define reference points for SA (e.g., setting learners' own previous performance or other people's performance as a reference point) will influence children's responses to the SA items. Unfortunately, we know little about how children rely on different reference points when they assess their L2/FL abilities. In fact, our knowledge of the self-assessing process is quite limited, even when considering adult learners. Moritz's (1995) exploratory study based on a think-aloud protocol and retrospective interviews revealed that college students of French as FL used a variety of reference points (both temporal and normative information) when self-assessing their French abilities.

We can also assume that the extent to which young learners of L2/FL understand the purpose of SA influences the accuracy of their responses. In an intervention study of SA that I conducted with Lee (Butler & Lee, 2010), one of the challenges that the participating primary school teachers reported was how to provide their

students with initial guidance in order for them to treat SA seriously. It was particularly challenging to implement SA in a competitive, exam-driven environment. A teacher who taught in a competitive environment told us that she believed that SA had to be tied to other assessments or grades in order to ensure the accuracy of her students' responses. However, a teacher who taught in a much less competitive environment did not see such measures as necessary. We know from the research on the development of self-appraisal among children that their motivation for responding to SA accurately seems to increase with age but not in a linear fashion. Moreover, their motivation for accurate SA is also influenced by the amount of domain-specific knowledge they have acquired as well as by the context in which SA is conducted. For example, children's positive bias is motivated if the context and culture value positive self-appraisal. Accuracy of response is also constrained (more likely negatively biased) if the child realizes that there is a social cost for aggrandizing self-appraisal (R. Butler, 2005). In any event, we need more studies on how best to situate SA so that children of different ages can understand the purpose of SA and are motivated to respond to SA accurately in their specific learning environments.

In addition to the issues related to item construction and task choice, various individual factors likely influence the accuracy of children's SA responses. Such factors include cognitive maturity, personality, motivation, proficiency in the target language, and experience with SA. The role of individual differences in children's responses in SA is an unexplored area of inquiry, and so I can offer no practical, research-based suggestions for ensuring the accuracy of SA of learning among children.

2.2 *Self-Assessment for Learning*

While research on SA to date has been conducted primarily from an *assessment of learning* orientation, researchers have been giving increasing attention to SA as a formative assessment, with the goal of discovering its potential for influencing learning.

In taking the assessment for learning approach, the relationship between validity and reliability may need to be conceptualized differently. According to Sadler (1989), in the traditional assessment of learning, higher reliability is necessary but not sufficient for ensuring higher validity; a test can be highly reliable but can be off target. Thus, reliability serves as a precondition for validity. In contrast, with assessment for learning, validity should be a precondition for reliability because, according to Sadler (1989), "attention to the validity of judgments about individual pieces of work should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement" (p. 122).

Validity and reliability can themselves be conceptualized very differently depending on which approach is used. In the assessment for learning orientation, assessment is considered as part of instruction and "is usually informal, embedded in all aspects of teaching and learning" (Black, Harrison, Lee, Marshall, & Wiliam,

2003, p. 2). In assessment for learning, *validity* refers to the extent to which both the content of the assessment and the assessments' methods and tasks are matched with instruction. Thus, assessment for learning is deeply embedded in the particular context of the assessment. In assessment for learning, learners are no longer merely objects being measured; they are active participants who make inferences and take actions, together with the teachers, for formative purposes. According to Brookhart (2003), the validity concerns of assessment for learning include the degrees and the ways in which learners can self-reflect and benefit from having assessment enhance their learning. Similarly, teachers' knowledge, beliefs, and practices are all part of the validity concerns as well. In assessment for learning, *reliability* refers to the degree of stability of "information about the gap between students' work and 'ideal' work (as defined in students' and teachers' learning objectives)" (p. 9).

By engaging learners in self-reflection, SA is considered to be effective for developing their self-regulation, which can be defined as "the self-directive process by which learners transform their mental abilities into academic skills" (Zimmerman, 2002, p. 65), and should enhance their motivation and learning. However, empirical studies examining the effect of SA on learners' motivation and learning have been limited, particularly in relation to L2/FL.

2.2.1 SA as Assessment for Learning Among Adult Learners

Among adult learners, intervention studies of SA indicate that learners' perceived effects of SA were generally positive. For example, Orsmond, Merry, and Reiling (1997) found that, out of 105 college-level biology students, 98 % of them thought that SA made them think more and 71 % thought that they learned more, and 90 % found SA beneficial. Similarly, in Stefani (1994), out of 87 college students who conducted SA and 67 students who conducted peer-assessment in biochemical studies, nearly 100 % said that SA or peer-assessment procedures made them think more, and 85 % said they could learn more using these procedures than using the traditional tutor-lead assessment.

A number of studies on adults employed objective measures, such as external tests, grades, and teachers' or tutors' evaluations, in order to examine the effectiveness of SA on learning, and they identified some factors that led to positive outcomes. Such factors included receiving sufficient training to conduct SA (McDonald & Boud, 2003), setting clear criteria or rubrics (Andrade, Wang, Du, & Akawi, 2009), and having feedback (Taras, 2002). To facilitate learners' understanding of criteria and rubrics, researchers have suggested that presenting descriptive statements along with examples (e.g., writing examples for writing rubrics) would be effective. Having opportunities to discuss the meaning of the criteria with the teachers and tutors made the learners think more. Learning outcomes were different when the learners were allowed to construct their own criteria and when they were given criteria (Orsmond, Merry, & Reiling, 2000). Because peer-assessment should help learners understand the criteria better, it has been suggested that peer-assessment be implemented before SA (e.g., Nicol & Macfarlane-Dick, 2006).

This may make sense, particularly considering that peer-assessment was found to be psychometrically more internally consistent and to have higher correlations with external measures than SA (Matsuno, 2009; Patri, 2002) but that SA helped to increase learning more than peer-assessment (Sadler & Good, 2006).

Feedback is an essential part of SA for it to be effective for learning (Sadler, 1989), but having feedback itself does not guarantee positive outcomes. Hattie and Timperley's (2007) meta-analysis on feedback showed that there were substantial differences in effect sizes across studies, indicating that the quality and timing of the feedback greatly influenced learners' performance. Nicol and Macfarlane-Dick (2006) listed seven principles for good feedback practice:

- (1) helps clarify what good performance is (goals, criteria, expected standards); (2) facilitates the development of self-assessment (reflection) in learning; (3) delivers high quality information to students about their learning; (4) encourages teacher and peer dialogue around learning; (5) encourages positive motivational beliefs and self-esteem; (6) provides opportunities to close the gap between current and desired performance; and (7) provides information to teachers that can be used to help shape teaching. (p. 205)

Nicol and Macfarlane-Dick also stated that once learners have developed their self-evaluative skills to the point where they are able to engage in self-feedback, they can improve themselves even if the quality of external feedback is "impoverished" (p. 204).

In order to benefit from SA, learners themselves need to meet certain conditions. Sadler (1989) identified three such conditions: "(a) possess a concept for the *standard* (or goal, or reference level) being aimed for; (b) compare the *actual* (or current) *level of performance* with the standards; and (c) engage in appropriate *action* which leads to some closure of the gap" (p. 121). From a constructivist view of learning, such as that of Vygotsky (1934/1978), such learners' abilities are cultivated through having dialogues with and receiving assistance from their teachers or capable peers. Orsmond et al. (1997) also showed that learners' thorough understanding of the subject matter makes the SA results more useful.

In the field of L2/FL, empirical studies on the effect of SA on learning are limited. Among adult learners of French in Australia, de Saint Léger (2009) found that SA had a positive influence on their perceived fluency, vocabulary, confidence, and sense of responsibility for their own learning. Similarly, de Saint Léger and Storch (2009) found that SA had a positive influence on adult learners' willingness to communicate in an FL (e.g., perceived participation of class activities).

It is important to note, however, that many studies that examined the effect of SA on learning conceptualized learning as one-dimensional, sequential, and largely knowledge-based. Sadler (1989) reminded us that not all learning can be conceptualized as such, and stated that "the outcomes are not easily characterized as correct or incorrect, and it is more appropriate to think in terms of the quality of a students' responses or the degree of expertise than in terms of facts memorized, concepts acquired or content mastered" (p. 123). Indeed, we need more research examining the effect of SA on learning when learning is conceptualized as multidimensional, nonlinear, and nonstatic processes.

2.2.2 SA as Assessment for Learning Among Young Learners

When applied to young learners, empirical studies on SA from an assessment for learning orientation are scarce, particularly in the context of L2/FL. Thus, it remains unclear if most of the basic issues addressed in the previous section apply to young learners.

Figure 2 illustrates a conceptual model of SA as assessment for learning for young learners. Compared with Fig. 1, which shows a model for SA as assessment of learning, there are a few important points to note. First, in assessment for learning, SA for learning is embedded in specific social and educational contexts. Second, the emphasis is placed on a circular process of SA, which is carried out through repeated interactions between children and their teachers or peers. We can assume that the teachers or other capable peers would play greater roles in the process for young learners than they would for adult learners. Third, by having learners engage in self-reflection, SA ultimately aims to help them be self-regulated and autonomous learners. While young learners may have limited abilities to self-regulate their learning, depending on their cognitive maturity and experience (Zimmerman, 1989), children generally show substantial development in self-regulatory abilities during the preschool and primary school years (Morrison, Ponitz, & McClelland, 2010).

Before implementing SA, teachers need to (a) make sure that the assessment is consistent with the instruction and (b) choose tasks for assessment carefully. Some tasks or domains may be more difficult for children to self-evaluate than others. In Dann's (2002) case study, primary school students (ages 10–11) found it particularly

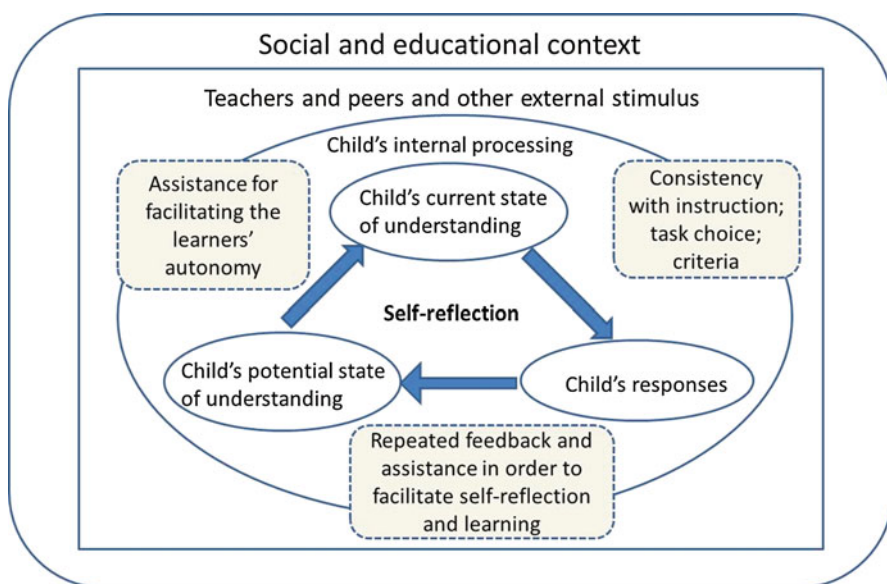


Fig. 2 The process of self-assessment for learning for young learners (Note: Components in SA described in *dotted squares* are key driving forces to facilitate learners' self-reflection processes)

difficult to assess listening compared with other domains. (Note, however, that Dann's study was conducted in a language arts context as opposed to an L2/FL context.) Unfortunately, we know very little about the kinds of tasks and performances that would be suitable for children—based on their cognitive maturity and experience—to engage in during SA.

As with adults, children need to understand the reasons for doing SA and have a clear understanding of the criteria. Children need to understand the goals and be invested in them in order to advance themselves (Torrance & Pryor, 1998). This appears to be the first hurdle to deal with, as indicated by Black et al.'s (2003) comment about young learners: “the first and most difficult task is to get students to think of their work in terms of a set of goals” (p. 49). In order to overcome this challenge, teachers may need to talk with children individually, perhaps on an ongoing basis. Although we have limited information on how children interpret the criteria for SA and make judgments using the criteria, it has been reported that children do not necessarily make judgments rationally—at least from the point of view of adults (Dann, 2002).

As suggested for adult learners, peer-assessment can help children understand the criteria better, and so it may be effective to implement peer-assessment before SA or along with SA (for a related discussion, see Hung, Samuelson, & Chen, 2016 in this volume). Dann's (2002) case study indicated that when children engaged in SA, they tended to draw on personal elements such as the *effort* that they had put into it in order to complete the work. Evaluating their peers' work (peer-assessment) seemed to help them objectify the criteria. In conducting peer-assessment with young learners, however, careful oversight is necessary. Research indicates that children who evaluate their peers' work and realize that their own progress and learning are limited compared to others are likely to lower their self-efficacy (Bandura, 1997), which in turn could negatively influence their further learning. In my studies in China (Butler, 2014, 2015), by the 8th grade (ages 13–14), some children started lowering their self-efficacy in FL learning at relatively early stages, and their level of self-efficacy turned out to be a major predictor of their FL performance.

It is also important to note that in assessment for learning, we do not necessarily adhere to the criteria in a strict sense. Instead, Dann (2002) suggested that “the priority given to pupil learning required a large degree of sensitivity in balancing the promotion of specific criteria with personal and individual factors” (p. 96–97). In other words, instead of considering the criteria to be absolute and fixed and expecting everybody to follow it uniformly, in assessment for learning the criteria should be flexible so that it can be adjusted according to the specific learning goals and needs of individual learners. Depending on the children's cognitive maturity and experience, they might even be able to actively participate in the process of developing criteria, in collaboration with their teachers.

SA can help teachers understand the gaps in a child's current state of understanding and his or her potential level of understanding (or an optimal goal for learning). It is important to note that children's judgment about their current understanding can be very different from the teachers' judgment, and thus dialogues are needed to

close the perceptual gaps between students and teachers. In order to become competent self-regulated learners, children have to develop metacognition to figure out *what they know* and *what they don't know*. As Harker (1998) stated, "only when students know the state of their own knowledge can they effectively self-direct learning to the unknown" (p. 13). And importantly, young learners are capable of monitoring their knowledge when they are provided with sufficient training. To facilitate the development of children's monitoring skills, SA should include items that capture the *process* of learning in addition to those that capture the learning outcome itself (Butler & Lee, 2010). After the gaps are understood by both the learner and the teacher, the teacher can help the learner set a goal within the zone of proximal development (ZPD, to use a Vygotskian term) and offer concrete assistance to help the learner reach the goal.

SA for learning is a recursive process. By repeating the process, SA ultimately aims to help children become self-regulated and autonomous learners. SA should be designed in such a way that learners can understand the goals of the tasks, self-reflect on their learning in relation to the goals, monitor their process of learning, and figure out what it takes to achieve the goals.

The teachers' role in the process of SA for learning is substantial. Y. G. Butler and Lee (2010) found that SA improved Korean primary school students' (ages 11–12) learning in English as well as their confidence but, importantly, the effects differed depending on individual teachers' attitudes toward assessment and their teaching context. When the teaching context was exam-driven and competitive, and if the teacher could not fully subscribe to the spirit of the assessment for learning, the effect of SA on the students' learning was limited. In other words, in order for SA to be effective, fostering a learning culture and the teachers' understanding of the assessment for learning appear to be indispensable.

3 Types of Major SAs

Various types of SA items have been developed for young learners in recent years. Some items are clearly designed for *SA of learning*, others are clearly designed for *SA for learning*, and still others can be used for either purpose, depending on the students' and teachers' needs and objectives. In this section, I examine major types of existing SAs, classifying them based on the following five dimensions and where they fall on the continua associated with those dimensions. These dimensions should be helpful for teachers and students as well as researchers when using existing SA items or developing their own items.

Domain setting

More general (open ended) ----- More specific

Scale setting

Fewer levels ----- More levels

More general (open ended) ----- More specific

Goal setting

More externally regulated ----- More self-regulated
 More static ----- More dynamic

Focus of assessment

More product oriented ----- More process oriented

Method of assessment

More individual based ----- More collaborative based

3.1 Domain Setting

SAs can vary in terms of domain specifications. In Example 1, the domain is defined very generally (i.e., speaking), and the assessment focuses only on fluency. Oskarsson (1978) called this type of SA “global assessment” (p. 13). It allows us to get only a rough picture of learners’ abilities.

Example 1 (Oskarsson, 1978, p. 37)¹

SPEAKING

Put a cross in the box which corresponds to your estimated level.

- 10 ← I am completely fluent in English
- 9
- 8
- 7
- 6
- 5
- 4
- 3
- 2
- 1
- 0 ← I cannot speak English at all.

However, in this format, the domain can be easily defined with increasing specificity, as in examples 2 and 3: “I can ask questions in class” (Example 2) is more specific than “speaking” (Example 1), and “I can ask where someone lives” (Example 3) is even more specific (ignore the scales of these examples for the time being).

¹This item did not include descriptions for each level of the scale, and was not meant for children. All other examples in this chapter were designed for young learners.

Example 2 (O'Malley & Pierce, 1996, p. 70)

I can ask questions in class

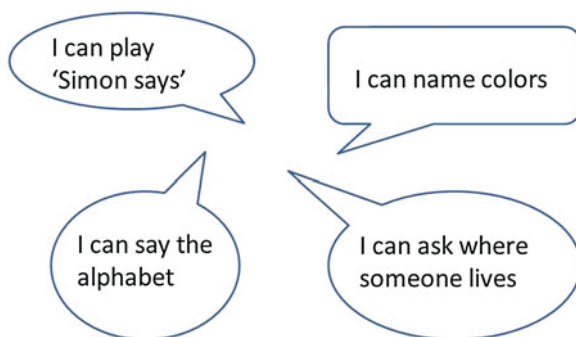
1. Not very well
2. Okay
3. Well
4. Very well

I can understand TV shows

1. Not very well
2. Okay
3. Well
4. Very well

Example 3 (CILT, European Language Portfolio, 2006, p. 11–12)²

Color in the speech bubbles when you can do these things.



From an assessment of learning perspective, the more concretely defined the domain specification, the more accurate the assessment, particularly among young learners. We can even set domains in a specific task that the children engaged in, as in Example 4. From an assessment for learning perspective, the assessment has to be embedded in context, as noted above; thus, contextualizing domain specificity is a critical condition for SA for learning.

²Some of the items in the European Portfolio are more or less specific. The items listed here are relatively specific.

Example 4 (Hasselgren, 2003, p. 79)

What I did in this task:

How true are these? Ring round the best number
(4= true, 3= more or less true, 2= partly true, 1= not true)

I managed to say what I wanted	4	3	2	1
I understood the others	4	3	2	1
I managed to 'keep the talk going'	4	3	2	1
I knew how to pronounce words	4	3	2	1
I knew enough words & phrases	4	3	2	1
I knew enough grammar	4	3	2	1
I managed not to mix languages	4	3	2	1
I liked doing this	4	3	2	1

Things I managed to do well:

Examples of words or phrases I learnt:

Things I still need to work on:

3.2 Scale Setting

The scale setting can be examined in two ways: (a) the number of levels and (b) the degree of specificity of each level. As I mentioned above, from the assessment of learning point of view, we don't know how many levels are optimal for young learners (i.e., yielding the most accurate responses). We can easily assume that the answer to this question depends, in part, on the degree of specificity of each scale level. Providing simple descriptions of each level, as in examples 2 and 4, may not necessarily contribute to higher accuracy. The scales still may be interpreted differently across children and, within a child, across items. It is important to make sure that children understand what each level means. While dichotomous SA items (can-do items), such as in Example 3, are increasingly popular at the primary school level, we still know very little about how children process and respond to dichotomous SA items, as discussed above.

Some SAs have detailed descriptions for each scale; such scales are often referred to as "descriptive rating scales" (Oskarsson, 1978, p. 16). In Example 5 (European Language Portfolio), each scale description corresponds to the *Common European*

Framework for Reference for Languages (CEFR, 2001). In general, the more detailed the descriptors, the easier it is for learners to respond. However, children may need assistance in comprehending the descriptors. Providing some concrete examples, as in Example 5, enhances children's comprehension of the descriptors.

Example 5 (CILT, European Language Portfolio, 2006, p. 32)

SPEAKING AND TALKING TO SOMEONE

A1 level: I can use simple phrases and sentences to describe where I live and people I know.

Grade 1: I can say/repeat a few words and short simple phrases

e.g., what the weather is like; greeting someone; naming classroom objects...

Grade 2: I can answer simple questions and give basic information.

e.g., *about the weather; where I live; whether I have brothers and sisters, or a pet...*

Grade 3: I can ask and answer simple questions and talk about my interests

e.g., *taking part in an interview about my area and interests; a survey about pets or favorite foods; talking to a friend about what we like to do and wear...*

From the assessment for learning point of view, scales can be useful if they are designed in such a way that learners can see the process or progress of their learning, or can identify the gaps in the current and potential levels of their learning. Scales can be set flexibly, according to individual learners' needs and learning trajectories.

3.3 Goal Setting

Goal setting refers to the process of identifying the goals of the SA, and it can be further divided into two sub-dimensions: (a) the extent to which learners have autonomy to identify the goals; and (b) the degree of flexibility with which goals can be defined. Granting autonomy and flexibility in goal setting may be a threat to the validity in the traditional assessment of learning approach, but it can be a critical feature for SA for learning, in order to help children to become autonomous and self-reflective learners. In Example 6, learners can choose from a list of predefined goals which goals they should aim for next. In Example 7, while some sample goals are listed, children can either come up with their own goals or choose their goals from the examples provided. The goals can be changed upon negotiation with the teacher.

3.4 Focus of Assessment

SAs can be designed to be more product-oriented or more process-oriented. SAs that are designed for assessment of learning are concerned mainly with what children can do (product), as exemplified in many can-do statements. Can-do items are also able to capture the degree of mastery by allowing for progressive responses (e.g., “I can do it all the time,” “I can do it most of the time,” “I can do it sometimes,” and “I can rarely do it”).

In assessment for learning, however, as we have seen already, it is critical to capture the process of learning—to make the learning process visible. We can see some attempt to capture the process in examples 4 and 7. Example 7 asks children to keep a record of their self-reflection on their performance. Upon receiving feedback from their teachers, the children can set a goal for the next class. By repeating this process and documenting it, the SA is designed to see the children’s progress over time.

Example 6 (Hasselgren, 2003, p. 78)

SPOKEN INTERACTION CHECKLIST: LEVEL A2.2

	Can you <u>usually</u> do these things? ³ Use these symbols: column 1 ✓=I think I can ✓✓=I know I can column 2 ✓=I aim to do this soon column 3 write the date when you’ve done an example of this	yes	myaim	example
1	I can understand what is said to me about everyday things if the other person speaks slowly and clearly and is helpful.			
2	I can show that I am following what people say, and can get help if I can’t understand.			
3	I can say some things to be friendly when I meet or leave someone.			
4	I can do simple ask-and-answer tasks with a partner in class, using expressions we have learnt.			
5	I can ask or tell the teacher about things we are doing in class.			
:				
:				

Example 7 (Kato, n.d., p. 6)⁴

1. Indicate today’s date
2. Write down your own goal(s) today
3. Indicate your performance in () using symbols below:
 ○ = super! ○ = Good Δ = Almost x = not done yet

³The underline was original. There are 12 items for each category.

⁴The original was in Japanese, and the select part was translated by the author.

4. *Write down your own reflection and submit it to your teacher*

Date	Your goal (write one or two)	Teachers' comments
	()	
	()	
	Your reflection	
Date	Your goal (write one or two)	Teachers' comments
	()	
	()	
	Your reflection	

Example goals

To try my best to engage in conversations, songs, and games in class

To speak (English) confidently

To talk to a foreign teacher

To effectively use gestures when speaking

To make eye contact to the partner when speaking

To used newly-learned words in conversation.....

3.5 *Method of Assessment*

SAs can be designed as an individual assessment activity or can be meant for more collaborative work. Although it is possible to use SAs for collaborative work even though they were originally meant to be carried out individually, SA items can also be designed in such a way that they invite other people's participation. This is particularly important for an assessment for learning orientation, in which it is critical to have a greater degree of collaboration (assistance from other capable individuals) in the SA process, especially during initial stages of children's SA practices. As children develop higher self-regulated skills, SAs can be conducted more independently.

4 **Conclusion and Implications**

Although recent policies often strongly encourage primary school language teachers to implement SA as a tool for helping children to gain greater ownership of their learning, many people continue to express concerns about the accuracy and subjectivity of SA. Such concerns, however, primarily originate from the traditional, measurement-based notion of assessment rather than learning-based notion of assessment. In addition, the age factor has not been sufficiently discussed in the previous research on SA. In this chapter, therefore, I clarified two notions of

assessment—*assessment of learning* and *assessment for learning*—while focusing on the case of SA among young learners. I also proposed five dimensions to characterize major SA items for young learners in order to help teachers and researchers to identify existing SA for use or develop SA items according to their own needs.

Research on SA among young learners of L2/FL is limited, and a number of important issues remain unresolved. With respect to *assessment of learning*, we need to uncover how item construction influences the way that children interpret and respond to items (e.g., what response bias we may observe depending on the number of scales and scale descriptors; how children use reference points; how the item wording may influence children's interpretation, etc.); and how various individual factors may influence the validity and reliability of SAs. From the *assessment for learning* point of view, we need to better understand children's *process* of engaging with SAs and its impact on their learning (e.g., how SAs enhance children's self-reflection, how both children and their teachers make inferences about the children's current and potential level of understanding, what kinds of actions were taken and their impact on children's learning, etc.). Importantly, we need more research that conceptualizes learning as a dynamic and non-linear process.

References

- AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System*, 32, 407–425.
- Andrade, H., Wang, X., Du, Y., & Akawi, R. (2009). Rubric-referenced assessment and self-efficacy for writing. *The Journal of Educational Research*, 102(6), 287–302.
- Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, 6(1), 14–29.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. New York: Open University Press.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7–74.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31.
- Blanche, P. (1990). Using standardized achievement and oral proficiency tests for self-assessment purposes: The DLIFC study. *Language Testing*, 6(1), 14–29.
- Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39(3), 313–340.
- Brantmeier, C., & Vanderplank, R. (2008). Descriptive and criterion-referenced self-assessment with L2 readers. *System*, 36(3), 456–477.
- Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System*, 40, 144–160.
- Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, 22(4), 5–12.
- Butler, R. (2005). Competence assessment, competence, and motivation between early and middle childhood. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 202–221). New York: The Guilford Press.

- Butler, Y. G. (2014). Parental factors and early English education as a foreign language: A case study in Mainland China. *Research Papers in Education*, 29(4), 410–437.
- Butler, Y. G. (2015). Parental factors in the children's motivation for learning English. *Research Papers in Education*, 30(2), 164–191.
- Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessment among Korean elementary school students studying English. *The Modern Language Journal*, 90(4), 506–518.
- Butler, Y. G., & Lee, J. (2010). The effect of self-assessment among young learners. *Language Testing*, 17(1), 1–27.
- Carless, D., & Lam, R. (2014). The examined life: Perspectives of lower primary school students in Hong Kong. *Education 3–13*, 42(3), 313–329.
- Cassidy, S. (2007). Assessing 'inexperienced' students' ability to self-assess: Exploring links with learning style and academic personal control. *Assessment & Evaluation in Higher Education*, 32(3), 313–330.
- Chambers, C. T., & Johnston, C. (2002). Developmental differences in children's use of rating scales. *Journal of Pediatric Psychology*, 27(1), 27–36.
- CILT (The National Center for Languages). (2006). *European language portfolio – Junior version: Revised edition*. Retrieved from http://www.primarylanguages.org.uk/resources/assessment_and_recording/european_languages_portfolio.aspx
- Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.
- Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. New York: Routledge.
- Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty. *Language Testing*, 2, 164–169.
- de Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, 42, 158–178.
- de Saint Léger, D., & Storch, N. (2009). Learners' perceptions and attitudes: Implications for willingness to communicate in an L2 classroom. *System*, 37, 269–285. doi:10.1016/j.system.2009.01.001.
- Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge, UK: Cambridge University Press.
- Dörnyei, Z. (2001). *Motivational strategies in the language classroom*. Cambridge, UK: Cambridge University Press.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.
- Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual Review of Psychology*, 50, 21–45.
- Hargan, N. (1994). Learner autonomy by remote control. *System*, 22, 455–462.
- Harker, D. J. (1998). Definitions and empirical foundations. In D. J. Harker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 1–23). Mahwah, NJ: Lawrence Erlbaum Associates.
- Harris, M. (1997). Self-assessment of language learning in formal settings. *System*, 51(1), 12–20.
- Hasselgren, A. (2003). *Bergen 'Can-Do' project*. Retrieved from <http://blog.educastur.es/portfolio/files/2008/04/bergen-can-do-project.pdf>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effect. *Language Testing*, 7(2), 174–201.
- Hung, Y.-J., Samuelson, B. L., & Chen, S.-C. (2016). The relationships 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*. New York: Springer.
- Kato, Y. (n.d.). Shogakko eigo katsudo-no hyoka-no arikata: jido-ga iyokuteki-ni torikumu jikohyoka-no kufu-to hyoka-no kanten-no meikakuka [How to evaluate English activities at

- primary school: Self-assessment that children are motivated to engage in and the clarification of evaluation criteria*]. Retrieved from <http://www.kyoiku-kensyu.metro.tokyo.jp/09seika/reports/files/kenkyusei/h18/k-31.pdf>
- Leach, L. (2012). Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education*, 37(2), 137–147.
- LeBlanc, R., & Painchaud, G. (1985). Self-assessment as a second language placement instrument. *TESOL Quarterly*, 19(4), 673–687.
- MacIntyre, P., Noels, K., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning*, 47(2), 265–287.
- Malabonga, V., Kenyon, D., & Carpenter, H. (2005). Self-assessment, preparation and response time on a computerized oral proficiency test. *Language Testing*, 22(1), 59–92.
- Marsh, H. W. (1986). Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22, 37–49.
- Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, 26(1), 75–100.
- McDonald, B., & Boud, D. (2003). The impact of self-assessment on achievement: The effects of self-assessment training on performance in external examination. *Assessment in Education*, 10(2), 209–220.
- Mihaljević Džigunović, 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*. New York: Springer.
- Moritz, C. E. B. (1995). *Self-assessment of foreign language proficiency: A critical analysis of issues and a study of cognitive orientations of French learners*. Doctoral dissertation, Cornell University, New York.
- Morrison, F. J., Ponitz, C. C., & McClelland, M. M. (2010). Self-regulation and academic achievement in the transition to school. In S. D. Calkins & M. A. Bell (Eds.), *Child development at the intersection of emotion and cognition* (pp. 203–224). Washington, DC: American Psychological Association.
- National Council of State Supervisors for Languages. (2014). *Lingua Folio®*. Retrieved from <http://www.ncssfl.org/LinguaFolio/index.php?checklists>
- Nicholls, J. G. (1978). The development of the concepts of effort and ability, perception of academic attainment, and the understanding that difficult tasks require more ability. *Child Development*, 49, 800–814.
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- Nicholls, J. G., & Miller, A. T. (1983). The differentiation of the concepts of difficulty and ability. *Child Development*, 54, 951–959.
- Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- O'Malley, J. M., & Pierce, L. V. (1996). *Authentic assessment for English language learners: Practical approaches for teachers*. New York: Longman.
- Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: Tutor and students' perceptions of performance criteria. *Assessment and Evaluation in Higher Education*, 22(4), 357–368.
- Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education*, 25(1), 23–38.
- Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, 6(1), 1–13.
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 175–187). Dordrecht, the Netherlands: Kluwer.

- Oskarsson, M. (1978). *Approaches to self-assessment in foreign language learning*. Oxford, UK: Pergamon Press.
- Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, 19(2), 109–131.
- Piaget, J. (1926/1930). *The child's conception of the world*. New York: Harcourt, Brace & World.
- Pierce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion and locus of control. *Applied Linguistics*, 14(1), 25–42.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experimental factors. *Language Testing*, 15(1), 1–19.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 145–165.
- Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment*, 11(1), 1–31.
- Shameem, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: The Wellington into-Fijians. *Language Testing*, 15(1), 86–108.
- Stefani, L. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education*, 19(1), 69–75.
- Stipek, D. J. (1984). Young children's performance expectations: Logical analysis or wishful thinking? In J. G. Nicholls (Ed.), *Advances in motivation and achievement: Vol. 3. The development of achievement motivation* (pp. 33–56). Greenwich, CT: JAI Press.
- Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment and Evaluation in Higher Education*, 22, 289–305.
- Taras, M. (2002). Using assessment for learning and learning from assessment. *Assessment and Evaluation in Higher Education*, 27(6), 501–510.
- Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning and assessment in the classroom*. Maidenhead, UK: Open University Press.
- Vygotsky, L. S. (1934/1978). *Thought and language* (Rev. and edited by A. Kozulin). Cambridge, MA: MIT Press.
- Zimmerman, B. J. (1989). Models of self-regulated learning and academic achievement. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated and academic achievement* (pp. 1–24). New York: Springer.
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory Into Practice*, 41(2), 64–70.
- Zoller, U., Tsaparlis, G., Fastow, M., & Lubezky, A. (1997). Student self-assessment of higher-order cognitive skills in college science teaching. *Journal of College Science Teaching*, 27, 99–101.

Relationships between Peer- and Self-Assessment and Teacher Assessment of Young EFL Learners' Oral Presentations

Yu-ju Hung, Beth Lewis Samuelson, and Shu-cheng Chen

Abstract As the traditional grammar translation approach is being gradually replaced by communicative approaches, paper-and-pencil tests do not meet the course goals. Thus, the purpose of this study is to investigate how two forms of alternative assessment, peer and self-assessment, can be implemented to evaluate young EFL learners' oral presentations and how the students perceive this experience. The study was conducted with 69 sixth graders (age 12) in Taiwan. The students formed groups of six to discuss and give grades after each individual student's oral report. Three types of data sources included evaluation rubrics, student survey, and a teacher interview. The results show that peer and teacher assessment had strong positive correlation, whereas self- and teacher assessment were moderately correlated. Though learners responded positively to the assessing experiences, they expressed concern that some grades assigned by peers were not fair and a few group members dominated the grading process. The findings shed light on benefits of combining peer and self-assessment and suggest training should emphasize self-assessment, evaluation criteria related to content of the presentation, and students' social skills to work in groups.

Keywords Peer assessment • Self-assessment • Young EFL learners • Observational learning • Social learning theory • Oral presentation

Y.-j. Hung (✉)

Foreign Languages Division, R.O.C. Air Force Academy, Kaohsiung City, Taiwan

e-mail: hung.yuju@gmail.com

B.L. Samuelson

Department of Literacy, Culture and Language Education, Indiana University Bloomington, Bloomington, IN, USA

e-mail: blsamuel@indiana.edu

S.-c. Chen

Sianbei Elementary School, Tainan City, Taiwan

e-mail: fredagotravel@gmail.com

1 Introduction

As the Ministry of Education in Taiwan has listed communication as one of the main objectives of English instruction in elementary school and encouraged alternative assessment (Ye, 2001), learner-centered instruction has started to gain popularity in EFL classrooms. Peer and self-assessment (hereafter PA and SA) are two forms of classroom assessment that involve students' participation to a great extent. PA is "an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners" (Topping, 2010, p. 62). In PA, students judge the work of their peers whereas students judge their own work in SA (Falchikov & Goldfinch, 2000). PA and SA have been found to motivate students and improve their learning (Butler, 2016 in this volume; Dochy, Segers, & Sluijsmans, 1999; Hung, Chen, & Samuelson, *under review*).

PA and SA can be reciprocal. Students' experiences of critiquing and evaluating in PA informs their SA (Topping & Ehly, 2001). On the other hand, SA unavoidably refers to viewpoints and judgments of others (Boud, 1995). Also, a combination of PA and SA has been suggested to prevent over-marking in rating peers and under-marking in students' rating of their own work (Dochy et al., 1999) though the issue of accuracy still remains questionable. We argue that a combination of PA and SA increases agreement between student and teacher assessment and benefits students' learning.

However, few classroom assessment studies that incorporate both PA and SA have been conducted in EFL contexts, particularly for young learners' oral presentation. SA of oral presentation is more difficult to practice, and Harris (1997) has suggested it be supplemented by PA. Therefore, the purpose of this study, grounded in observational learning in social learning theory (Bandura, 1971), is to investigate how PA and SA can be implemented to evaluate young EFL learners' oral presentation and how students perceive this assessment experience. The two research questions are

1. What are the relationships like between peer, self-, and teacher assessment?
2. How do students and the teacher perceive the assessment experience?

2 Observational Learning in Social Learning Theory

This study is situated within the framework of observational learning in social learning theory (Bandura, 1971), later reconceptualized as social cognitive theory (Bandura, 1991). In this framework, human behavior is neither driven by inner forces, nor is it shaped by trial and error, as proposed in the conditioning view. Rather, the causes of behavior are cognitively mediated by means of continuous reciprocal interaction between behavior and environmental forces. New patterns of behavior are the causal consequences arising from cognitively mediating the influences of stimuli of given activities. Among the stimulus determinants, learning first

occurs through direct experience or by observing the behavior of others. Thus, providing an appropriate model of the target learning behavior is indispensable in the process.

2.1 Learning Through Modeling

Social learning theory does not accept that learners simply imitate a model's actions, but that they form new response patterns by organizing behavioral elements they observe. This modeling learning is governed by four processes. The first is attentional processes. Learners select from the model's numerous characteristics and attend to the most relevant ones. Associational preferences are another essential factor. Learners associate with members in their social groups. In other words, learners relate to their peers in classroom settings. The second is retention processes. Verbal coding of the observed information facilitates cognitive processing and storage. Also, rehearsals, or actually performing or mentally rehearsing, enhance long-term retention. The third involves motoric reproduction processes. Learners first acquire symbolic representations of modeled activities; thus, they achieve approximations of the desired behavior. They refine the new patterns of behavior through self-corrective adjustments according to feedback from their own performance. The fourth is reinforcement and motivational processes. Positive feedback or incentives activate the acquired skills to actual performance. Anticipation of positive consequences is one of the best motivators to reinforce and generate an effective, high level of observational learning (Bandura, 1971).

Similarly, students rated their peers' performances based on the criteria in the evaluation rubrics in the present study, so they selectively attended to features of their peers' oral presentations. After each presentation, they discussed and decided the scores on individual assessment criteria as a group. Each group and the teacher then gave oral feedback about the strengths and weaknesses of the presentation. This verbalizing process helped them understand and retain the criteria. The assessing experiences also provided students opportunities for self-reflection by casting themselves in a similar context, a form of mental rehearsal to facilitate their future performance. Afterwards, their SA reinforced their assessment ability for their own presentation and benefited their learning. Self-observation and self-judgment in the process of SA informed learners how well they were progressing toward their goals and motivated behavioral change (Schunk, 2001).

2.2 Functions of Reinforcement

Within the framework of social learning theory, an effective, high level of observational learning of modeled behaviors is shaped and activated by three functions: the informative function, motivational function, and cognitive function. Informative

function of reinforcement indicates learners observe modeled behaviors and conceive what they must do to obtain beneficial consequence. When doing ratings, students reflect by thinking, comparing, contrasting what they observe (Topping, 1998). For motivational function, anticipated consequences and affective factors, such as being empowered to do ratings, serve as best incentives. Cognitively mediated reinforcement offers students opportunities concerning what to selectively pay attention to and what to reward or ignore. Using evaluation criteria and peer group discussion of the criteria reinforce students' understanding of standards of high quality presentations.

3 PA and SA in L1 and L2 Contexts

Relevant studies of PA and SA have been carried out extensively in various fields in L1 higher education contexts, but fewer studies combine both forms of assessment of target oral performance in L2 contexts, especially with young learners. This section reviews PA and SA in higher education first and then narrows the scope to discuss empirical studies incorporating both forms of student-assessment with young learners.

3.1 *Reviews of PA and SA*

The PA process, in which students benefit from social interaction between assessors and assessees, enhances development of cognition and meta-cognition, affect, and social skills (Topping, 1998). Reviews of PA studies find general agreement between student and teacher ratings. Falchikov and Goldfinch (2000) analyzed 48 quantitative studies in L1 settings from 1959 to 1999 and found the mean value of correlation coefficients was 0.69, indicating general agreement between peer and teacher ratings. Consistent with the previous findings, van Zundert, Sluijsmans, and van Merriënboer (2010) reviewed 26 studies of L1 PA from 1990 to 2007 and further pointed out that peer feedback helped students revise their work, higher achievers were more skillful in PA than lower achievers, and students had mixed attitudes toward PA. The problems of friendship marking (Pond, UI-Hag, & Wade, 1995), also referred to "reciprocity effects" (Panadero, Romero, & Strijbos, 2013, p. 195), and insufficient differentiation (Murphy & Cleveland, 1995), where learners gave ratings higher than their peers deserved and tended to give their peers a narrower range of ratings to avoid inaccurate evaluations, were commonly shown in adult learners.

Given opportunities to assess and reflect on their individual progress by engaging in SA, learners focus on their own learning, locate their strengths and weaknesses, and take responsibility for their own learning (Harris, 1997).

The review of SA research shows self-appraisal improves students' achievement, though the correlations for self- and teacher agreement are not as good as for PA

(Blanche & Merino, 1989; Ross, 2006). SA of oral skills is found to be more difficult because speaking can be highly intangible (Harris, 1997). Self-ratings may be affected by subjective errors due to past academic record, and peer or parental expectations (Blanche & Merino, 1989). Cultural factors, such as the pressure to display overt modesty, which is valued in Chinese culture, may make students more critical of their own performance (Chen, 2008; Oscarson, 1997). In contrast, Iranian students are lenient when rating themselves since overt or false modesty concerning one's accomplishments is not emphasized in their culture (Esfandiari & Myford, 2013). Young children tend to over-estimate due to their wishful thinking and lack of the cognitive skills needed to evaluate their abilities accurately (Ross, 2006).

The above reviews show benefits as well as potential problems of PA and SA. Dochy et al. (1999) argued that incorporating both types of student assessment could overcome the defects of over-marking and under-marking. However, the following studies show that this proposal still remains questionable and that additional empirical studies are needed to verify this argument.

3.2 *Combination of PA and SA*

In studies that combine PA and SA of oral performance in L1 universities, student and teacher ratings show disagreement (De Grez, Valcke, & Roozen, 2012; Fallows & Chandramohan, 2001; Langan, et al., 2008) and agreement (Lanning, Brickhouse, Gunsolley, Ranson, & Willett, 2011). The disagreement between student and teacher ratings might be due to different interpretations of evaluation criteria between them (De Grez et al., 2012). Particularly in Asian contexts, low achievers over-marked, and high achievers under-marked. Students' hesitation or lack of confidence in distinguishing their peers' performances resulted in a narrower range of rating their peers. Students also reported that they could not pay full attention to their peers' performance because they needed to do peer-marking while watching the performance (Langan et al., 2008). This result contradicts the findings of previous studies and calls into question the idea of learning from modeling because students are so focused on assessing their peers that they may not be able to observe the performance for the purpose of improving their own presentation.

The positive effects of proper training, involving students in constructing evaluation criteria, providing more opportunities for student assessments, and combining PA and SA with teacher feedback have been shown in other empirical studies in L1 and L2 contexts though the tasks are not on oral performance. Orsmond, Merry, and Reiling (2002) found that using exemplars to discuss criteria helped students understand what was expected. Exemplars could also be used to create agreement between students and teachers, although better agreement was observed between PA and teacher assessment than SA and teacher assessment. Students appeared to be more objective and more focused on product—the presentation itself—when rating their peers, but more subjective and more focused on process—how they prepared for the presentation—when rating themselves. Nevertheless, though benefits of student

assessments were recognized, they should not replace teacher assessment. The appropriate combination of PA, SA, and teacher assessment had the best impact on student learning of assessment skills as well as on target learning outcome (Birjandi & Tamjid, 2012; Murakami, Valvona, & Broudy, 2012).

3.3 *PA and SA with Young Learners*

Student assessment has been found to have a positive effect on young learners' achievement, but an age-related difference appears to be a factor. Ross, Hogaboam-Gray, and Rolheiser (2002) found that 5th and 6th graders who received self-evaluation training had a higher math achievement than who did not. In another study of 6th graders in English class in Korea, repeated SA improved students' assessing ability as well as English performance on objective tests (Butler & Lee, 2010). Butler (1990) compared ratings by children at ages 5, 7, and 10 with adult judges after they copied drawings. Young learners were interested in and capable of comparing drawings using agreed upon standards. However, when the young learners were put in a competitive condition, the desire to outperform others and difficulties in evaluating relative abilities caused inflated perceptions of their own work and decreased their interest.

Butler and Lee (2006) compared 4th and 6th graders' responses to an off-task SA and an on-task SA with teacher assessment and results of standardized tests in Korea. In the off-task SA, learners self-evaluated their general performance in a decontextualized way. The on-task SA was in a contextualized format, in which learners self-assessed their performance in a specific task. The study showed that the validity of SA in the contextualized format was higher than SA in the decontextualized format. The results also indicated that the 6th graders out-performed 4th graders in terms of student assessment accuracy. Though age-differences in SA were found in this study, the reasons behind the differences remained unclear.

In Mok's (2010) study, four secondary students expressed serious concerns that they were not good enough to evaluate their peers, even though they agreed PA helped them reflect upon their own performances. Mok called for preparation of the students both methodologically and psychologically for the role of peer assessor. Hung, Chen, and Samuelson (under review) examined group PA of 4th to 6th graders' oral performance in EFL classes in Taiwan. The results showed that the 5th and 6th graders were able to assess their peers much as their teacher did, whereas the 4th graders were not. The majority of the students in all levels reported they enjoyed playing the role of assessor and indicated this process benefited their subsequent performance and English learning. However, challenges of accepting diverse opinions and conducting discussions of evaluating their peers within groups, particularly for the 4th graders, were indicated.

Though there are some preliminary findings of practicing PA and SA with young learners in the related literature, the effect of combining the two remains uninvestigated and therefore is the main focus of this empirical study.

4 Research Method

This classroom-based research used both quantitative and qualitative data to examine the assessment process as well as the opinions of the students and their teacher. Chen worked collaboratively with two university researchers to plan and implement student assessment procedures in her class. Hung observed all classes in which student assessment was conducted. Samuelson assisted with research data analysis, and her prior experience as an English teacher in southern Taiwan helped her to be familiar with the educational context of the study.

4.1 *Setting and Participants*

The setting for this study was a public elementary school in southern Taiwan. The school was established in 1996 to serve a new high socioeconomic status (SES) suburban community. The total student population was about 800 students, divided into 30 classes (grades 1–6). This school was regarded as a high performing school where the teachers as well as the students had received awards for excellence from the local government and the national Ministry of Education.

Approximately 90 % of the students were Taiwanese; 10 % were Hakka (an ethnic Chinese group comprising 15–20 % of Taiwan's population) or immigrants from provinces in Mainland China or other countries. When the study was conducted, students were required to study English from 3rd grade in elementary school (age 9) in accordance with the national policy. However, local educational policy promoting English proficiency required all students at this school to start English courses from the 2nd grade (age 8).

4.2 *The Teacher and the Students*

Chen held a MA degree of English teaching and had been teaching English at elementary school for 14 years. After attending a workshop on student assessment held by the Ministry of Education, she carried out the PA and SA activities in two 6th-grade classes (age 12). These intact classes were selected because they were taught by the same teacher. Sixty-nine students participated in the study, with three students excluded due to absences. Forty-two were female students and twenty-seven were male. All of the students began learning English in the 2nd grade and received two 40-min English classes every week. In addition to the formal English instruction in elementary school, 58 % of the students (N=40) started to learn English from tutors or in private institutes before entering elementary school, and an additional 16 % of them (N=11) started in 1st grade. Approximately 96 % of the participants (N=66) learned English out of class when this study was conducted.

Based on routine placement tests in the beginning of the semester and the students' final English grades the previous semester, all 6th graders had been divided into advanced, intermediate, and basic levels and separated into different classes. The participants in the current study were assigned to advanced classes. For the purpose of the study, the students were arranged in groups of six for PA. There were twelve peer groups in the two classes, six groups in each class.

4.3 The Classroom Atmosphere

Chen emphasized communicative competence through simple daily conversations. Grammar was not focused on. The students were required to take an oral exam and a written exam to fulfill the course requirement. The instructional approach involved a lot of teacher-student and student-student interaction, role-plays, and English games. Because the majority of the students had also been taught by Chen in 5th grade, they were quite accustomed to these activities and felt comfortable talking and participating in their English class.

4.4 PA and SA Procedures

Training students to ensure they are aware of the objectives and procedures of the assessment and understand evaluation criteria is the key to successful PA and SA activities. Several important steps mentioned in the literature include clarifying the purpose of the kind of assessment done and expectations of the students as assessors; involving participants in developing assessment criteria; providing practice and examples of student performance; providing written checklists or guidelines, specifying activities and timescale; giving feedback; and examining the quality of feedback (Oscarson, 1997; Topping, 2009). Accordingly, the researchers designed the following procedure. The entire procedure of PA and SA lasted seven weeks to complete for each class: two class periods per week and 40 min per class period. After Chen taught the textbook content in each class, she set aside approximately one third of the course time for the student assessment activity. Training took one whole class period. The process writing activity took 3 weeks. Presentations took 3 weeks. Six to eight presentations were done per class.

Step 1. Introducing PA and SA

Chen informed students that PA and SA would be used to evaluate their oral presentations. Students' final grades would include peer, self- and teacher ratings. The purpose and rationale of student assessment were introduced. Students were told that evaluation should be decided from different perspectives, not only by their teacher, but also by their fellow students. When they did PA, they were learning English from others at the same time. They could reflect on their own performance

by rating others and themselves and improve their own future presentation. Chen encouraged the students to take responsibility for the process and learn from the assessing process. After she introduced PA and SA, students moved on to prepare for their oral presentations.

Step 2. Preparing oral presentations

This class used the English textbook, *Enjoy 10*, issued by the local Bureau of Education (Shen et al., 2001). The first unit covered the topic of traveling, after students had just returned from their summer vacation. Chen decided to use “My Summer Vacation” as the presentation topic. Since the English level of this group of students was still at the beginner’s stage, she guided the students to draft their presentation content via process writing. After the students composed draft 1 at home and submitted it to Chen, she indicated the parts that the students could elaborate and taught them how to look up English words online. In the second draft, Chen underlined obvious language errors. In the final draft, she corrected language errors that the students could not revise by themselves. Figure 1 was a final draft by one of the students. In the presentation, the student memorized the content and recited it in front of their classmates.

Step 3. Discussing evaluation criteria

Involving students in the development of evaluation criteria has been recommended in the literature to help learners understand what constitutes a good presentation and to develop a sense of ownership (Harris, 1997; Topping, 2009). Chen discussed the evaluation criteria with the whole class who decided on the criteria together (see Fig. 2). The students agreed that the four criteria should be weighted differently. From Chen’s previous experience of practicing student assessment, students tended to focus on their peers’ weaknesses instead of strengths, so strengths and suggestions were used in the comment to lead the students to pay more attention to their peers’ strengths and give feedback constructively. Finally, she discussed with the students what should be considered the standard for each criterion.

Step 4. Presenting and evaluating

Right before the first presentations, the students reviewed again the evaluation criteria. After each presentation, the audience discussed their classmate’s performance

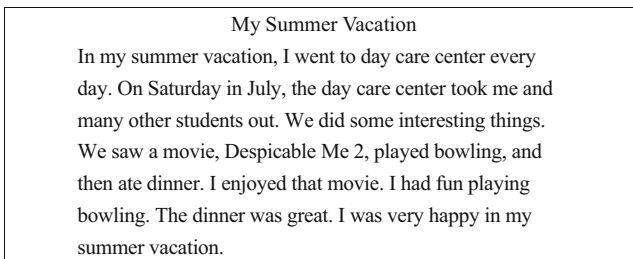


Fig. 1 Student writing sample

Evaluation Rubric	
Voice (6 points)	
Content (6 points)	
Interaction with audience (6 points)	
Body language & facial expression (2 points)	
total (20 points)	
Strength:	
Suggestion:	

Fig. 2 Evaluation criteria

within their groups and assessed their peer by deciding the grades as a group. Meanwhile, each presenting student sat apart and did a SA using the same rubric. Then the teacher and each student group gave oral feedback on the performance. The assessment of all presentations followed the same pattern. Since the students' English abilities were developing, the discussion within groups and in the whole class was conducted in their native language, Chinese.

Step 5. Reflecting

Chen calculated the final scores across groups and compiled all the comments from each group. In the next class, she gave each group its results. She then led the whole class in a reflective discussion on the assessment process.

4.5 Data Sources

In addition to peer, self-, and teacher ratings for each presentation, data included a post-assessment survey filled out by the students and a teacher interview. The survey items and their Chinese translations were examined by Chen to establish the content validity, based on the premise that a subject matter expert's judgment of whether a measure includes the appropriate content for the construct it aims to measure is an acceptable way to establish validity (Cohen, & Swerdlik, 2005). Chinese versions of the questionnaire along with a parental consent form were given to the students. Only students who completed both the survey and returned the consent form were included (N=69). The design of the five-point Likert scale questionnaire for the ratings and interactions between assessors and assessees as well as among team members was framed by social learning theory (Bandura, 1971). In addition to students' demographic information, the items were constructed on the basis of three functions of reinforcement in observational learning, including informational function (Items 1–7), motivational function (Items 8–11), and cognitive function (Items 12–16). One open-ended question elicited the students' general reflection on this process (see Table 1).

Table 1 Student survey

Information function
1. I paid more attention to my classmates' presentations when I evaluated them.
2. I learned English from evaluating my classmates' presentations.
3. I learned how to do a good oral presentation from rating my classmates.
4. My classmates' feedback was helpful to my presentation.
5. I could reflect on my own presentation and think how to improve from evaluating myself.
6. I learned how to give clear concrete suggestions from giving my classmates feedback.
7. I learned how to encourage the presenter from giving my classmates feedback.
Motivational function
8. I liked this assessing activity.
9. I could assess my classmates objectively.
10. I could assess myself objectively.
11. My classmates could assess me objectively.
Cognitive function
12. The whole-class discussion of evaluation criteria helped me understand how to prepare my oral presentation.
13. Each member had a chance to express their own opinions in group discussions.
14. My group members accepted each other's opinions in group discussions.
15. My opinions were accepted in group discussions.
16. I accepted my group members' opinion in group discussion.

The semi-structured teacher interview probed the teacher's perceptions of this assessment practice. The questions included the benefits and difficulties she encountered and how she would expect it to be modified in future classes. The interview was recorded and transcribed.

4.6 Data Analysis

4.6.1 Rubric Data

Paired samples t-tests were used to compare differences between mean scores of peer, self-, and teacher ratings to reveal whether students' perception of their performance accorded with their teacher (Isaac & Michael, 1995). Correlation was used to analyze agreement of total scores and scores on each evaluation criterion between peer, self-, and teacher ratings. Agreement was confirmed if the peer or self-ratings lay within one standard deviation of the teacher's ratings (Kwan & Leung, 1996). The maximum and minimum scores of PA, SA, and teacher assessment were also compared to examine the range of their ratings.

4.6.2 Questionnaires

Descriptive analysis was used to tabulate numbers, percentages, and mean scores of the results of the questionnaires. Cronbach's alpha coefficient for the 16 items is .873, suggesting high reliability of the questionnaire.

4.6.3 Open-Ended Question

Students' responses to the open-ended question in the survey and the teacher interview were coded using the three functions of reinforcement of observational learning given above. Hung and Chen coded all the data independently. A Kappa measure of the two raters' coding was greater than 0.85, indicating acceptable inter-rater reliability (Landis & Koch, 1977). Agreement on each coding was reached through discussion.

5 Results

We present our findings in terms of each of the research questions given at the beginning of this article. The peer, self-, and teacher ratings are used to show the correlations between their evaluations, and the student survey and teacher interview are used to delineate their perceptions.

5.1 Agreement of PA, SA, and Teacher Assessment

The analyses of PA, SA, and teacher assessment reveal peer, self-, and teacher ratings were correlated to a certain extent in the present study. Over-marking, under-marking, and range restriction, which appeared in previous studies of PA or SA, did not exist in this study. As Table 2 shows, the ranges of peer and self-ratings are 9–20 and 7–20, respectively; whereas the range of teacher rating is 12–20. The ranges of both peer and self-ratings are larger than the teacher ratings. The mean differences between peer and teacher ratings and between self- and teacher ratings lay within one standard deviation of the teacher ratings, which indicates agreement between peer and teacher ratings as well as self- and teacher ratings (Kwan & Leung, 1996).

Table 2 Descriptive statistics for peer, self-, and teacher ratings

	Mean	Std. deviation	Minimum	Maximum	N
Peer rating	16.51	1.61	9	20	69
Self-rating	16.09	2.51	7	20	69
Teacher rating	16.66	2.15	12	20	69

Though the mean scores of peer- and self-ratings are slightly lower than the mean score of the teacher ratings, paired sample t tests reveal no significant differences between peer and teacher ratings ($p > .05$) and between self- and teacher ratings ($p > .05$). As displayed in Table 3, the Pearson correlation coefficient between peer and teacher ratings is .73 ($p < .01$), while the correlation coefficient between self- and teacher ratings is .48 ($p < .01$). A correlation of 0.5 is large, 0.3 is moderate, and 0.1 is small (Cohen, 1988). The results show that PA and teacher assessment had a strong positive correlation, whereas the correlation between SA and teacher assessment was moderate and positive. Both correlations were significant.

In the interview, the teacher stated she had noticed the difference between PA and SA. She speculated some students might have over-marked themselves because they were more aware of and took into account their effort. Chen thought that the students' self-assessment of their effort was a good supplementation to other assessments, since it was difficult for the instructor to evaluation the students' preparation process. As she stated in the interview,

When a student rated their peers' performance, he watched the performance of the student critically. When the presenter evaluated himself, he thought 'How much effort did I put into

Table 3 Correlation between peer, self-, and teacher ratings

Total			
		Peer	Self
Teacher	Pearson correlation	.73**	.48**
	Sig. (2-tailed)	.00	.00
	N	69	69
Voice			
		Peer	Self
Teacher	Pearson correlation	.76**	.49**
	Sig. (2-tailed)	.00	.00
	N	69	69
Content			
		Peer	Self
Teacher	Pearson correlation	.44**	.34**
	Sig. (2-tailed)	.00	.00
	N	69	69
Interaction with audience			
		Peer	Self
Teacher	Pearson correlation	.60**	.28*
	Sig. (2-tailed)	.00	.02
	N	69	69
Body language and facial expression			
		Peer	Self
Teacher	Pearson correlation	.40**	.25*
	Sig. (2-tailed)	.00	.04
	N	69	69

this? How was my performance from my point of view?' He evaluated his own performance from his own perspective, not from the perspective of an outsider. I could compare the differences of the evaluations from two perspectives. (Teacher Interview)

當學生幫學生打分數時，他們站的是比較批判的角度去看這個學生的表現，可是如果是站在這個 presenter 的角度，我從中努力了多少，我看到我自己的表現是多少，那我想要看中間的不同點，而不是只是站在 outsider 的角度，要從我自己的角度去看。

Unlike previous studies that indicated students tended to under-mark themselves because modesty was valued in Chinese culture (Chen, 2008), only a small number of students under-marked themselves in this study, and that may have been partially because they set high standards for themselves. The teacher also commented on this phenomenon in her interview,

When judging oneself, one always knew all the hard work done and this prompted a student to rate him/herself more generously. Few students marked themselves really low. These were special cases. They gave themselves really low grades, but their performances were very good according to the teacher's scores. This might be because they had high expectations of themselves. It might also reflect their desire to display humility about their accomplishments. But these were the minority. From what I observed, most of the students did not rate themselves very differently from their peers' evaluations of them. (Teacher Interview)

有些學生會把自己的分數打得比較高，中間的差異性在於他知道自己付出多少，有些學生把自己的分數打得特別低，這幾個特別 case，打得特別低的，但表現得很好的，就是自我要求很高，就是他很習慣性的 too humble，但是這種情況比較少，大部分打出來的，觀察之下，其實跟同儕打的差距性也沒有很大。

Table 3 also shows correlations between peer, self-, and teacher ratings for each evaluation criterion. Though all of the criteria are positively correlated between peer, self-, and teacher correlation, slight differences exist in correlation between PA and teacher assessment. For the criteria of voice and interaction with audience, PA and teacher assessment are strongly correlated ($r=.76$ and $r=.60$); in contrast, the correlations of content and body language and facial expression are relatively weak ($r=.44$ and $r=.40$). The criteria of voice and interaction with audience are probably easier to observe and evaluate. For the content, the students might not have comprehended their peers' presentation completely or they might have had different standards from the teacher. The total number of points for the criterion of body language and facial expression was only 2, which may also help to explain the weak correlation.

5.2 Mutual Reinforcement Functions of PA and SA

5.2.1 Informative Function

The students clearly recognized what they had learned from the assessing activity. Approximately 95 % of the students strongly agreed or somewhat agreed that they paid attention to their peers' presentation, learned some English because of it,

Table 4 Informative function of reinforcement

	Strongly agree	Somewhat agree	Neutral	Somewhat disagree	Strongly disagree	Total
1. I paid more attention to my classmates' presentations when I evaluated them.	72.46 % 50	24.64 % 17	1.45 % 1	1.45 % 1	0.00 % 0	100 % 69
2. I learned English from evaluating my classmates' presentations.	55.07 % 38	39.13 % 27	1.45 % 1	4.35 % 3	0.00 % 0	100 % 69
3. I learned how to do a good oral presentation from rating my classmates.	78.26 % 54	17.39 % 12	1.45 % 1	2.90 % 1	0.00 % 0	100 % 69
4. My classmates' feedback was helpful to my presentation.	72.46 % 50	26.09 % 18	0.00 % 0	1.45 % 1	0.00 % 0	100 % 69
5. I could reflect on my own presentation and think how to improve from evaluating myself.	73.91 % 51	21.74 % 15	2.90 % 2	1.45 % 1	0.00 % 0	100 % 69
6. I learned how to give clear concrete suggestions from giving my classmates feedback.	69.57 % 48	24.64 % 17	1.45 % 1	4.35 % 3	0.00 % 0	100 % 69
7. I learned how to encourage the presenter from giving my classmates feedback.	63.77 % 44	30.43 % 21	4.35 % 3	1.45 % 1	0.00 % 0	100 % 69

learned how to do a presentation, and gave and got feedback to improve themselves (see Table 4). As one student stated in the survey,

This was a great activity! By rating our classmates' presentations, we gave ratings, and we also learned to accept others' opinions. When others evaluated us, they gave us some suggestions. Their suggestions made us understand our strengths and weaknesses. We could reflect on our presentations and think how to improve ourselves. It also let us experience doing a presentation in front of others. We improved our performance on the stage. We learned more and more broadly, not just limited to the content of the textbook. (Student 7)

這是很棒的一個活動·透過同學們報告·讓我們為他評分·評分的過程·也讓我們學習接受別人的意見·別人為我們評分時·會給一些建議·同學的建議可使我們了解自己的優缺點·反省自己的報告並且去思考如何改進·也可以讓我們有上台報告的經驗·讓台風變得更好·也使我們學習更多、更廣·不再只有學習課本上的東西而已。

Three of the 69 students reported that they did not learn any English from doing the PA (Item 2), but that they did learn to give suggestions (Item 6). Since these students only experienced this type of student assessing activity once, they might

need practice doing PA and SA before they would be able to identify the long-term improvement in their English abilities. Also, giving concrete suggestions is relatively more difficult than giving ratings and therefore needs more guidance.

5.2.2 Motivational Reinforcement

The majority of the students enjoyed being empowered to be assessors, and therefore they tried to fulfill the responsibilities of assessors and learn to be fair. In Item 8 and Item 9, the students reported they liked the assessing activity and they were able to assess their peers objectively (see Table 5). They knew they were playing the role of a teacher.

When doing peer assessment, I felt like a judge because I could evaluate my classmates. (Student 27)

同儕評分時·我覺得我像個評審一樣·因為可以幫同學評分。

I think peer assessment has to be fair and just. We can't favor a particular classmate because he is a friend. Peer assessment is also a process to test whether I can give ratings in the stance of a teacher, so I think this is a very good activity. (Student 40)

我覺得同儕評分一定要公平·不能因為他是自己的朋友而偏袒他·所以同儕評分是在考驗是否能以老師的立場去評分·所以我覺得這個活動很好。

As Chen mentioned above, she thought most students could assess their peers and themselves objectively whereas only a few of them could not. In Table 5, five students reported that they could not assess themselves objectively (Item 10), and three students reported that they disagreed with the statement that their peers assessed them objectively (Item 11). One student doubted the fairness of PA and their group played safe by giving a restricted range of ratings for all of the presenters:

I don't oppose this activity, but honestly a little more than half of the class didn't take giving ratings seriously. It was always the same students [in the group] doing ratings. Some of the students couldn't get the standard, just like our group. We were terrible in assessing. We gave two thirds of our classmates 16 [out of 20 possible points]. Once the teacher said one presenter was good, they changed the rating to 18. Also, friends and enemies influenced ratings more or less (I am not sure whether my class has this problem or not). (Student 13)

Table 5 Motivational function of reinforcement

	Strongly agree	Somewhat agree	Neutral	Somewhat disagree	Strongly disagree	Total
8. I liked this assessing activity.	56.52 % 39	39.13 % 27	2.90 % 2	1.45 % 1	0.00 % 0	100 % 69
9. I could assess my classmates objectively.	60.87 % 42	34.78 % 24	2.90 % 2	1.45 % 1	0.00 % 0	100 % 69
10. I could assess myself objectively.	53.62 % 37	37.68 % 26	1.45 % 1	4.34 % 3	2.90 % 2	100 % 69
11. My classmates could assess me objectively.	68.12 % 47	21.74 % 15	5.80 % 4	4.35 % 3	0.00 % 0	100 % 69

我並不反對這項活動，但老實說班上半數再多一點的人在打分數上有點隨便，打分數時幾乎都是那幾個在打，部分的人在打分數上找不到標準，像我們那一組打分數有夠兩光，班上三分之二的人都16分，有次老師說她不錯，他們就把16分改為18分，另外朋友和仇人多少影響分數(我還不知道班上有無這個習慣)。

5.2.3 Cognitive Reinforcement

The majority of the students agreed whole-class discussion of evaluation criteria helped them understand how to prepare for their presentations (Item 12) and that they had opportunities to talk about these criteria in their groups (Items 13–16) (see Table 6). The within-group discussions provided them opportunities to cultivate rapport, improve presentation, and assess others accurately, as one student stated during the interview:

I feel group discussion was a very good task because it could build rapport among group members. Most important of all, we could absorb each other's opinions. That helped us do a better presentation. It could also help me to increase accuracy of my evaluation of others. So I think we should have more group discussions. It helped me and others improve our abilities. (Student 66)

我覺得各組討論是一件很好的事情，因為可以培養組員的感情，最重要的事，可以吸收別人的意見，讓報告更完整，還可以增加自己評判別人的精準，所以我覺得應該多做各組討論，讓自己也讓別人提升自己的程度。

Table 6 Cognitive function of reinforcement

	Strongly agree	Somewhat agree	Neutral	Somewhat disagree	Strongly disagree	Total
12. The whole-class discussion of evaluation criteria helped me understand how to prepare my oral presentation.	57.35 % 39	38.24 % 26	2.94 % 2	1.47 % 1	0.00 % 0	100 % 69
13. Each member had a chance to express their own opinions in group discussions.	76.81 % 53	11.59 % 8	1.45 % 1	8.70 % 6	1.45 % 1	100 % 69
14. My group members accepted each other's opinions in group discussions.	71.01 % 49	18.84 % 13	2.90 % 2	7.25 % 5	0.00 % 0	100 % 69
15. My opinions were accepted in group discussions.	60.87 % 42	30.43 % 21	2.90 % 2	2.90 % 2	2.90 % 2	100 % 69
16. I accepted my group members' opinion in group discussion.	78.26 % 54	20.29 % 14	0.00 % 0	1.45 % 1	0.00 % 0	100 % 69

Through discussion, the students learned how to accept diverse opinions and to work together to decide on a rating as a group.

When we gave ratings through group discussion, we learned not to raise or lower the standard because of particular people. (Student 50)

透過組內討論幫同學評分，我們就可以學會如何不因對象而提高或降低評分標準。

Sometimes everyone had different opinions. After discussion, we could give a rating that everyone was satisfied with. (Student 31)

有時候大家意見很不合，但經過討論後，就會討論出大家都滿意的分數。

However, some students did not learn how to participate in and conduct an effective group discussion. A few students reported not every member was given a chance to express their opinions, and that some of them did not accept each other's opinions (Items 13–15) (See Table 6). As Student 38 said, “Some people didn't respect others' opinions. They didn't learn to how to work well with each other.” [有人不尊重別人的意見，沒辦法學會合作。]

6 Discussion

The finding of a strong positive correlation between PA and teacher assessments and the finding of a moderate positive correlation between SA and teacher assessments together imply that PA has a positive impact on SA. This is similar to what was suggested by Topping and Ehly (2001). In the combination of both PA and SA, challenges that appear in either PA or SA alone in the previous studies are overcome. Contrary to previous arguments that young learners are not able to evaluate themselves fairly due to age-related issues of under-development of cognition and wishful thinking (Ross, 2006), this group of learners demonstrated that they were able to conduct PA and SA as their teacher did, at least to a moderate extent. The problems of over-marking and under-marking were minimized, as Dochy et al. (1999) argued, though subjective issues still appeared in a few SA cases and therefore should be discussed and eliminated in training.

As suggested in social cognitive theory, learning is regulated by interaction between external influence and self-directedness (Bandura, 1991). The integration of group PA and SA serves informative, motivational, and cognitive functions to reinforce students' learning to assess and assessing to learn (Bandura, 1991). For the informative function, the reflecting experience was amplified and had a positive impact on students in terms of being an assessor as well as a language learner. In this context combining both PA and SA, the students observed their peers' performance from the perspective of an outsider whereas they scrutinized their own performance from the viewpoint as an insider. The process of comparing, contrasting, and cross-checking the perceptions of an outsider, an insider, and other outsiders crystalized the standard of each evaluation criterion for the students, who therefore benefited from the experience and developed the abilities to be assessors in both PA and SA. Meanwhile, attending to and reflecting on their peers' as well as their own

presentations helped these students' future performance and English learning although it was suggested that more experience with student assessment might be needed by some students to recognize the long-term effects of improving their English abilities (Butler & Lee, 2006). Also, the results suggest students need guidance to interpret feedback, so they can bridge the connection between feedback obtained and their work to improve their future performance (Sadler, 1998).

As to the motivational function, playing the role of the teacher motivated the students to become fair assessors. The concept of the authoritative role of teachers in Chinese culture empowered the students when they accepted ownership of classroom assessment, and this served as the best motivation to learn to assess fairly, just as a teacher would. Nevertheless, the traditional authoritative role of the teacher is a double-edged sword. Besides inspiring the students to be competent assessors, the teacher's role affected the students' judgment of their peers' performance. One student indicated that his group changed the score they had decided on in order to conform to the teacher's opinion. In other words, the teacher might still dominate the assessing process, and the teacher was likely to be viewed as the only standard in the classroom. As the power of assessment was surrendered by the teacher to the students, and the classroom culture moved from being teacher-center to student-center.

In terms of the cognitive function, the students applied the evaluation criteria that they agreed on to evaluate and reflect on their classmates' performances and then to improve their own presentations. Students' familiarity with the criteria enhances the validity (Falchikov & Goldfinch, 2000). Furthermore, discussion within groups enabled the students to share opinions with each other and analyze their observations collaboratively. Peer-assisted learning has been found to foster social interaction and develop interpersonal skills (Topping & Ehly, 2001), but learning from collaboration should not be taken for granted. Students need help in carrying out exploratory talk to try out and re-organize ideas and therefore benefit from talking to learn (Wells & Wells, 1984).

It is also noteworthy that the incorporation of PA and SA helped the teacher to understand the students' learning and made the assessment more comprehensive than teacher assessment alone or either one of the student assessments by itself. From PA, the perceptions of the majority of the students could be told from their grades, written comments, and oral feedback, all of which deepened the teacher's understanding of whether or to what extent the students knew the criteria of high-quality performance. SA revealed each student's own point of view regarding his or her performance and the effort put into the preparation of the performance. As the teacher pinpointed in her interview, not only the product but also the process should be valued, and she appreciated SA uncovering what she could not tell from the student's performance as product only.

The reciprocal nature of integrating PA and SA in the present study sheds light on the feasibility of implementing student assessment with young EFL learners. Being aware of students' traditional culture and avoiding romanticizing democratic practice of collaborative discussion empower every student, foster autonomy, and orient the learning and assessing process towards learner-centeredness.

6.1 Limitations and Suggestions for Future Research

Firstly, although the results of this study shed light on benefits of combining PA and SA, future experimental research can be undertaken to compare practice of both forms of assessment with each individual form. Secondly, the validity of SA appeared to be lower than that of PA; thus, how to facilitate learners to self-assess their performance needs to be investigated. Finally, differences between group and individual implementation of PA and SA can be examined for future practitioners to successfully implement various approaches to using PA and SA in their classrooms.

References

- Bandura, A. (1971). *Social learning theory*. New York: General Learning Press.
- Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes*, 50, 248–287.
- Birjandi, P., & Tamjid, N. H. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment and Evaluation in Higher Education*, 37, 513–533.
- Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39, 313–338. doi:10.1111/j.1467-1770.1989.tb00595.x.
- Boud, D. (1995). *Enhancing learning through self-assessment*. London: Kogan Page.
- Butler, R. (1990). The effects of mastery and competitive conditions on self-assessment at different ages. *Child Development*, 61, 201–210. doi:10.1111/1467-8624.ep9102040554.
- 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*. New York: Springer.
- Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal*, 90, 506–518. doi:10.1111/j.1540-4781.2006.00463.x.
- Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, 27, 5–31. doi:10.1177/0265532209346370.
- Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, 12, 235–262. doi:10.1177/1362168807086293.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurement*. Boston: McGraw Hill.
- De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education*, 13, 129–142. doi:10.1177/1469787412441284.
- Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, 24, 331–350. doi:10.1080/03075079912331379935.
- Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing*, 18, 111–131. doi:10.1016/j.asw.2012.12.002.

- Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*, 287–322.
- Fallows, S., & Chandramohan, B. (2001). Multiple approaches to assessment: Reflections on use of tutor, peer and self-assessment. *Teaching in Higher Education, 6*, 229–246. doi:[10.1080/13562510120045212](https://doi.org/10.1080/13562510120045212).
- Harris, M. (1997). Self-assessment of language learning in formal settings. *ELT Journal, 51*, 12–20. doi:[10.1093/elt/51.1.12](https://doi.org/10.1093/elt/51.1.12).
- Hung, Y.-J., Chen, S.-C., & Samuelson, B. L. (under review). *Peer assessment of oral English performance in a Taiwanese elementary school*.
- Isaac, S., & Michael, W. (1995). *Handbook in research and evaluation for education and the behavioral sciences* (3rd ed.). San Diego, CA: Educational and Industrial Testing Services.
- Kwan, K.-P., & Leung, R. (1996). Tutor versus peer group assessment of student performance in a simulation training exercise. *Assessment and Evaluation in Higher Education, 21*, 205–214.
- Landis, J. R., & Koch, G. D. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
- Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., & Wheeler, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment and Evaluation in Higher Education, 33*, 179–190. doi:[10.1080/02602930701292498](https://doi.org/10.1080/02602930701292498).
- Lanning, S. K., Brickhouse, T. H., Gunsolley, J. C., Ranson, S. L., & Willett, R. M. (2011). Communication skills instruction: An analysis of self, peer-group, student instructors and faculty assessment. *Patient Education and Counseling, 83*, 145–151. doi:[10.1016/j.pec.2010.06.024](https://doi.org/10.1016/j.pec.2010.06.024).
- Mok, J. (2010). A case study of students' perceptions of peer assessment in Hong Kong. *ELT Journal, 65*, 230–239. doi:[10.1093/elt/ccq062](https://doi.org/10.1093/elt/ccq062).
- Murakami, C., Valvona, C., & Broudy, D. (2012). Turning apathy into activeness in oral communication classes: Regular self- and peer-assessment in a TBLT programme. *System, 40*, 407–420. doi:[10.1016/j.system.2012.07.003](https://doi.org/10.1016/j.system.2012.07.003).
- Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks, CA: Sage.
- Orsmond, P., Merry, S., & Reiling, K. (2002). The use of exemplars and formative feedback when using student derived marking criteria in peer and self-assessment. *Assessment and Evaluation in Higher Education, 27*, 309–323.
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (Vol. 7, pp. 175–187). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation, 39*, 195–203. doi:[10.1016/j.stueduc.2013.10.005](https://doi.org/10.1016/j.stueduc.2013.10.005).
- Pond, K., UI-Hag, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovation in Education and Training International, 32*, 314–323.
- Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research and Evaluation, 11*, 1–13.
- Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5–6 mathematics effects on problem-solving achievement. *Educational Assessment, 8*, 43–59.
- Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in Education: Principles, Policy and Practice, 5*, 77–85.
- Schunk, D. H. (2001). Social cognitive theory and self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Shen, C., Lin, F., Xu, Y., Guo, W., Guo, F., Chen, M., ..., & Liu, S. (2001). *Enjoy 10*. Tainan, Taiwan: Bureau of Education of Tainan City Government.

- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276.
- Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*, 20–27.
- Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*, 339–343.
- Topping, K. J., & Ehly, S. W. (2001). Peer assisted learning: A framework for consultation. *Journal of Educational and Psychological Consultation, 12*, 113–132.
- van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*, 270–279.
- Wells, G., & Wells, J. (1984). Learning to talk and talking to learn. *Theory Into Practice, 23*, 190–197.
- Ye, X. (2001). Alternative assessment. In Y. Shi (Ed.), *English teaching and assessment in primary and middle schools* (pp. 42–73). Taipei, Taiwan: Ministry of Education.