# 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

M. Nikolov (ed.), *Assessing Young Learners of English: Global and Local Perspectives*, Educational Linguistics 25, DOI 10.1007/978-3-319-22422-0_6

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 (Bredekamp

& 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 |
| | | selects specific information from texts, referring to young learner dictionaries or encyclopaedias as required |
| | knows all letters of the alphabet, reads and understands short and simple texts | 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 (*Family at home*) | 11 |
| | 3 | True/False (*In the park*) | |
| | 4 | True/False (*In the classroom*) | |
| Reading 1 | 1 | Multiple choice | 18 |
| | 2 | | |
| Reading 2 | 1 | Picture and text matching (*The story of cat and mouse*) | 10 |
| | 2 | Picture and text matching (*Computer*) | |
| | 4 | Picture and text matching (*TV*) | |
| Reading 3 | 1 | Title and text matching (*Too many sweets*) | 5 |
| | 2 | Title and text matching (*Play with animals every day*) | |
| | 4 | Title and text matching (*Holiday hobby*) | |

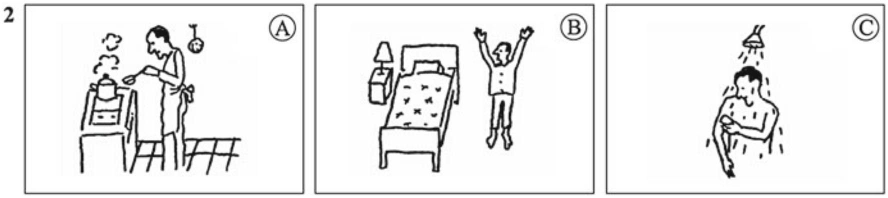
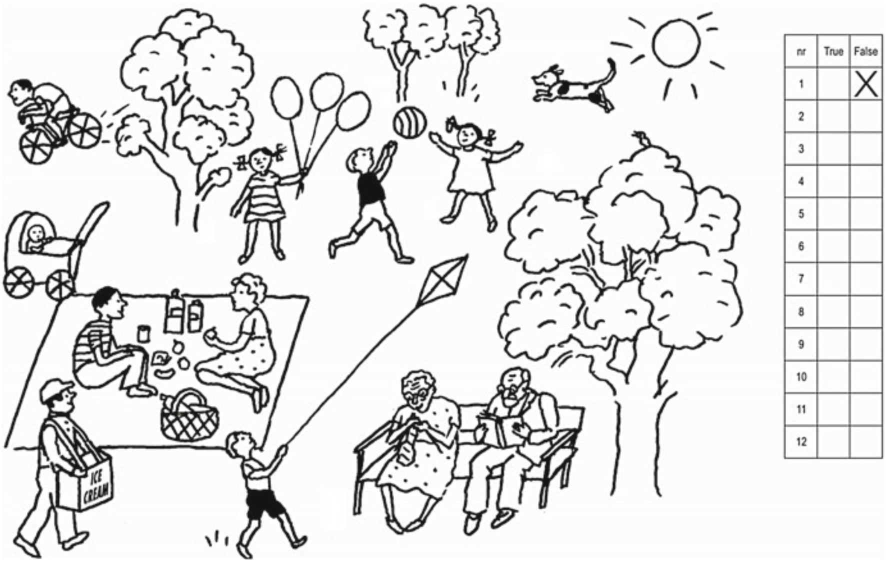


**Fig. 1** Example of listening comprehension items in task 1: multiple choice

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*)

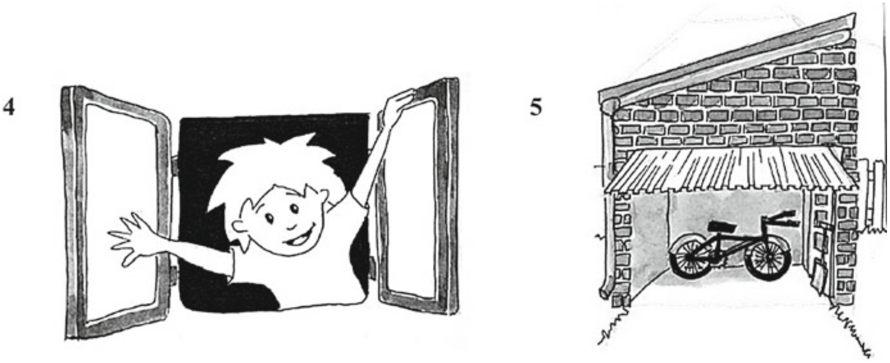
Translation of the instruction in Polish*: Look carefully at the scene. Listen to the sentences or brief dialogues and mark the appropriate box according to whether what you hear is True or False with a cross (x). You will hear the recording twice.*

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.*



□ What a boring game!

□ What a lovely day!

□ What terrible weather!

□ Go and brush your teeth!

□ Dinner's ready!

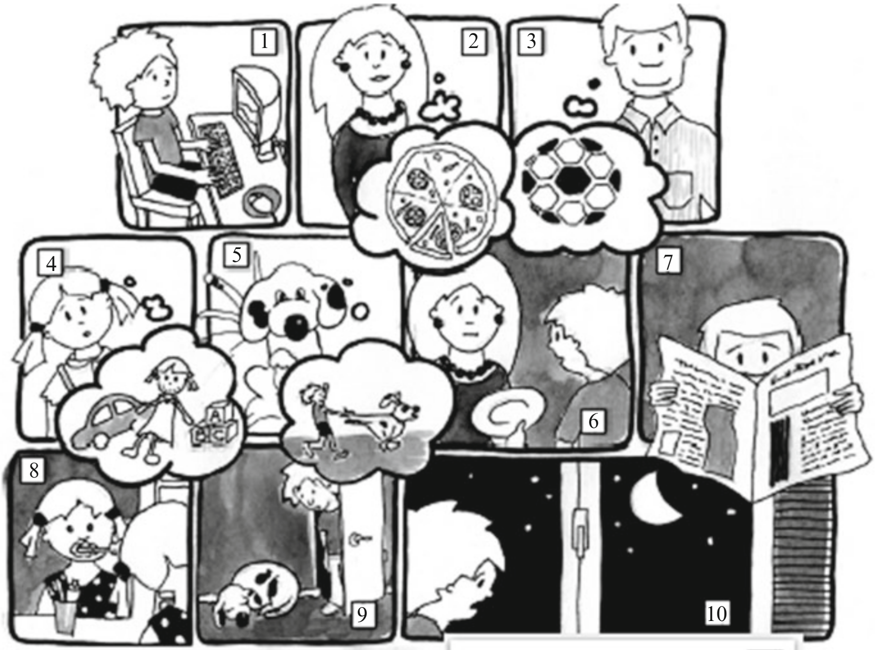
□ Let's go for a ride!

**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 *washback* 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.



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

Spot, the dog, wants to go for a walk with Tom. □

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

Tom is at home. He is playing on the computer. 1

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

It is time to go to bed. □

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

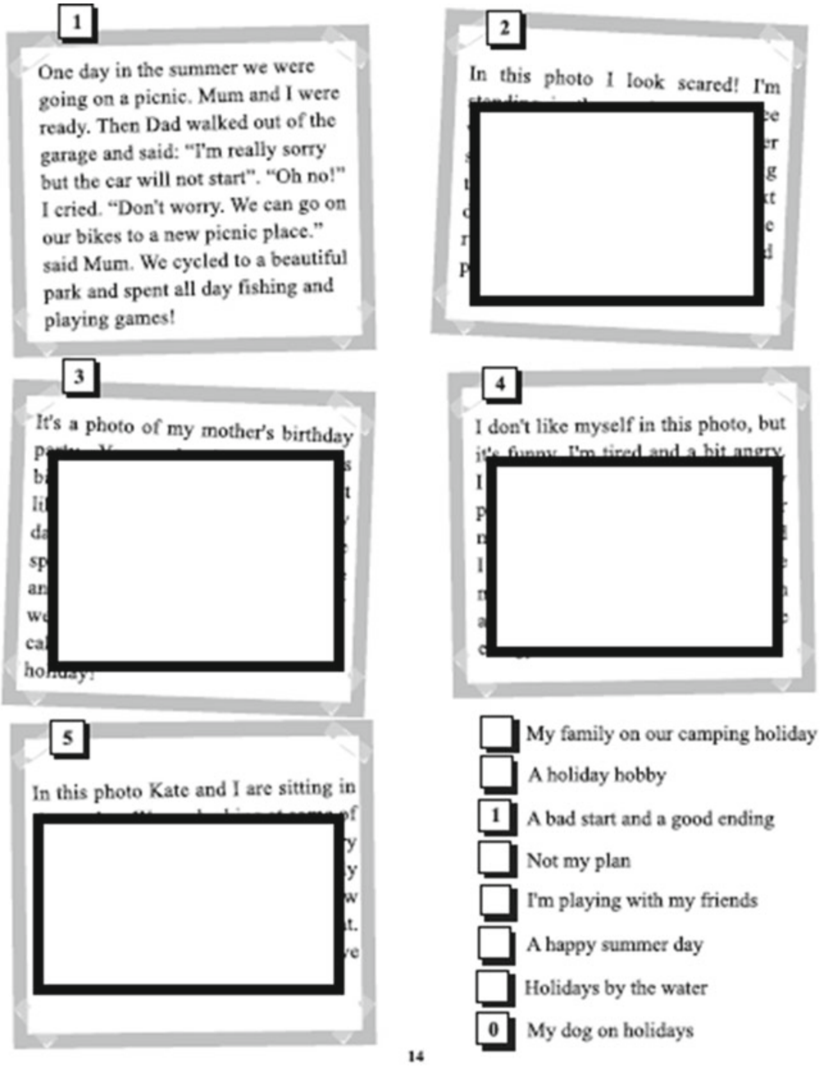
Anna is brushing her teeth. It's too late to play. □

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

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

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 [redacted]

3

It's a photo of my mother's birthday [redacted]

4

I don't like myself in this photo, but [redacted]

5

In this photo Kate and I are sitting in [redacted]

□ 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 Ⓐ łatwe Ⓑ ani łatwe ani trudne Ⓒ trudne

2 Ⓐ znane Ⓑ trudno powiedzieć Ⓒ nieznane

3 Ⓐ podobało mi się Ⓑ nie wiem Ⓒ 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

| | Test version | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | 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 |

[a] Numbers 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)

| | Test version | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task | 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.

```
          PERSON - MAP - ITEM
              <more>|<rare>
   4           ###  +
                    |
                    |
                    |
                    |
                    |
                 .  |
                    |
   3             .  +
                    |
                    |
                    |
               .##  |
                    |
                 .  |
                 . T|
   2                +  L1V1Z8B
                ##  |
                 .  |T
                    |
              .###  |  L1V1Z10B
                 .  |  L1V1Z5B
                 . S|
            .#####  |
   1             .  +
             #####  |S
                 .  |
           .######  |
                 .  |  L1V1Z20B
           .######  |
                 . M|  L1V1Z19B  L1V1Z2B   L1V1Z9B
        .#########  |  L1V1Z6B
   0             .  +M L1V1Z13B  L1V1Z18B
      .###########  |  L1V1Z17B
                 #  |
           .######  |  L1V1Z7B
                 .  |  L1V1Z15B
              .###  |  L1V1Z11B  L1V1Z12B
                 . S|
               ###  |S L1V1Z14B
  -1             .  +  L1V1Z16B  L1V1Z4B
                 .  |
                .#  |  L1V1Z3B
                 .  |
                .#  |
                 . T|
                    |T
                 .  |
  -2             .  +
                    |
                    |
                 .  |
                    |
                    |
                 .  |
                    |
  -3            .#  +
              <less>|<frequ>
EACH "#" IS 5. EACH "." IS 1 TO 4
```

**Fig. 7** IRT person/item map analysis of L1v1 items

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.

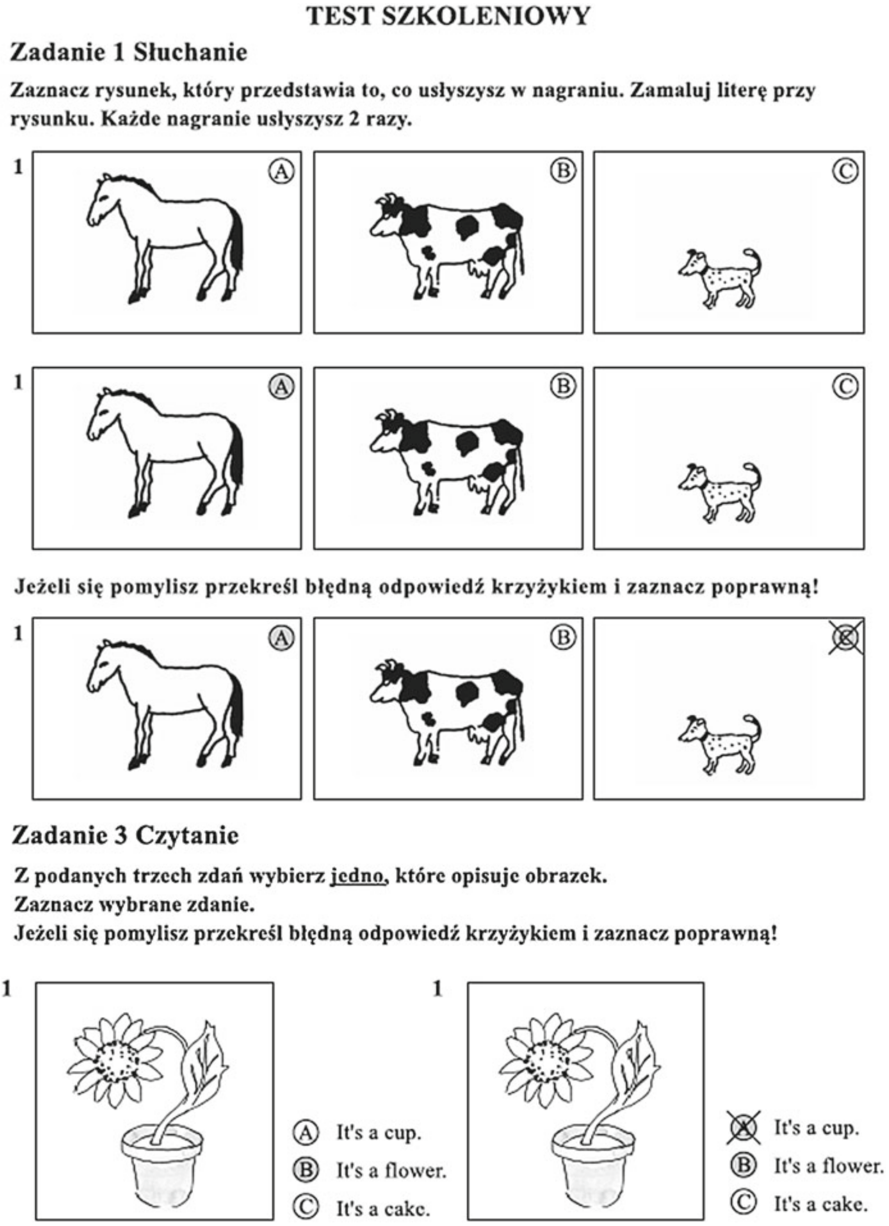
Several aspects of the design process need to be particularly emphasised. The first is the careful analysis of items using IRT to ensure a choice with the best psychometric qualities. The second is the enormous value of cognitive laboratories to obtain young learners' perspectives on planned tests. These 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 Ⓐ Ⓑ Ⓒ

1 [✓ Ⓐ] Ⓑ Ⓒ

**Jeżeli się pomylisz przekreśl błędną odpowiedź krzyżykiem i zaznacz poprawną!**

1 [✓ Ⓐ] Ⓑ ~~Ⓒ~~

**Zadanie 3 Czytanie**

**Z podanych trzech zdań wybierz <u>jedno</u>, które opisuje obrazek.**
**Zaznacz wybrane zdanie.**
**Jeżeli się pomylisz przekreśl błędną odpowiedź krzyżykiem i zaznacz poprawną!**

1

Ⓐ It's a cup.
[✓ Ⓑ] It's a flower.
Ⓒ It's a cake.

1

~~Ⓐ~~ It's a cup.
[✓ Ⓑ] It's a flower.
Ⓒ 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.

Bredekamp, 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.