



Fangfang Du* and Jens F. Beckmann

Dynamic testing of language learning aptitude: an exploratory proof of concept study

<https://doi.org/10.1515/jccall-2023-0028>

Received November 29, 2023; accepted January 19, 2024; published online March 5, 2024

Abstract: This paper provides an example of how online technology can be utilised to increase efficiency and validity of assessment procedures beyond simple computerization of testing. We report the first steps of the development of an online assessment procedure for the measurement of language learning aptitude that is based on the concept of dynamic testing. Online adaptive dynamic tests provide learning opportunities in the test situation, allowing test takers demonstrate their ability to benefit from feedback. Test performance shown in such test procedures has been demonstrated to be more useful in predicting future learning and to represent a more valid indication of test takers' learning potential. The newly developed online assessment procedure aims at measuring young learner's ability to improve performance based on inductive reasoning across lexical, morphological and syntactic rules of an artificial language using pictorial stimuli. The test was administered on ten mainland Chinese children aged between 9 and 13. The tentative analysis indicates that the newly developed test is feasible and has the potential to be a useful diagnostic tool for measuring language learning aptitude. We report and discuss insights gained and outline how these will be utilised for the further refinement of this online assessment tool.

Keywords: dynamic testing; language learning aptitude; online adaptive test

1 Introduction

Learning a second language is important for adapting to today's society. Learners differ in their aptitude to learn an additional language. Some seem to be able to learn languages easier, faster, and more effectively than others (Brown, 2014).

*Corresponding author: Fangfang Du, Durham University, Durham, UK,

E-mail: fangfang.du@durham.ac.uk. <https://orcid.org/0000-0002-7711-2428>

Jens F. Beckmann, Durham University, Durham, UK, E-mail: j.beckmann@durham.ac.uk. <https://orcid.org/0000-0002-4006-9999>

Psychological research has a long tradition in describing and explaining individual differences in general, and in foreign language learning in particular. Part of these research efforts is the development of theories and assessment tools related to language learning (Grigorenko et al., 2000). For instance, Carroll and Sapon devised the *Modern Language Aptitude Test* (MLAT, Carroll & Sapon, 1959) by using an artificial language to assess learner's auditory ability, phonetic coding ability, grammatical sensitivity, memory, and inductive language learning to predict a person's likely success in learning a foreign language. In 1966, Pimsleur introduced *The Pimsleur Language Aptitude Battery* (PLAB, Pimsleur, 1966) comprising motivation scales, verbal ability scales (vocabulary and language analysis), and auditory ability scales (sound discrimination and sound-symbol association). Other well-known language aptitude tests include the *Army Language Aptitude Test* (ALAT, Horne, 1971), and the *Defense Language Aptitude Battery* (DLAB, Perterson & Al-Haik, 1976). These instruments are rooted in psychometric traditions, are focussed on English language learners, and aim at predicting language learning outcomes (Sternberg & Grigorenko, 2002).

Even though these tests have been extensively discussed in research literature and are widely used by language educators and administrators (Brown, 2014), they are not without problems. In this paper we start with highlighting some of those problems. We then present a proposal on how some of these problems can be overcome. Finally, we will present an example for a test procedure that – as we will argue – allows for a more adequate operationalisation of language learning aptitude.

1.1 Three problems in contemporary language aptitude testing

The first problem related to language learning aptitude tests relates to the claim of measuring “aptitude”. To elaborate on this problem, we briefly revisit Snow's work on aptitude and its testing. Snow (1992) argued in his aptitude theory that “the concept is especially close to readiness, but also to suitability, susceptibility, and proneness. ... The common thread through these and other related terms is potential.” (Snow, 1992, p. 6). However, “... applied psychologists increasingly ignored the substantive roots of the concept. Aptitude became nothing more than the predictions made from conventional ability tests” (Snow, 1992, p. 7). As has been highlighted before, the problem is that the test procedure employed in conventional aptitude tests does not allow test takers to demonstrate their susceptibility to feedback to improve their performance, i.e., to show their ability to learn. In short, conventional tests do not measure what they claim to measure.

The second, and related problem is that conventional tests measure performance that is interpreted as a result of a learning process. As Beckmann (2014) argues, performance levels shown in a conventional ability test can be the product of

qualitatively different processes. For instance, low performance could be indicative of a low level of aptitude, but it also could be a result of a lack of learning opportunities in a test taker's developmental or educational past. If we cannot be certain about the processes that led to the performance levels observed in ability tests, any inferences drawn based on test scores in regard to a test taker's aptitude will be uncertain too. Such uncertainty is bound to also affect the confidence with which one might want to predict future performance (i.e., one of the practical purposes of language ability testing in terms of selection or admission decisions). In short, unawareness of this issue by simply relying on how a test is labelled, creates the consequential risk of underestimating learner's "true ability to learn" (Beckmann, 2014, p. 34).

The third problem, again a consequence of the one just mentioned, relates to the criticism towards traditional or static test as being potentially biased against ethnic minorities or social-economically disadvantaged learners (Elliott, 2003; Haywood & Lidz, 2006). As Sternberg and Grigorenko (2002) and others have argued, test performance in ability tests is multi-causally determined. Contributing factors include, educational environment, developmental opportunities, parental support, test taking skills, and many more. Minority groups are more likely being misjudged based on standardized tests (Utley et al., 1992). Minorities and social-economically disadvantaged tend to be exposed to sub-optimal learning environments when developing cognitive (and other) abilities or skills. Their performance in conventional tests is therefore more likely an indication of their access to resources, rather than an indication of their learning potential, or aptitude.

1.2 A possible solution?

Critical reflections of that nature are anything but new. The idea of dynamic testing has its origins in the early 70s of the 20th century. Some might even argue that it relates back to the early 20s of the same century (see PhD thesis of De Weerd, 1923). Early literature on Dynamic Testing includes: Guthke (1982), Guthke and Harnisch (1986), Guthke et al.(1986), Guthke (1990), Guthke (1992), Guthke and Wingenfeld (1992), Hamers et al.(1994), Guthke et al. (1995), Wiedl et al.(1995), Guthke, et al. (1997), Grigorenko & Sternberg (1998).

In a dynamic testing approach, test takers are provided with learning opportunities as part of the testing procedure itself. These enable the test taker to demonstrate their ability to benefit from feedback and thinking prompts. Test performance shown under these conditions has been demonstrated to be more useful in predicting future learning and to represent a more valid indicator of test takers' learning potential (Beckmann, 2006; Elliott et al., 2018; Lidz & Elliott, 2000).

This article reports on first steps to address the issues briefly outlined above in the context of language learning. To that end we introduce a newly developed test procedure that aims at detecting learners' language learning potential in a more direct way as conventional approaches offer. This test is named *Dynamic Test of Language Learning Aptitude – Chinese (DToLLA – C)*, which combines the concept of Dynamic Testing in psychometric assessment with language testing and benefits from the functionality that a computer-based online test administration offers. The use of computer allows for the application of an adaptive algorithm, that helps tailoring the complexity level of items, the order of items, the number of items, and the level of feedback to the response behaviour of any given test taker. This enables the creation of individualised learning opportunities and allows to register test takers' levels of responsiveness or susceptibility to learning stimulation. This paper reports preliminary results from an explorative proof of concept study related to the first steps in the development process of this test.

1.3 Aims

This exploratory proof of concept study reported here pursued three aims: (1) to identify potential procedural issues, including comprehensibility of instruction and feedback; (2) to establish whether the test materials and test procedures are suitable for the targeted age group; and (3) to gain insights to inform the computerisation of the test.

2 Theoretical framework

2.1 Definition

Dynamic testing is “a methodological approach to psychometric assessment that uses systematic variations of task characteristic and/or situational characteristics in the presentation of test items to evoke intraindividual variability in test performance” (Beckmann, 2014, p. 310; also see Guthke & Beckmann, 2000b; Guthke & Wiedl, 1996; Guthke et al, 2003). The central characteristic of a dynamic testing approach is “the combination of testing and instructional intervention” (Sternberg & Grigorenko, 2002, p. 23). Beckmann (2014) listed intervention features: “feedback, hints, thinking prompts, retries, retesting after training phases, and so forth” (p. 309). The purpose of these interventions is to help test takers to take a step forward from where they are currently struggling (or their developed level of ability) to the area where they can

reach their maximum potential (or their to be developed level of capacity), or in the vernacular of Vygotsky's, their zone of proximal development (ZPD, Vygotsky, 1978).

For contextualisation: Vygotsky defined the Zone of Proximal Development (ZPD) as “the distance between the actual development level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peer” (Vygotsky, 1978, p. 86). This notion aligns with our dynamic testing principles to detect the maximum potential for development through systematic help and intervention. ZPD operates by quantifying the systematic prompts provided during the testing.

Elliott (2003) stated that the degree of standardization of the feedback or hints provided in the instructional interventions depends on the purpose of the dynamic testing. For example, for the purpose of educational selection or resource allocation, standardized feedback systems are utilised to obtain data to inform systematic and meaningful comparisons among test takers (see also Guthke & Beckmann, 2000a). Navarro and Mourgues-Codern (2018) also suggested that standardized interventions lend themselves to be implemented in form of computerised algorithms. The intended function of the newly developed test is to identify learners who have learning potential but do not necessarily perform well in traditional language tests.

The application of dynamic testing in the domain of language learning test is very limited so far. The authors are aware of two examples. Grigorenko and associates have developed a dynamic test of language learning ability, the Cognitive Ability for Novelty in Acquisition of Language (Foreign) test (CANAL-F). This test measures language learners' cognitive processing through immediate and delayed recall of lexical, morphological, semantic, and syntactic information. The test was designed to capture “... the ability to learn at the time of test” (Grigorenko et al., 2000, p. 392). Guthke and Harnisch (1986) have designed a diagnostic program explicitly utilising a Dynamic Testing approach called “Acquisition of Syntactic Rules and Lexis” to measure learners' ability to learn foreign languages.

2.2 Test format

Guthke and Beckmann (2000a) discussed two different forms of dynamic testing procedures in relation to learning tests. They distinguish between “long-term learning tests” and “short-term learning tests”. Long-term learning tests have the structure of a pre-test, followed by an instruction or learning phase, followed by a post-test. The difference in performance between pre and post-test is interpreted as indicative of a test taker's ability to learn. In contrast, short-term learning tests comprise only of one test session in which items serving two different purposes are

combined within the test procedure. Test items fulfil the purpose of testing the current level of capabilities, while training items serve the purpose of providing structured learning opportunities. The performance in the subsequently presented test items is seen as reflecting the accumulated effects of the learning episodes worked through during the test so far. Sternberg and Grigorenko (2002) propose arguably more instructive category labels for the distinction between long-term learning tests and short-term learning tests. The so-called sandwich format refers to long-term learning tests that employ a sequence of pretest-instruction-posttest. The so-called cake format refers to short-term learning tests in which test and training items are mixed within the test procedure.

The newly developed language learning test presented here follows the principle of a short-term learning test, that is one test session in which test items and training items are combined. In such test procedure (see Figure 1), if test takers successfully solve a test item, they will move to the next (usually a more complex) test item. This is not different to a conventional approach in ability testing, where the number of correct responses is aggregated to a test score. Such test score reflects the test takers' zone of actual development (Vygotsky, 1978), or their developed level of ability. In case of an incorrectly solved item, however, a dynamic test procedure deviates from a conventional approach. In a dynamic test procedure, incorrect answers will be followed by the presentations of so-called training items where a graduated system of feedback and thinking prompts will be provided until the test taker is able to provide a correct answer. This creates a form of an intermediating mini training session which is followed by a return to the test items, that, if learning took place will

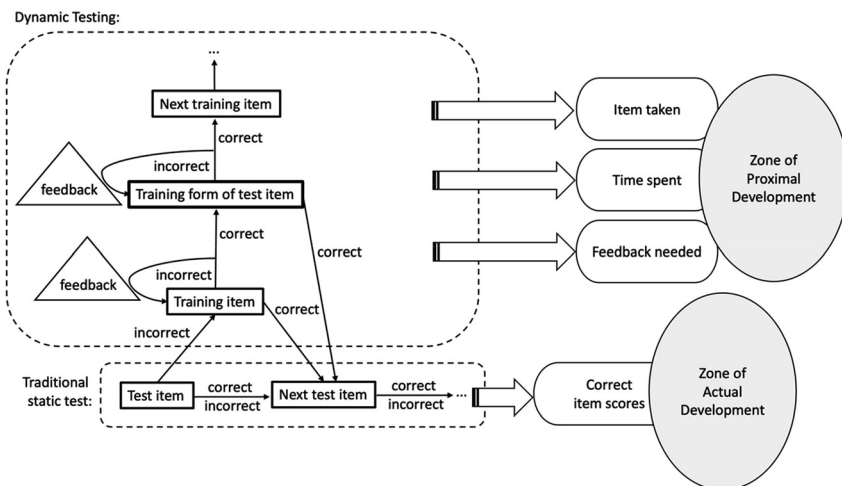


Figure 1: Test functioning mechanism.

then be tackled more successfully than before. In this dynamic testing approach, instead of registering the number of correctly solved items, the test score reflects the amount of training opportunities taken up to progress through the test. The test score therefore represents a test takers' zone of proximal development (Vygotksy, 1978), or their to be developed level of capacity.

2.3 Target population

According to the National Bureau of Statistics' announcement on the seventh Population Census in 2021, China's migrant population reached 375 million, an increase of 69.74 percent since the sixth Population Census in 2010. The number of migrant children in compulsory education reached 14.3 million. The education of these migrant children has been of concern by policy makers, the wider public, and researchers. However, despite years of government efforts to ensure that these children receive equal compulsory education, they remain at a disadvantage on multiple levels. With regard to the school education level, a large portion of migrant children are still denied access to public schools. Although the central government has urged the destination cities to address the education problem of migrant children and to enroll them in public schools, the local government reversely increased the restrictions on the education of migrant children under the consideration of population control (Chen et al., 2019; Liang et al., 2020). As a result, a large number of migrant children are forced to seek alternative provision, often in form of low fee-paying private migrant schools. These school tend to struggle with effective management, lack of recourses, and poor teaching quality (Chen et al., 2019). Migrant children tend to receive less family support than local children in all aspects due to the constraints in their social, educational, and financial capital (Jin et al., 2017). Part of such disadvantage is migrant children's lower quality of parent-children attachment which increases the risk of depression (Shuang et al., 2022), levels of mental health problems, and results in lower life satisfaction (Gao et al., 2015).

All in all, the learning environment of migrant children in migrant schools is very challenging. Using conventional unified assessment tools to assess them will only contribute to masking their potential to learn. As a consequence, poor performance in conventional tests will further reduce their chances to strive in China's exam-oriented education. It is therefore important to employ assessment approaches that are more sensitive to otherwise hidden potential. We consider this as small but meaningful step towards educational equity as it is advocated by the government. In this regard, the DTOLLA – C aims at providing pupils developing under sub-optimal learning conditions a fairer form of assessment.





3 Test content



3.1 Inductive reasoning


Language learning (either first language or second language learning) is largely an inductive reasoning process “in which learners must infer certain rules and meaning from all the data around them” (Brown, 2014, p. 104, see also Sternberg & Gardner, 1983). The newly developed test is designed to measure children’s responsiveness to feedback and learning prompts whilst engaging in processes of inductive reasoning.

Vygotsky (1978) stated that the zone of proximal development is most likely to appear when the child is playing, since this is the time when a child will behave beyond his or her average age, and above his or her daily behavior, or in Vygotsky’s words “a head taller than himself” (p. 102). The activity setting of this test aims to be fun and playful, which is mainly in the form of “figuring out” a secret language using depictions of animals, plants, vehicles, and other common objects in daily life that are of interest and familiar to school-age children. The items comprise two parts, examples and tasks. The examples have pictures and text translations, while the tasks only provide pictures. The examinee needs to solve the task by selecting the correct word or words to describe the pictures from the answer options (see Figure 2 for an example item). The analogical reasoning problems in the test are implemented by asking the examinees to deduce the text translation of the picture given in the task through finding out the similarities and differences between pictures given in the example and the corresponding text translation.

Item 1

Example:  :   : 

 : 

Task:  : _____

Characters to choose:








      

Figure 2: Screen shot of Item 1.

3.2 Language used in the test

The newly developed test makes use of an artificial language, which does not resemble any existing language. This is to ensure fairness and to minimise the potential advantages for test takers who might have had prior experiences or exposure to an otherwise “real” language. The language used in the DTOLLA – C differs from stimulus material in tests designed by many Western or American Scholars in that the languages they utilise are mainly alphabetic word-formation, which is found in Germanic languages such as English, French, German, etc. The language used in the DTOLLA – C test uses morphological word-formation, which is found in Chinese–Tibetan languages such as Chinese, Japanese, etc. Each examinee will learn the language from the simplest single word construction to the composition of phrases, to the expression of whole sentences. Therefore, from the dimension of language, test takers’ abilities in vocabulary, word formation, morphology and syntax will be tested. Topics range from object classifications, number expressions, outline of objects, object state descriptions, location descriptions, action descriptions, etc.

3.3 Item pool

There are 36 items in the item pool, divided into 4 complexity levels, with 9 items in each complexity level. The 9 items were classified into test items and training items. The first three items in each complexity level (e.g., Item 1, 2, 3) are test items, and the last six items (e.g., Item 4, 5, 6 and 1f, 2f, 3f) are training items. Item 1 and 1f have the same content, but the difference is that Item 1 is a test item without feedback assistance. While Item 1f, like Items 4, 5, and 6, is a training item with standardized feedback assistance. The same rules are applied to Item 2 and 2f, and Item 3 and 3f, and so on. In addition, the test item and the corresponding training item (e.g., Item 1 and Item 4) used the same information cue combination rules. More detail descriptions about information cue combination rules are covered in Section 3.4.

3.4 Structures of complexity levels

The structures of complexity levels (CL) in this test are set in two ways (as shown in Table 1). The first way is the increasing complexity of language construct in the items, from the measurement of single word to the measurement of two-word phrase, to the measurement of three-word phrase or sentence, and then to the measurement of

Table 1: The structure of complexity level.

Complexity level	Information cue	Language structure	Item content
1	2	One word	One object
2	3	Two-word phrase	One object and one attribute
3	4	Three-word phrase or sentence	One object and two attributes
4	5	Five-word sentence	Two objects and two attributes


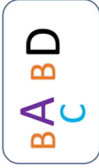


five-word sentence, as to realize the leap from word, phrase, and sentence. The second is the increase of information cues in the task. As shown in Table 1, the information cues increase successively from 2 at the first complexity level to 5 at the fourth complexity level, as to increase the number of logical points for inductive reasoning. More detailed combination rules are shown in Table 2.

The task in items at complexity level 1 (CL1) comprises one object represented by a single character. This character consists of three information cues representing different the object's features, categories, and environment respectively. The combination rules of the information cues in the test items and the corresponding training items are the same, but the combination rules of different test items are different. Table 2 shows one of the combinations for each CL. There are other combinations, but the number of information cues examine in the task is the same for the same CL. To take it more specifically, in CL1, one of the three information cues is controlled and the other two are examined. In the example of Item 1 in Figure 1, the information cue A representing the pig's special feature is controlled, so each character in the answer options have a symbol of “♂”. Information cues B and C, as examination information cues, have different symbols and positions in each answer option. The test taker is required to find out the correct answer from the clues in the example through inductive reasoning (note: the correct answer for Item 1 is ♂♂).

3.5 Online adaptive technology

An effective realisation and utilisation of the features of dynamic testing requires the computerised (i.e., online) implementation of the test. Whilst the test as used in this proof-of-concept study was not yet computerised, we briefly outline how we envision the ultimate version of the test to benefit from an effective utilisation of computer technology. The implementation of online technology will comprise an adaptive algorithm which includes four modules: “item selection algorithm”, “feedback item

Table 2: Examples of information cue combination.

Complexity level	Number of information cues	Information cue combination rules	Language constructions
1	2		Word combination: A Stand for object feature, B Stand for object category, C Stand for object environment
2	3		Subject-predicate phrase: A Stand for object feature, B Stand for object category, C Stand for object environment, D Stand for movement
3	4		Subject-predicate object phrase: A Stand for object feature, B Stand for object category, C Stand for object environment, D stand for movement, E Stand for status
4	5		Sentence that includes adverbials: A Stand for object feature, B Stand for object category, C Stand for object environment, D stand for movement, E Stand for status, F Stand for position

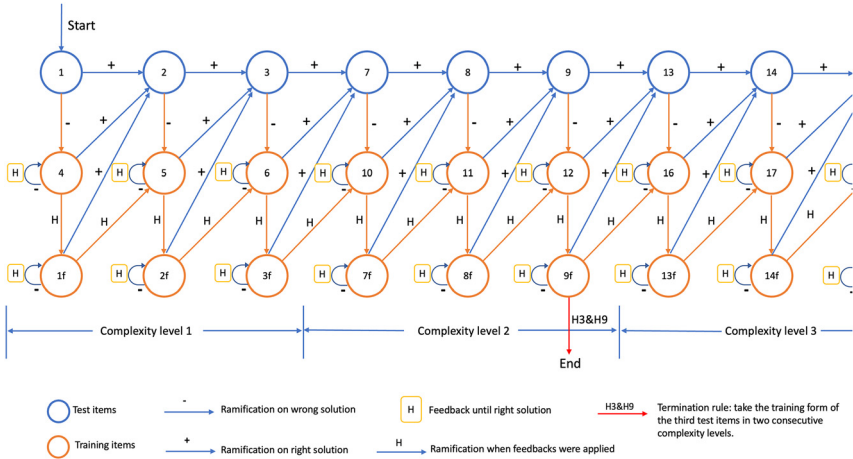


Figure 3: The ramification rules of the test.

selection algorithm”, “termination condition algorithm”, and “scoring algorithm” (The ramification chart is presented in Figure 3). These algorithm modules determine the progression to the respective next complexity level, the order and number of items, the order and amount of feedback based on the test taker’s answers.

The “item selection algorithm” introduces the logic of transition between different types of items and different levels of complexity. As shown in Figure 3, all examinees start with Item 1 in complexity level 1 (CL1). After a correct response the test continues with Item 2. In case of an incorrect answer, however, the corresponding training item will be presented (e.g., from Item 1 to Item 4). The ramification rule stipulates that if the training item is responded to correctly in the first attempt, the test continues with the next test item (e.g., from Item 4 to Item 2). An incorrect response to a training item, however, will be followed up by feedback and thinking prompts. Subsequently another training item will be presented to determine whether more learning opportunities are needed (e.g., from Item 4 to Item 1f; from Item 1f to Item 5).

This adaptivity in terms of item sequence as well as error-specific feedback and thinking prompts is realized across the four complexity levels. For instance, if Item 3 is answered correctly, the test taker proceeds to Item 7 in CL2. Or, if the answer to Item 6 is correct at the first attempt, the test taker “jumps” to Item 7 in CL2. If supportive hints and thinking prompts are needed for Item 3f – because of an incorrect answer to Item 6 previously – the next item to tackle will be Item 10 in CL2, and so on.

Feedback, hints, or thinking prompts are features of training items. The kind of feedback given after an incorrect response to a training item depends on (a) the kind of error that underpins the incorrect answer and (b) the number of preceding incorrect responses. The number of information cues that are to be considered increases from complexity level 1 to complexity level 2. An incorrect answer to a training item in complexity level 2, for instance, can be a result of having ignored one or two information cues. Feedback and thinking prompts refer to the information cue(s) that were apparently missed in providing an incorrect answer.

To mitigate the risk of overwhelming test takers with presenting them with items they might not be able to solve (despite feedback and learning prompts) a termination rule will be implemented in the ultimate version of the test. That is when a test taker has utilised three training items across two consecutive complexity levels then the test is terminated. Test performance is operationalised via the number and type of items tackled, the amount of feedback needed, the level of feedback required, and the time spent taking the test.

3.6 The feedback system



Campione and Brown (1987) realised a graduated feedback approach to provide a scaffolding framework for test takers so that they were enabled to successfully solve a given test problem. The feedback system used in the test presented here is based on the feedback systems employed in the dynamic test discussed by Guthke and Beckmann (2000a), Shabani (2018), Ebadi et al. (2018), Pileh Roud and Hidri (2021) and others. Table 3 provides more detailed information. The feedback system of the DTOLLA – C is divided into five layers (see Table 3). The first layer is accuracy feedback only providing “correct” or “incorrect” in response to a test taker’s answer to an item. The latter is followed by an opportunity to try again. The second, third, and fourth layers of feedback are error-specific (i.e., elaborate feedback). The second layer of feedback takes the form of a question which refers to the common features, asking examinees to think about the similarities and differences of objects and symbols in the examples (thinking prompts). The third layer of feedback is presented in the form of answers to questions posed at the second layer, identifying the common features in the examples. The fourth layer provides the information of relating common features to reasoning rules. The fifth and final layer of feedback shows the correct answer to the test taker together with a short explanation for why this is the correct answer. The feedback is error-specific feedback which means feedback from the same layer (such as “a” and “b”) targets specific information cues and incorrect answer options. For instance, in Item 1 (see Figure 2), if “” or “” was chosen

Table 3: Example of the feedback system.

Item 1 language points		a. 𐀀 on the left represents four legs animals b. 𐀁 at the bottom right represent on the farm
Level of assistance	Type of assistance	Feedback
0	Accuracy feedback	I Correct. (Move to next item) II Incorrect. Please try again.
1	Specific hint 1 (refer to common features)	Incorrect. Please think of the following questions: a. What do the dog, the lion, and the pig have in common, while different from the bird? What do 𐀀 and 𐀁 have in common, while different from 𐀂? b. What do the lion and the bird have in common, while different from the dog and the pig? What do 𐀁 and 𐀃 have in common, while different from 𐀀?
2	Specific hint 2 (identify the common features)	Incorrect. a. The dog, the lion and the pig have four legs, while the bird has two legs. 𐀀 and 𐀁 both have 𐀀 on the left, while 𐀂 has 𐀀 on the left. b. The lion and the bird are in the wild, while the dog and the pig are in the farm. 𐀁 and 𐀃 both have 𐀁 at the bottom, while 𐀀 has 𐀁 at the bottom.
3	Specific hint 3 (relate common features to reasoning rules)	Incorrect. a. The dog (𐀀) and the lion (𐀁) are four legs animals, and both have 𐀀 on their left, so 𐀀 on the left means four feet animals. b. The lion (𐀁) and the bird (𐀃) are in the forest, and both have 𐀁 at the bottom, so 𐀁 means in the forest. In the same way, the dog and the pig are in the farm, and the dog (𐀀) has 𐀁 in the same position as 𐀁, so 𐀁 means in the farm.
4	Stating the correct answer and reason behind	Incorrect. The correct answer is 𐀀. Because the pig is a four legs animal in the farm. The symbol for four legs animal is 𐀀 and for in the farm is 𐀁. Only 𐀀 meets these conditions.

(incorrect answers), subsequent feedback will refer to “a”. If “𐀀” or “𐀁” was chosen, the test taker will get feedback “b”. If “𐀂” is selected, both feedback “a” and “b” will be given to the test taker. This rule applies to layer 2, 3, and 4.

4 Materials and methods

4.1 Participants

In this exploratory proof of concept study, 10 participants were recruited, but 1 dropped out in the middle of the test. This participant was relatively young, at the age of 8. When the test reached to the fourth complexity level, she felt it was too difficult to complete, so she chose to quit. The other 9 participants (6 females) completed the test. They ranged in age from 9 to 13 and were enrolled in grades four through eight (for more details see Table 4). Those participants were recruited through word of mouth via friends and relative's recommendation and advertisements on WeChat.

4.2 Data collection

For the purpose of this proof-of-concept study the test was administered in the form of one-on-one sessions using an electronic tablet with the researcher present. This form of test administration does not represent the ultimately envisioned form (i.e., independent work on a computer, which also makes a group administration of the test possible). This approach was deemed necessary to facilitate information gathering in relation to the aims of the study (i.e., to identify potential procedural issues, including comprehensibility of instruction and feedback; to establish whether the test materials and test procedures are suitable for the targeted age group; to gain insights to inform the computerisation of the test). Response behaviour was recorded on a performance sheet (see Appendix I for an example) and test taker's verbalised thinking processes were audio recorded for later in-depth analyses. The test was presented in Chinese.

Information recorded on the test performance sheet include the number of test items and training items taken, the amount of feedbacks needed, feedback scores, time spent, and miscellaneous notes for test revision and refinements. The test taker's shared thoughts while working on the test items were audio-recorded. We used these thinking aloud protocols (e.g., Gilhooly & Gregory, 1989; Lüer et al., 1990;

Table 4: Participants.

N	Age			Grade			Sex	
	Range	Mean	SD	Range	Mean	SD	Female	Male
9	9–13	11.2	1.2	4–8	6	1.2	6	3

Short et al., 1991; Veenman et al., 1997) to primarily identify any remaining ambiguities in the items and instructions.

4.3 Procedure

Prior to test administration and data recording, informed parental consent was obtained. The test situation started by collecting general demographic information before the purpose of the session was explained and the general instruction about what to do was provided. Then started the test including the audio recording. The test items were presented using a tablet, and participants were able to use a stylus to mark and circle answers on the screen. There were two or three hyperlinks in the lower right corner of the page of each item, to ensure that the order in which the item material and feedback was presented followed the ramification rules described above. While participants were taking the test, the researcher used the test performance sheet to record their performance, provided oral feedback if needed, or helped participants to navigate to the next item.

4.4 Data analysis

The thinking aloud protocols were analysed to identify potential misunderstandings of instructions, ambiguous items, or indications of fatigue during the course of the test taking process. Data from the test performance sheet were also analysed in form of simple descriptive statistics, including time spent on the test and feedback items used in each complexity level.

5 Results and discussion

5.1 Time spent

Insights regarding the overall time needed for the test feed into judging its feasibility in general. Analysing the time spent across each of the four complexity levels also allows to gauge whether the complexity manipulation (i.e., systematic increase in the number of relevant information cues) is appropriately reflected in the levels of difficulty test takers are likely to experience.

Table 5 provides an overview of the average amount of time spent by participants at each complexity level. The time taken by the first three complexity levels almost doubled which increased from 6 min and 31 s on the first complexity level to

Table 5: Descriptive statistics for time spent.

Complexity level (CL)	Time duration		
	Range (in seconds)	Mean (in mins:seconds)	SD (in seconds)
CL1	129 – 1,020	6:31	266.9
CL2	240 – 1,422	12:30	326.3
CL3	617 – 2,824	26:30	798.1
CL4	1,560 – 2,943	33:03	484.2
Total	2,906 – 6,786	1: 28:41	1,363.5

12 min and 30 s on the second complexity level to 26 min and 30 s on the third complexity level, the time taken by the first three complexity levels almost doubled. The fourth complexity level took about 7 min longer than the third complexity level, at 33 min and 3 s. This pattern corroborates the effectiveness of the complexity manipulations across items, which is fundamental for providing learning challenges at different levels of proficiency.

The mean duration in whole test completion was around an hour and a half, which must be considered too long for the majority of potential test takers. It also would render the test's administration as a group assessment within an ordinary classroom schedule unfeasible. This insight will inform our approach to explore (a) the implementation of a termination rule, and (b) reduce the number of items, both test items and training items within each complexity level, all without sacrificing the test's potential to provide high quality diagnostic information about test takers' language learning aptitude.

No particular pattern was detected in terms of age differences in time spent (see Figure 4). The youngest and oldest students spent almost the same amount of time in the whole test, both around 80 mins. One 11 years old and one 12 years old participants took longer time than the other two 9 years old and 10 years old. Although the sample size is far too small to inform any form of generalisation, we tentatively conclude that the test is suitable for 9–13 years olds.

5.2 Feedback used

The amount of feedback used across items in each CL was also analysed to examine the effectiveness of the graduated feedback system implemented in the test and to establish whether the complexity manipulations will be reflected in the levels of difficulty test takers are likely to experience. Figure 5 depicts the frequency of

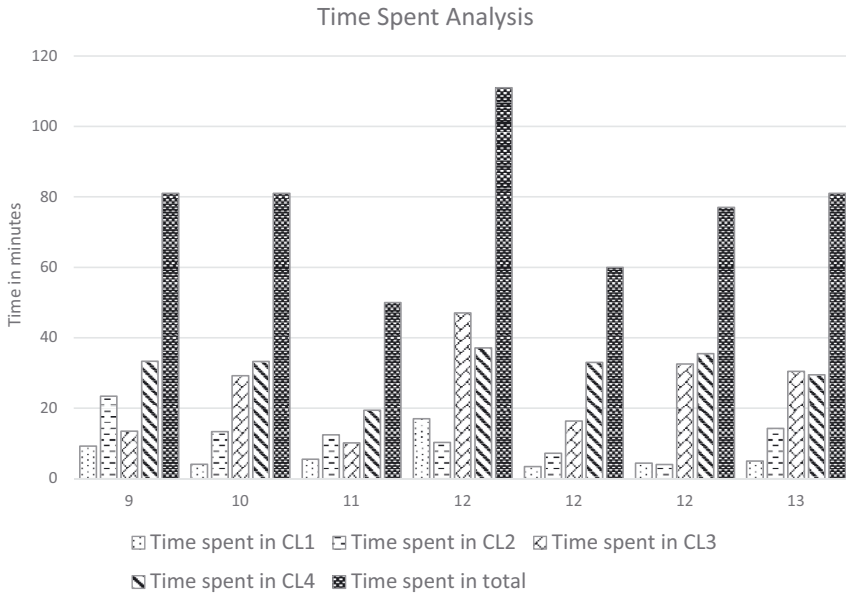


Figure 4: Time spent across age groups.

feedback use at each of the 5 feedback levels across training items. In these diagrams the horizontal axis represents the feedback items in each complexity level. For instance, 4_0 indicates feedback level 0 for training Item 4 (note the system incorporates 5 levels of feedback). Similarly, “1f_1” indicates feedback level 1 for training Item 1f. The vertical axis represents the frequency with which each feedback item was used by the 9 participants.

As an inspection across Panel A to D signifies the frequency of use of feedback increases gradually from CL1 to CL4. In CL1 (see Panel A), feedback items were rarely used, ranging from 0 to 2 times. Feedback level 3 and 4 were not used in CL1. This suggests that the challenges imposed by items in at this complexity levels tends to sit comfortably within the actual level of ability of the test takers. For subsequent complexity levels the frequency of use of feedback items increases, ranging from 0 to 5 times. Feedback level 3 and 4 were used in Item 8f and Item 12. In CL3 (see Panel C), the frequency of feedback use increased further, and did even more so for CL4.

This result is consistent with the previous analysis of the time spent in each complexity level. As expected, the more frequently feedback is used, the more time it takes to answer the task.

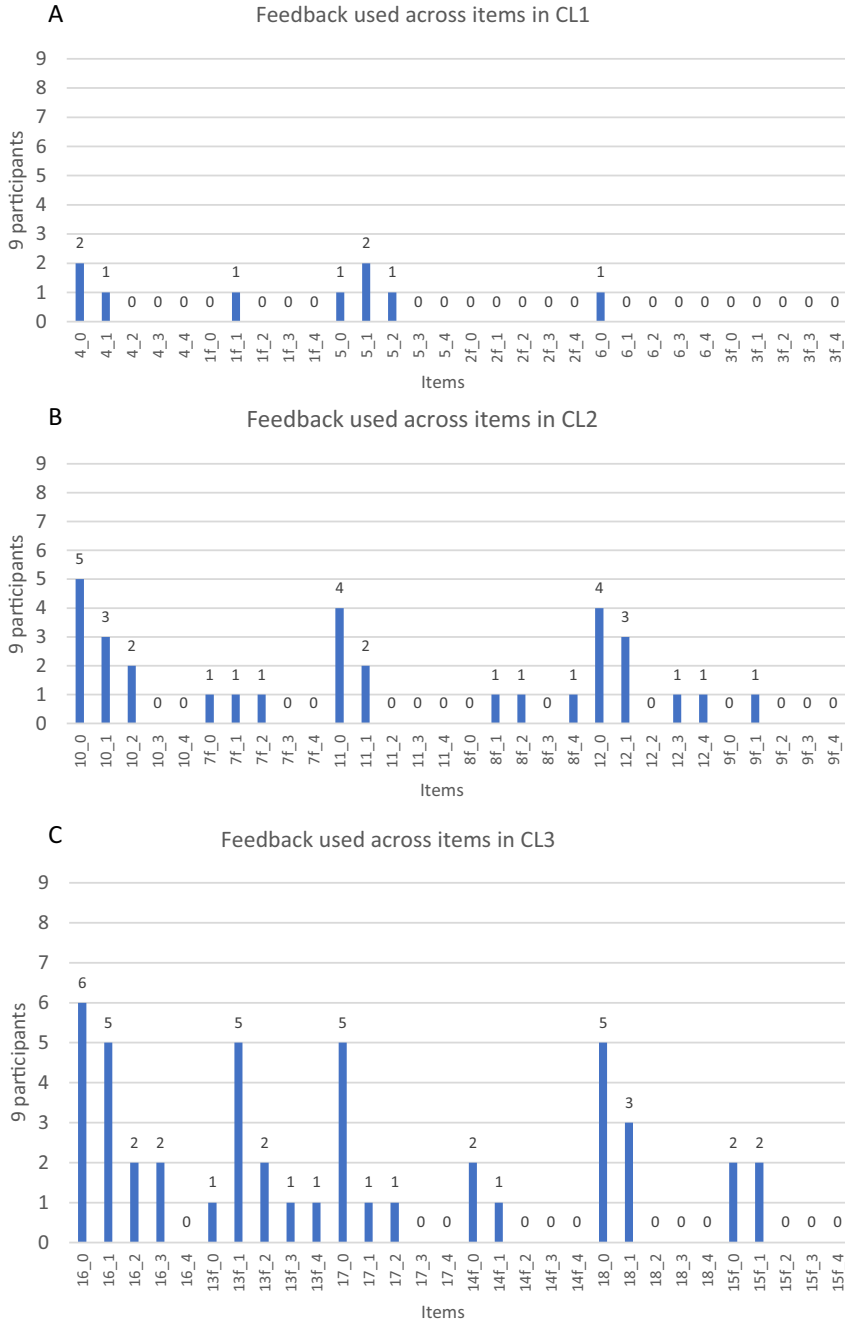


Figure 5: Feedback used across training items in four complexity levels.

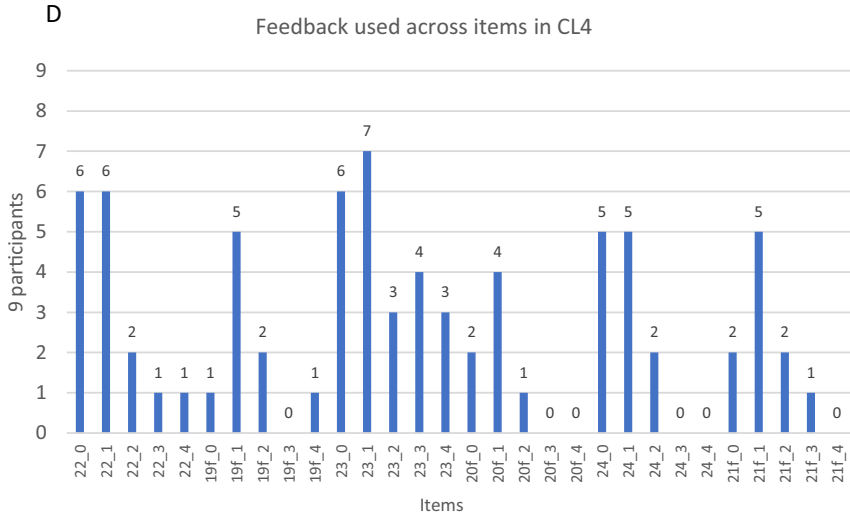


Figure 5: Continued.

This result pattern is also consistent with the intended systematic increase in task complexity. We anticipated an increase in the frequency of feedback needs from complexity level to complexity level.

To study the effectiveness of the graduated feedback system implemented for the training items in the test (see Figure 6). We can see that all “f” training items (such as 1f, 2f, 3f, etc.) required much less feedback than the previous training item. This shows the effectiveness of feedback learning. The amount of feedback at a new complexity level will have a sharp climb to a peak, and then the overall trend of decline. For instance, in CL1, the number of feedback dropped from 3 feedback

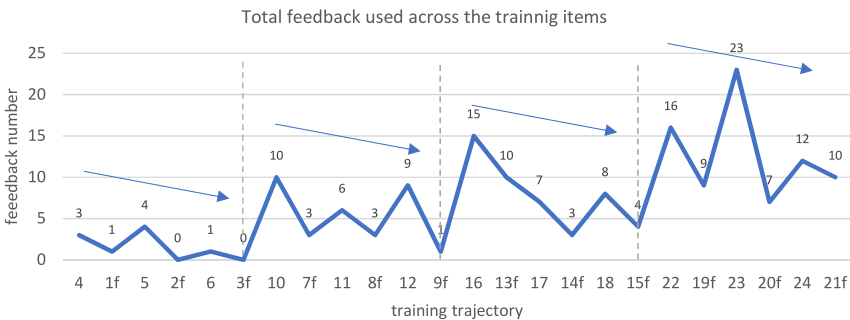


Figure 6: Trends of feedback use across complexity levels.

required for the first training Item 4 to 0 feedback required for the training Item 2f and 3f. Same situation happened in CL2. In training Item 10, the number of feedback rose sharply to 10, and then decreased to 1 for training Item 9f. Also, in CL3, the number of feedback dropped progressively from 15, then rose to another peak of 16 for training Item 22, and then to 9 for Item 19f. Finally, the feedback in CL4 gradually decreased from the peak of 23 to 10 for the last training Item 21f.


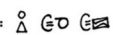


This result is encouraging because it demonstrates both the effectiveness of complexity level design and the effectiveness of feedback learning. We predicted that higher levels of complexity would require more feedback, and that if feedback learning was effective, the feedback required for the same complexity level would gradually decrease. Because if the learning is effective, test takers working on the new item at the same complexity level will benefit from what was learned in the previous item. Therefore, we can infer that the feedback used in this test can provide effective learning for test takers. Although the sample size is small, the result in this study still provides a good indication for the future main experiment.




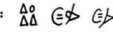
5.3 Thinking aloud protocols


To identify potential sources of ambiguity in instructions or items that could result in confusion for the test taker, the audio recordings of the thinking aloud protocols were analysed. As is to be expected, not every participant was able to engage in thinking aloud. Of the nine participants, only two provided information about their thinking processes during the test. Both participants’ shared reflections reflect very well the intended structure of reasoning necessary to solve the items.

For example, participant “H” shared their thinking process related to Item 15 (see Figure 7 for reference):

Item 15

Example:  :   : 

 :   : 

Task:  : _____

Characters to choose:




Figure 7: Screen shot of Item 15.

H: 三只，二，四，六，这个是五个。这个是第一个。第二个，等于，两条腿。爬着的。不对。两条腿，两条腿，一个飞着，一个这个。跑的，这是在吃东西。第二个是... 蝴蝶有腿吗?好像没有腿。这个两个横代表等于。嗯? Butterflies. 第二个是这个。第三个，诶，信，这是在吃吗?这是吃东西，跑步，飞。诶，不对啊，这不是... 这是第三个。第二个，第二个... 这个。

Researcher: 对了。

(Translations:










H: Three (chicks). Two, four, six. These (butterflies) are five. This (the character for five) is the first one. The second one, this equal (symbol) means two legs. Crawling (referring to the symbol for insects). No (it doesn't mean that). Two legs (referring to the chicks), two legs (referring to the hens), one flying (referring to the birds), one this (referring to the symbol of flying). Running (referring to the hens are running). They are eating (referring to the chicks are eating). The second one is ... Do the butterflies have legs? Seems to have no legs. These two lines mean equal. En? Butterflies (suddenly speaking English). This is the second one. The third one, Ei? Envelop (the symbol for eating looks like an envelope). Are they eating? They are eating. Running. Flying. Oh, no, this is not (the second one), this should be the third one. The second one, the second one ... This one.










Researcher: Correct.


H's reflections are aligned with the stimulus content and the design of the task, including the correct consideration of the number of the objects, the movement ascribed to the objects, and the object's features.

Another example, this time from participant "J" whilst working on item 17 (see Figure 8 for reference) is presented in the following:

Item 17

Example:  :      :   

 :      :   

Task:  : _____

Characters to choose:

𠄕
𠄖
𠄗
𠄘
𠄙
𠄚
𠄛
𠄜
𠄝

Figure 8: Screen shot of Item 17.

J: 这是个啥?两个, 四个吗?这个数字考过了, 数字, 先搞数字吧。一, 二, 三, ... (counting). 十... 这是个严峻的问题。第一个是这个, 第一个是不是这个?尖的, 圆的, 这个是尖的。这个, 这个是 第二个。然后, 这是什么?球, 铁球, 篮球, 铁针, 那应该是跟那个相似的。是不是这个?这个。

Researcher: 不对。再试一次。

J: 是不是这个错了?这个?

Researcher: 对了。

(Translations:

J: What's this (referring to the character of four)? Two (triangles), is it four? This number has been tested. Numbers. I should find out the numbers first. One, two, three, ... (counting the metal nails). Ten, ... This is a serious problem. The first one is this one. Is this the first one? Pointed, rounded. This (referring to the metal nails) is sharp. This one. This (referring to the character for sharp) is the second one. And then, what's this? Balls. Iron balls, basketballs, iron needles. Then it should be similar to that one. Is it this one? This.

Researcher: Incorrect. Please try again.

J: Is this one (referring the image of metal nails) incorrect? Should be this one?

Researcher: Correct.

Participant J correctly encoded the relevant information from the stimulus material (e.g., “iron balls, basketballs, iron needles”). Their approach to solving the item followed the intended logic. This included to focus on finding out the quantity of objects, followed by identifying similarities (and differences) amongst the depicted objects.

The examples presented here are indicative of the overall alignment of the (externalised) thinking processes with the intended test and item structure. This lends support to the impression that there are no major issues with the instruction, the item material, and the item logic.

6 Discussion

This exploratory proof of concept study sought to examine the test's feasibility, to explore whether the complexity manipulations were accordingly reflected in

difficulty or performance scores, whether the graduated feedback system is effective, and to identify potential procedural issues, and to gauge the general time requirement for the test.

Drawing from the data collected in the test performance sheet and the audio recordings of the thinking aloud protocols, the test appears to be feasible in terms of its complexity structure, instruction, items, stimulus material, and feedback system. The increase of time spent at each complexity level indicates that the construction of complexity structures shows the intended effects. The feedback usage analysis corroborates this interpretation. From the audio recordings, no ambiguity regarding test content and item material became apparent. The data also suggest that the test seems generally suitable for 9–13 years old language learners.

The average time taken to complete the test in its entirety was 1 h and 20 min. As could be inferred from the audio recordings too, there was the risk for test takers to experience fatigue and subsequent disengagement in the second half of the test. Therefore, the following adjustments will be implemented in the fully computerised version: First, an extended instruction with example items will be presented prior to starting with the test. Second, observations made during the testing recommend reducing the feedback system to four levels. A too fine-grained feedback system creates the risk of becoming counterproductive in form of overwhelming test takers with well-intended, yet too detailed information. Third, adjust the feedback provision rules in the training form of the test items. This can be realised as follows: If the test taker presented with the training version of test (i.e., 1f, 2f, 3f; 7f, 8f, 9f; ...) the feedback level 0 (i.e., accuracy feedback) can be skipped. This can be justified as the test taker already has received accuracy feedback as a result of the incorrect answer they provided to the item in its test form. In short, being asked to answer the item in its training version represents their second attempt. This change not only reduces the number of possible attempts per item to four, it also will reduce the overall testing time.

For the purpose of pilot testing the test was administered in form of one-on-one sessions. The final version of the test, however, will be fully computerised. The benefit of this approach will not only be in terms of efficiency (enabling group administration), it also will maintain standardisation and objectivity. A computerisation of test administration also enables a precise measurement of item-specific response times, i.e., the time needed to respond and the time taken to process the feedback information. These data have the potential to provide valuable insights into the dynamics of learning *processes* individual test takers are engaged in. For

instance, a reduction of response times across items within a complexity level (i.e., increase in efficiency) could be explored as further indications of a test taker's ability to learn.

7 Conclusions

This paper presents a report of piloting a conceptually informed development of a language learning aptitude test that utilises the principles of Dynamic Testing. These principles include (a) the provision of item-by-item feedback, (b) the integration of learning opportunities in a test situation, and (c) a system of graduated hints and thinking prompts. Test performance is operationalised in terms of the amount of additional training opportunities needed in order to successfully tackle items of increasing complexity.

This exploratory proof of concept study provided first encouraging evidence for the feasibility of the test procedure, the quality of the test stimuli, and the effectiveness of the system of graduated hints and thinking prompts. The insights gained from this study will help informing next steps in the continuing development and refinement of this test and its presentation procedure. These include will efforts an optimisation of the ratio between test and training items, the overall test duration without sacrificing the benefit of an interactive test procedure. However, due to the small number of participants, the current study is, of course, not able to answer all the questions that need to be addressed before the test can be recommended for use in larger scale “real-life” context. The tentatively encouraging outcomes so far will inform the necessary further work.

We have argued that a valid assessment of aptitude requires a dynamic testing approach. The efficient and effective implementation of such test procedure relies on the use of computer technology. In fact, the utilisation of computer technology in this context is instrumental to better align our conceptual understanding of language learning ability (as a dynamic construct) and practically relevant assessment practice.

Ethical approval: The study was conducted in accordance with the Declaration of Helsinki, and approved by the ethics committee of School of Education of Durham University (reference number: *EDU-2022-01-25T14_13_27-nsxh58*).

Appendix I

An Example of Filled Test Result Record

②

Dynamic test result record

Name	Age	Grade	Total test items taken	Total training items taken	Total hints needed	Time started	Time end
Cindy	12	6	11	10	10	10:05	11:57

Item	Taken?	Hints needed					Notes
		0	1	2	3	4	
Complexity level 1							
Test item 1	✓						10:05 0:20 - 1:50
Training item 4	✓						2:04 - 3:34
Training item 1							
Test item 2	✓						3:38 - 4:40
Training item 5	✓	✓	✓	✓			05:34 - 14:05 Do we need to set a time zone for each item?
Training item 2	✓						
Test item 3	✓						15:00 -
Training item 6							
Training item 3							
Complexity level 2							
Test item 7	✓						(should we tell the testee whether it is correct or not items?) - 20:40
Training item 10							
Training item 7							
Test item 8	✓						8
Training item 11	✓						- 24:30
Training item 8							
Test item 9	✓						24:36 - 27:31
Training item 12							
Training item 9							
Complexity level 3							
Test item 13	✓						27:45 - 33:00
Training item 16	✓	✓	✓				33:20 - 38:45 - 42:18 should we tell the testee which character is correct, which is not?
Training item 13	✓	✓	✓				- 39:40
Test item 14	✓						

References

- Beckmann, J. F. (2006). Superiority: Always and everywhere? On some misconceptions in the validation of dynamic testing. *Educational & Child Psychology, 23*(3), 35–49.
- Beckmann, J. F. (2014). The umbrella that is too wide and yet too small: Why dynamic testing has still not delivered on the promise that was never made. *Journal of Cognitive Education and Psychology, 13*(3), 308–323.
- Brown, H. D. (2014). *Principles of language learning and teaching* (6th ed.). Pearson Education.
- Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–115). Guilford Press.
- Carroll, J., & Sapon, S. (1959). *Modern language aptitude test*. Psychological Corporation.
- Chen, Y., Feng, S., & Han, Y. (2019). Research on the education of migrant children in China: A review of the literature. *Frontiers of Economics in China, 14*(2), 168–202.
- De Weerd, E. A. H. (1923). *A study of the improvement of fifth grade children* [Doctoral dissertation]. Yale University, New Haven, CT.
- Ebadi, S., Weisi, H., Monkaresi, H., & Bahramlou, K. (2018). Exploring lexical inferencing as a vocabulary acquisition strategy through computerized dynamic assessment and static assessment. *Computer Assisted Language Learning, 31*(7), 790–817.
- Elliott, J. G. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review, 51*(1), 15–32.
- Elliott, J. G., Resing, W. C. M., & Beckmann, J. F. (2018). Dynamic assessment: A case of unfulfilled potential? *Educational Review, 70*(1), 7–17.
- Gao, Q., Li, H., Zou, H., Cross, W., Bian, R., & Liu, Y. (2015). The mental health of children of migrant workers in Beijing: the protective role of public school attendance. *Scandinavian Journal of Psychology, 56*(4), 384–390.
- Gilhooly, K. J., & Gregory, D. J. (1989). Thinking aloud performance: Individual consistencies over tasks. *Current Psychology: Research & Reviews, 8*(3), 179–187.
- Grigorenko, E. L., & Sternberg, R. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75–111.
- Grigorenko, E. L., Sternberg, R. J., & Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability: The Canal-F theory and test. *The Modern Language Journal, 84*(3), 390–405.
- Guthke, J. (1982). The learning test concept - An alternative to the traditional static intelligence test. *German Journal of Psychology, 6*, 306–324.
- Guthke, J. (1990). Learning tests as an alternative or completion of intelligence tests: A critical review. *European Journal of Psychology of Education, 5*, 117–133.
- Guthke, J. (1992). Learning tests: The concept, main research findings, problems and trends. *Learning & Individual Differences, 4*(2), 137–151.
- Guthke, J., & Beckmann, J. F. (2000a). The learning test concept and its application in practice. In C. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 17–69). Elsevier Science.
- Guthke, J., & Beckmann, J. F. (2000b). Learning test concept and dynamic assessment. In A. Kozulin & B. Y. Rand (Eds.), *Experience of mediated learning: An impact of Feuerstein's theory in education and psychology* (pp. 175–190). Elsevier Science.
- Guthke, J., & Harnisch, A. (1986). Die Entwicklung eines Diagnostischen Programms “Syntaktischer Regel- und Lexikerwerb” - ein Beitrag zur Psychodiagnostik der Fremdsprachenlernfähigkeit. [The development of the diagnostic program “Acquisition of syntactic rules and lexis”: A contribution to

- the psychological assessment of foreign language aptitude]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 7(4), 225–232.
- Guthke, J., & Wiedl, K.-H. (1996). *Dynamisches Testen. Zur Psychodiagnostik der intraindividuellen Variabilität* [Dynamic testing: On the assessment of intra-individual variability]. Hogrefe.
- Guthke, J., & Wingenfeld, S. (1992). The learning test concept: Origins, state of the art, and trends. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 64–94). Springer.
- Guthke, J., Beckmann, J. F., & Dobat, H. (1997). Dynamic testing – problems, uses, trends and evidence of validity. *Educational & Child Psychology*, 14, 17–32.
- Guthke, J., Beckmann, J. F., & Stein, H. (1995). Recent research evidence on the validity of learning tests. In J. S., Carlson (Series Ed.), *Advances in cognition and educational practice. Vol. 3. European contributions to the dynamic assessment* (pp. 117–143). JAI Press.
- Guthke, J., Beckmann, J. F., & Wiedl, K. H. (2003). Dynamik im dynamischen testen [Dynamics in dynamic testing]. *Psychologische Rundschau*, 54, 225–232.
- Guthke, J., Harnisch, A., & Caruso, M. (1986). The diagnostic program of “syntactical rule and vocabulary acquisition” - a contribution to the psychodiagnostic of foreign language learning ability. In F. Klix & H. Hagedorf (Eds.), *Human memory and cognitive capabilities* (pp. 903–911). North Holland.
- Hamers, J., Pennings, A., & Guthke, J. (1994). Training-based assessment of school achievement. *Learning & Instruction*, 4(4), 347–360.
- Haywood, H. C., & Lidz, C. S. (2006). *Dynamic assessment in practice: Clinical and educational applications*. Cambridge University Press.
- Horne, K. M. (1971). Differential prediction of foreign language testing. In *Meeting of the Bureau of International Language Coordination, London*. Bureau of International Language Coordination.
- Jin, X., Liu, H., & Liu, L. (2017). Family education support to rural migrant children in China: Evidence from Shenzhen. *Eurasian Geography & Economics*, 58(2), 169–200.
- Lidz, C. S., & Elliott, J. G. (Eds.). (2000). Introduction to dynamic assessment. In C. Lidz & J. G. Elliott (Eds.), *Dynamic assessment: Prevailing models and applications* (pp. 3–15). Elsevier Science.
- Liang, Z., Yue, Z., Li, Y., Li, Q., & Zhou, A. (2020). Choices or constraints: Education of migrant children in urban China. *Population Research & Policy Review*, 39, 671–690.
- Lüer, G., Ruhlender, P., Klettke, W., & Lass, U. (1990). The construction of procedural knowledge independent of declarative factual knowledge: An experimental study. In R., Groner, G., d'Ydewalle, & R., Parham (Series Eds.), *From eye to mind: Information acquisition in perception, search, and reading. Vol. 1. Studies in visual information processing* (pp. 141–152). North-Holland.
- Navarro, J. J., & Mourgues-Codern, C. V. (2018). Dynamic assessment and computerized adaptive tests in reading processes. *Journal of Cognitive Education & Psychology*, 17(1), 70–96.
- Pimsleur, P. (1966). *The Pimsleur language aptitude battery*. Harcourt, Brace, Jovanovich.
- Perterson, C. R., & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). *Educational & Psychological Measurement*, 6, 369–380.
- Pileh Roud, L. F., & Hidri, S. (2021). Toward a sociocultural approach to computerized dynamic assessment of the TOEFL iBT listening comprehension test. *Education & Information Technologies*, 26(4), 4943–4968.
- Shabani, K. (2018). Group dynamic assessment of L2 learners' writing abilities. *Iranian Journal of Language Teaching Research*, 6(1), 129–149.
- Short, E. J., Evans, S. W., Friebert, S. E., & Schatschneider, C. W. (1991). Thinking aloud during problem solving: Facilitation effects. *Learning & Individual Differences*, 3(2), 109–122.
- Shuang, M., Yiqing, W., Ling, J., Guanzhen, O., Jing, G., Zhiyong, Q., & Xiaohua, W. (2022). Relationship between parent-child attachment and depression among migrant children and left-behind children in China. *Public Health*, 204, 1–8.

- Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist*, 27(1), 5–32.
- Sternberg, R. J., & Gardner, M. K. (1983). Unities in inductive reasoning. *Journal of Experimental Psychology: General*, 112(1), 80–127.
- Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge University Press.
- Utley, C. A., Haywood, H. C., & Masters, J. C. (1992). Policy implications of psychological assessment of minority children. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 445–469). Springer.
- Veenman, M. V. J., Elshout, J. J., & Meijer, J. (1997). The generality vs domain-specificity of metacognitive skills in novice learning across domains. *Learning & Instruction*, 7(2), 187–209.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological progress*. Harvard University Press.
- Wiedl, K. H., Guthke, J., & Wingenfeld, S. (1995). Dynamic assessment in Europe: Historical perspectives. In J. S., Carlson (Series Ed.), *Advances in cognition and educational practice. Vol. 3. European contributions to dynamic assessment* (pp. 33–82). JAI Press Inc.

Bionotes

Fangfang Du

Durham University, Durham, UK

fangfang.du@durham.ac.uk

<https://orcid.org/0000-0002-7711-2428>

Fangfang Du is a PhD student at Durham University School of Education. Her research focuses on how the concept of dynamic testing can be utilised for the assessment of language learning aptitude. She is interested in dynamic, computerised adaptive testing, and language learning and teaching.

Jens F. Beckmann

Durham University, Durham, UK

j.beckmann@durham.ac.uk

<https://orcid.org/0000-0002-4006-9999>

Dr Jens F. Beckmann is Professor of Educational Psychology at Durham University, UK. He conducts research related to learning, problem solving, and cognitive flexibility. This research includes not only conceptual work, but also focuses on methodological approaches to the measurement of cognitive abilities as dynamic processes.