Interaction Design and Architecture(s) Journal - IxD&A, N.23, 2014, pp. 50-67

Designing TEL products for poor comprehenders: evidences from the evaluation of TERENCE

Maria Rosita Cecilia[1], Tania Di Mascio[2], Laura Tarantino[2], Pierpaolo Vittorini[1]
[1] MeSVA, University of L'Aquila, Piazzale S. Salvatore, Ed. Delta 6
67100 L'Aquila, Italy
[2] DISIM University of L'Aquila, Via Vetoio 1
67100 L'Aquila, Italy
mariarosita.cecilia@graduate.univaq.it;{tania.dimascio,laura.tarantino,pierpaolo.vittorini}@univaq.it

Abstract. Developing the capabilities to read and comprehend texts is fundamental for the development of children and for their full participation in society. The FP7 European project TERENCE faced the problem of poor text comprehenders and created the first adaptive learning system for text comprehension for primary school children. The paper, after a brief introduction to the research problem behind TERENCE and an overview of the system, reports on the findings of four round of evaluations aimed at assessing both the usability and the psycho-pedagogical effectiveness of the system, and report them as hints useful for researchers and designers.

Keywords: Technology enhanced learning, psycho-pedagogical effectiveness, usability, poor comprehension

1 Introduction

For all children, developing the capabilities to comprehend written texts is key to their development as young adults. From the age of 7–8 until the age of 11, children develop as independent readers. Nowadays, more and more children in that age range turn out to be poor (text) comprehenders: they demonstrate difficulties in deep text comprehension, such as integrating distant information in texts. The comprehension process may be stimulated by educational intervention carried out by primary school educators; experiments show that inference-making questions centred on a number of identified skills, together with adequate visual aids, are pedagogically effective in fostering deep comprehension of stories. While traditionally the psycho-pedagogical intervention is carried out by primary school educators by means of paper-based learning material, the advent of Learning Management Systems opened new possibility with respect to the support to both teachers and learners.

Generally speaking, a Learning Management System (LMS) is a suite of functionalities designed to deliver, track, report on and manage learning content, learners' progress and learners' interactions, applying to very simple course management systems, or highly complex enterprise-wide, distributed environments.

Anyhow, despite extensive implementation in a number of educational contexts, LMSs continue to be little more than a support tool for education that does not allow for the automation of the educational process. With the aim of solving this shortcoming in current LMS, a great deal of focus has been placed on research related to Adaptive Learning Systems (ALS's), able to tailor their behavior to the individual learner [1]. To achieve this adaptation capability the conceptual model of an ALS [2] is usually made up of:

- a *model component* describing (1) the student relevant information (*Student Model*), (2) the repository of the learning material (*Domain Model*), (3) the description of the user hardware/software skill/capabilities (*Environment Model*), (4) the inferential rules that, given the previous models, provide the actual adaptation (*Adaptation Model),*
- an *engine component* that actually personalises the learning process (*Adaptation Engine*).

As to the former point, different aspects of user modelling have been studied independently from various different viewpoints. Aside from distribution, scalability and performance aspects [3] as well as context information [4], the principal motivations for the development of user models are (i) to characterise an individual user and (ii) to have a generic representation of different types of users and their learning styles. The former approach has received greater attention in research and proof-of-concept implementations. For instance, the KBS-Hyperbook [5] and TRAILS projects [6] base their modelling on (reasoning over) logged user actions. In the AHA! project [7], user actions are not typically logged but are immediately translated into higher-level user model information. There are, however, few ontologies described in literature, the primary ontology being the generic user model GUMO [8] and another more specific in the TERENCE project [9, 10]. There is also a great deal of research on different learning models for students and how these models are closely related to the characteristics of each student and their surroundings and characteristics (e.g., age, country, culture, gender).

With regard to engines, different developmental techniques have been proposed (e.g., intelligent analysis of the learner's solutions, interactive problem-solving support, and example-based problem solving), each of which related to artificial intelligence. The works in [11, 12, 13] are particularly relevant since they are tailored to the specific needs of their users in order to be pedagogically effective.

Nowadays, a few Adaptive Learning Systems (ALSs) promote reading interventions. However, existing ALSs are developed for old children or adults, and not specifically for younger poor comprehenders. Filling this gap was the main intent of the TERENCE project (`www.terenceproject.eu`). TERENCE was an FP7 EU multidisciplinary project that developed the first intelligent ALS for primary-school children in the 7-11 years range and their educators. The system presents to children adequate digital stories, organised into difficulty categories and collected into books, along with instructional smart games for reasoning about stories. The presentation of the learning material is actually organised as a cognitive stimulation designed by the neuro-psychologists involved in the project, also according to organisational constraints set by the schools [14].

In such a context, this paper summarises the main evidences related to the system evaluation, in terms of both usability and psycho-pedagogical effectiveness, achieved during the design process through formative evaluation and at the end of the project through summative evaluation (not reported in any previous publication on TERENCE). The paper hence contributes to the general discourse centred around children-oriented design&evaluation and provides experimental data on children's abilities as addition to interaction design ingredients of Technology-Enhanced Learning products in general and specifically for poor comprehenders.

The remainder of the paper is organised as follows. Section 2 introduces the problem by discussing prevalence along with cognitive and metacognitive difficulties of poor comprehenders. Section 3 briefly summarises the main features of the TERENCE system in terms of architectural model, stimulation plan and learners' interaction environment. Section 4 reports on evidences gained through evaluations of the interaction environment and the stimulation plan, conducted in parallel. Finally, in Section 5, a discussion about the reported evidence is positioned within the general discourse on children-oriented system evaluation and conclusions are drawn.

2 The problem: cognitive difficulties of poor comprehenders

Reading is a complex cognitive activity that transforms print to speech and print to meaning trough a negotiation of meaning between the text and its reader, as an activity of problem solving [15]. According to the "Simple View of Reading", first articulated by Gough and Tunmer, reading is a multidimensional process, including *decoding* and *comprehension* [16]. The two abilities are correlated and take time to develop: children become skilled and independent reader around 7-11 years old [17]. However, while most children learn to read and spell with very little explicit instruction, many learners experience two very different forms of reading problems, namely decoding difficulties and reading comprehension difficulties [18], thus failing to reach functional levels of reading (for example, in the United States, about 2.6 million children aged 6-11 years have a learning disability [19]). Actually, decoding and comprehension skills, although correlated, depend on different cognitive and linguistic skills [20] and thus researchers classify poor readers in *poor decoders* and *poor comprehenders,* who show distinct cognitive and linguistic profiles [21]. Poor decoders, often defined as dyslexics, have difficulties with learning to read fluently, yet manage to comprehend what read reasonably well [22]. Poor comprehenders read words and sentences accurately, fluently and at age-appropriate levels, but have serious difficulty understanding what they have read [23].

There are many different experimental assessments that purport to measure reading comprehension in children. However, in order to ensure that a child has a specific deficit in reading comprehension, it is important to obtain, as far as is possible, independent assessments of the decoding and the comprehension. For instance, the Neale Analysis of Reading Ability - NARA [24] has been used widely in the UK in studies where measures of both word decoding accuracy and reading comprehension are required [25]. In Italy, decoding (the correctness and the speed) and reading comprehension were independently assessed with the MT standardized tests [26], the

most commonly used psychometric Italian instrument to measure these factors [27]. In detail, for measuring comprehension, the child is asked to read a story and answer a set of questions. The comprehension variable reflects the number of correct answers readers select from a list of alternate choices: the higher the score, the better the comprehension. The child performances are compared with the score reported in the conversion tables for the Italian population and can be classified as normal or as risk condition in the comprehension. More precisely, depending on the class and on the score, a child can be assigned to one of the following clusters: "Need for immediate intervention" (NI), "Attention is needed" (AN), "Sufficient performance" (SP), "Complete performance" (CP). If a child belongs to one of the first two clusters, he/she can be considered as a poor comprehender. According to such assessments, more and more children of 7-11 years old turn out to be poor comprehenders [28].

According to studies, an aspect that may affect text comprehensions is the ability to hear. For example, poor comprehenders without hearing impairments comprise up to 10% of 7-11 years-old in UK schools [29]. The estimate dramatically increases when the whole population of young deaf people is considered. According to Wauters, Van Bon, and Tellings, only 19% out of 504 hearing impaired 7-20 olds showed reading comprehension scores above the third grade level [30]. The children's comprehension of spoken texts is poor [31] and their ability to produce coherent narratives is impaired [32]. In general, hearing and deaf poor comprehenders show relatively age-appropriate word recognition skills [30], but text comprehension difficulties become apparent when children need to answer questions that require more than recall of simple facts of the text [33]. They also have difficulties in using cohesive markers that signal relations in text. They are poor at making inferences when reading and listening to language and coherently integrating information from different parts of a text [34]. They also have great difficulty in identifying errors and inconsistencies in texts, which are taken as an indication of difficulties in active metacognitive processes, such as comprehension monitoring [35]. Further studies have shown that these differences in reasoning skills cannot be attributed to differences in general knowledge [36], but as for deaf children, a lack of the vocabulary and grammar knowledge used in texts may also affect text comprehension [37]. Deaf students spend more time on reading than their hearing peers, however taking more time to read, not reading more materials [38]. Indeed, it is well documented that deaf children generally achieve lower levels of reading attainment in comparison to their hearing peers [39]. In particular, Wauters et al., showed that reading comprehension scores of deaf children were far below the scores of hearing children [30].

The inability to understand what they have read is a major obstacle for student to learning. This has huge costs for the people affected and also for wider society: reading difficulties may have long-term educational, social and economic consequences and increase the risk of developing psychological and emotional problems [40] and can persist for a lifetime [41]. Early recognition, evidence based on evaluation, and treatments are necessary to achieve the best possible outcome.

3 An overview on the TERENCE system

To guide the design and development of the TERENCE system, we followed an iterative user-centred design approach [42] mixed with evidence-based design [43] which stresses the role of empirical evidence gathered from experts in order to attain pedagogical effectiveness. The main idea behind TERENCE is that the stimulation by the system integrates with the traditional stimulation by teachers, while taking into account the individual learner's skills, styles and profiles.

The TERENCE system is developed as an ALS (the latest release is available at http://hixwg.univaq.it/terence/3rd-release/). The *model component* is structured as follows:

(1) the *student model* contains all demographic information on learners along with logs about reading and playing activities (e.g., stories chosen while interacting with the ALS and performances related to the smart games, such as correct answers and mistakes);

(2) the *domain model* (i.e., the repository of learning material) includes stories and associated smart games, and accessory material: *stories*, organized in books, are ordered and actually written into four different versions with increased cognitive difficulty [44]; instructional *smart games* are of three types: factual (e.g., "guess who did something", temporal (e.g., "what happened before/after this event?") and causal (e.g., ",what caused this?", "which is the effect of this?"). Figure 2 illustrates two cases of factual and temporal games (for more information about the instructional smart games used in TERENCE we refer to [45]); the *accessory material* includes elements designed in order to make the learning experience appealing, such as avatars available for the children, cards illustrating the characters of the books, relaxing games that can be played by children after the stimulation for entertainment and relaxing purposes;

(3) the *environment model* contains information about learners' technological skills (though traditionally distinct, in TERENCE the environment model and the student model are actually integrated from an implementation point of view);

(4) the *adaptation model* includes rules that associate learners with the correct versions of stories and smart games, according to their comprehension level.

The *engine component* is implemented as a rule-based expert system that, according to the rules of the adaptation model and the information in the other models [46], provides a learning experience adaptive in terms of lists of the available *avatars*, ordered according to the child gender, *books*, depending on the child age, proper *story versions* and *smart games*, challenging for the child, but not too difficult [10].

As to the psycho-pedagogical stimulation plan, according to the advice of the experts of the project, in TERENCE we adopted the constructivist pedagogical approach arguing that "learning takes place in contexts" [47] and suggesting the ideas of "training via iterations" and of "rewarding structures" [48]. Accordingly, the system is conceived so to integrate in regular school activities, and its stimulation plan is inspired by a traditional teaching strategy including reading the story and analyzing the text via inference-making question answering. In TERENCE the

stimulation plan is implemented via learning sessions (two or three per week) mirroring a customary warm-up, peak, and relaxing phases structure, composed of reading and playing activities; specifically: (1) reading a story, silently – warm-up, (2) resolving related smart games for analyzing the story – peak, and, finally, (3) playing with other games able to relax the learners according to a their performances in the previous step – relaxing. Figures 1 and 2 provide a flavor of the learner's experience.

(a) (b)

Fig. 1. Sample screenshots from the the TERENCE system related to reading activities: (a) browsing the characters' list; (b) reading a story episode. All displays are based on a simple common template including system communication on the left (carried on by means of an avatar) and interaction with content in the main area. Stories are structured as sequence of illustrated episodes presented according to a focus+context carousel pattern that allows children to focus on single episodes while maintaining a global vision on the whole story and on the order of episodes [14].

(a) (b)

Fig. 2. Sample screenshots from the the TERENCE system related to playing activities: (a) a factual game – who game; (b) a temporal game – befor-after game. The content area of all games is divided into three portions: a lower bar displaying three cards corresponding to the three possible choices, a middle area displaying the question to be answered, and an upper part depending on the specific games. In all cases the interaction is based on "drag&drop": the child has to select the card corresponding to the correct answer and drag it into specific elements of the upper part.

4 Evidences from the evaluation of the TERENCE system

As pointed out in [49], educational systems have to provide *curricular materials* (e.g., books in TERENCE) and *assessment strategies*, not to be confused with "evaluation": while evaluation refers to the system, assessment refers to learners, judging their performances in terms of the psycho-pedagogical outcomes. In an adaptive learning system, like TERENCE, assessment is a core concept of the system itself, while what is to be evaluated, then, is the psycho-pedagogical effectiveness of the system strategy, i.e., the educational value of the system. This has to be done within the context of an iterative design, thus including formative evaluation throughput the entire project and summative evaluation to be done when the system 'goes live' [50]. The design is in fact conducted as a "test and make changes" process, according to *formative evaluation*, which is "user testing with the goal of learning about the design to improve its next iteration" [51]. After formative evaluation and iterative design are complete, a final *summative evaluation* serves to document the effectiveness of the design and justify its use by learners and teachers [49].

In both cases, evaluation must pay attention first to *usability*, and second to *learning outcomes*: if students cannot use the system, they certainly will not learn through its use [49]. Usability is evaluated according to customary methods [52], while for evaluation of learning outcomes a variety of quantitative and qualitative techniques are commonly used [53].

Table 1. The TERENCE project evaluations.

Evaluation Characteristics	Issues	Release (Month)	Involved Users
1st expert-based Formative Qualitative April-May 2012	Usability Curricular material (stories, books and illustrations).	Prototypes (March 2012)	about 10 domain experts of text comprehension and interaction design
1st user-based Formative Qualitative June-Sept 2012	Usability Learning outcomes	1st release (June 2012)	about 170 learners deaf and hearing
2nd expert-based Formative Qualitative Nov 2011-Jan 2012	Smart Games revision and production	2nd release (September 2012)	about10 domain expert of pedagogy
2nd user-based Summative Qualitative&Quantitative March-June 2013	Usability Learning outcomes	3rd release (March 2013)	About 830 learners deaf and hearing

Thus, according to the state-of-the-art, evaluating TERENCE required assessing both its usability and the psycho-pedagogical effectiveness (Sections 4.1 and 4.2 highlight evidences coming from the two evaluations, while Section 4.3 highlights the relation between them). More specifically, the evaluation articulated in two expert-based evaluations and two user-based evaluations: the expert-based evaluations involved domain experts of (poor) text comprehension and human computer interaction; the two user-based evaluations involved about 170 and 830 users respectively. Table 1 offers a synoptic view of the entire TERENCE project evaluation: the first column specifies the characteristics of the evaluation (expert-based/user-based, formative/summative, qualitative/quantitative), the second and third column specify the issues anf the release under evaluation, respectively, and finally the last column indicates the number of users involved (we notice that figures related to formative user-based evaluation fulfill the prescription from Bailey, who calculated that to be 90% confident of finding usability problems that will affect 99% of users requires about 112 representative test participants [54]).

4.1 Evidences coming from the usability evaluations

As described in Table 1, the usability evaluations were performed via two expert-based evaluations and two user-based evaluations. The main goals of these four usability evaluations have been:

- Examining if the issues raised during evaluation at stage *I* were satisfactorily fixed in the TERENCE release at evaluation at stage *i+1*;
- Examining the quality of the interaction;
- Assessing the satisfaction of the learners in using the system and the interest in using it.

In particular, during the expert-based evaluation, experts evaluated if the interface (i) followed the general visual design guidelines, (ii) well supported users, and (iii) provided appropriate feedback. During the user-based evaluation, investigators examined users performing the main tasks of the TERENCE systems, reported in Table 2.

Table 2. The tasks evaluated during the TERENCE project evaluations.

Task order	Task Description
1	Accessing the system via the login page
2	Choosing an avatar
3	Choosing a book
4	Choosing a story in the spatial map of the book
5	Browsing and reading the cards of characters
6	Browsing and reading a story
7	Browsing and playing with smart games
8	Browsing and playing with relaxing games

The people involved in the user-based evaluations were real users; in particular, as shown in Table 1, during the 1st user-based evaluation people involved were about 170 and during the 2nd user-based evaluation were involved about 830 users. Different methods were used: heuristics evaluation and cognitive walkthrough for the expert-based evaluation and direct observation and controlled experiment for user-based evaluations [52]. During the user-based evaluations qualitative data were thus gathered as follows:

- Via direct observations, e.g., of facial expressions, and by tracking comments per post (that means per user) using a structured schema. At the end of the week, the supervisor tutors made a summary of the overall gathered schema.
- Via inquiry (indirect questions to children) at critical points, e.g., if the child explicitly asked for help, if the child seemed lost, if the child seemed upset.

Quantitative data were gathered during the 2nd user-based evaluation, when the last prototype went live. The quantitative data gathered through log files were:

- The story titles chosen by learners;
- For the reading task, the start and end time for reading the selected book;
- For each game instance, the time for its resolution before the game was over.

Results and evidences are reported considering the goals defined so far. First of all, we report the evidences come up during the expert-based evaluation and the evidences per each task performed by learners (see Table 3 and 4) during the two user-based evaluations analysing both qualitative and quantitative data (in-depth detailed discussions on evaluations and their findings can be found in the project deliverables [55, 56, 57, 58]).

Table 3. The tasks evaluated during the TERENCE project evaluations.

Task Description		Evidences Gathered during the evaluation
Accessing the system via log-in page	a)	Children loved human presence and animation of the visual page; in general, the login page was appreciated from the graphical viewpoint.
	b)	In general, 7-year olds children needed assistance in inserting login data.
Choosing avatars	a)	In general, male learners chose the male avatars; female learners chose the female avatars.
	b)	In general, no children had problem in browsing avatars using both carousel effect and/or arrows.
	c)	Children were in general interested in avatars, but for none in particular—learners changed avatar choices across sessions.
	d)	Children appreciated the role of the avatar—discovering the chosen avatar in all the pages of the TERENCE was very exciting for children, especially when the avatar was happy after a successful game play.
	e)	The majority of children considered very nice to browse through avatars using the carousel effect.
	f)	During the small-scale evaluation, the relation between avatar and points resulted not clear.
Choosing Books	a)	Children appreciated the layout and the carousel effect to browse books.
	b)	Children reported that in the first prototype of TERENCE, the font used for the book title (below 10 points) was too small

Choosing stories	a)	Children appreciated the use of a spatial map for choosing stories of a book.
	b)	In general children reported that in the first prototype of TERENCE, the font used for the story titles was too small(below 10 points).
Browsing and reading the cards of characters	a)	Children appreciated the layout and the carousel effect to browse cards.
	b)	Older children seemed more interested in reading cards than younger one.
	c)	During the small-scale evaluation, the information about the characters was not easily readable due to the fact that the used font was too small (below 10 points).
	d)	During the small-scale evaluation, cards resulted too numerous (about 20), for these reason in the last prototype of TERENCE cards were associated to the chosen story and not the chosen book (the number of cards reduced to about 10).
	e)	During the large-scale evaluation, some children said that it is boring to always read the same cards—due to system constraints, cards are the same for all the stories of a book.
	f)	During the large-scale evaluation the cards were read only during the first 2 weeks, afterwards they were completely skipped.
Browsing and reading stories	a)	Deaf children first read then looked at images for fixing in mind what they had read.
	b)	Story plots were generally judged funny and creative, instructive and with a deep meaning.
	c)	In general children liked the illustration style.
	d)	The majority of younger children complained about the font size or the type of font (below 10 points); the main difficulties emerged with 7 year olds—in those cases, children often used the finger as pointer to hold the sign while reading and they read aloud, as typical of their age.
	e)	All children watched closely illustrations, and complained when they noticed any perceived incoherence between the story text and its illustration.
Understanding story illustrations	a)	Some children complained about incoherencies between story texts and illustrations or badly resized images.
	b)	Some older children judged complaints of too small fonts (below 10 points) or not nice font type illustrations good for younger children.
	c)	Many deaf children complained about lack of vivid colours, and the characters being always the same or the illustrations not being realistic.
	d)	Many deaf children complained about lack of sufficiently visible page number.
Navigating the system	a)	In general, children appreciated the layout and the carousel effect to browse objects.
Browsing and playing with smart games		See Table 4.
Browsing and playing with relaxing games	a)	Learners complained about the fact that the games were not contextualised with the latest read story.
	b)	Learners asked to add more relaxing games.
	c)	Due to few bugs, learners were sometimes unable to spend all the points they gained to play with relaxing games.
	d)	Almost all types of relaxing games were appreciated; a qualitative ranking for relaxing game placed monkey at the first position, then slice the fruit, diamonds, find the differences and, finally, find the way.

Table 4. The smart game playing task.

Playing smart Games subcategories		Evidences Gathered during the evaluation
Choices	a)	Children easily used captions, they resulted usable, no problems emerged
	b)	Children had no problems in managing time and causality game choices
	c)	Children had no problems in understanding if the choices were available to move or not
	d)	Children appreciated the grey-out effect for highliting the unavailable choices
Interaction modalities	a)	Children had no problems in using the drag and drop
	b)	In general, children appreciated the easily affordance of all games
Feedback	a)	Many deaf children complained about the time for responding to smart games; they found it too fast;
	b)	When the solution feedback appears, children noted it, and were interested in understanding what was the correct solution (e.g., deaf children complained when it disappeared without allowing them to read it);
	c)	In general childred appreciated the consistency and explanatory feedback that resulted in general clear.
Points and instructions	a)	Children in general paid attention to points, because they learned that points give them extra coins and time for playing with relaxing games;
	b)	In general, male learners were interested in their score, they told everybody about the points they gained upon returning in their classroom;
	c)	In general, both male and female children remember their total scores and, in the very few cases in which the system lost score information, children were very sad;
	d)	During the small-scale evaluation instructions in general, and game instruction in particular, were not read or not sufficiently clear;
	e)	During the small-scale evaluation points were not noticed or not sufficiently clear.
User satisfaction	g)	During the large-scale evaluation instruction resulted in general clear, thanks to the system tutorial introduced in the last release;
	a)	During the large-scale evaluation children loved the visual metaphors of the feedback;
	b)	In general, children asked for more relaxing games;
	c)	At the end of the project, during the large-scale evaluation, children asked to play TERENCE at home, to conclude reading “their” TERENCE book;
Times performance	a)	Times for reading and playing decrease while using the system;
	b)	The learner precision in resolving games increases in time

4.2 Evidences coming from the psycho-pedagogical effectiveness evaluations

This section describes the psycho-pedagogical data collected during the large-scale evaluation among hearing students in Italy and the related research findings. The study was carried out in 3 months (March/June 2013). The sample of participants, summarised in Table 5, was made up of three schools, the Comprehensive Institute “Mazzini-Fermi” of Avezzano, the Comprehensive Institute “Fontamara” of Pescina,

made up of three complexes, named “Pescina Centro”, “Pescina Oriente” and “Cerchio”, and the Comprehensive Institute “Roccasinibalda” of Rieti. The experimental group was made up of the Comprehensive Institute “Mazzini-Fermi” of Avezzano and the complex of “Pescina Centro”. The other complexes of “Pescina Oriente” and “Cerchio” did not participate in the TERENCE stimulation plan because of Internet connectivity issues, as well as limits of time and resources. They played with other challenging activities. However, during the pre-evaluation all children were tested. The control group was made up of the Comprehensive Institute “Roccasinibalda” of Rieti. In the experiment design, the experimental group used TERENCE together with standard school activities, while the control group only performed standard school activities. Inclusion criteria was all students 7-11 aged, while exclusion criteria were inadequate knowledge of Italian language, lack of informed consent, or diseases/health conditions that did not allow the assessment of reading performances. Table 5 summarizes the number of the Italian hearing participants during the large scale evaluation and the available data for each group.

Table 5. Number of the Italian hearing participants during the large scale evaluation and the available data.

School/complex	Participants	Available data
C.I. “Mazzini-Fermi” of Avezzano	254	Pre/post
C.I. “Fontamara” of Pescina	186	
· “Pescina Centro”	68	Pre/post
· “Oriente”	61	Pre
· “Cerchio”	57	Pre
C.I. “Roccasinibalda” of Rieti	183	Pre/post

The aims of this evaluation were to investigate (i) a pre/post difference in the experimental group and in the single schools that made up of the experimental group, (ii) a pre/post difference of the experimental group and the single schools with respect to a control group, (iii) whether a different effect can be identified in poor comprehenders that in good comprehenders, and (iv) the prevalence of poor comprehenders among all Italian sample and their socio-demographic characteristics. In the following, we report the results of the aforementioned research aims wrt (i), (ii) and (iii), while for issue (iv) we refer to [59, 60]. We notice that, as to psycho-pedagogical effectiveness, evidences coming from the studies of deaf learners are not reported here, since the findings are still under evaluation (the data and a preliminary analisis are delivered in [57]).

(i) Investigate a pre/post difference in the experimental group and in the single schools that made up of the experimental group. As for the first objective we considered the students of Avezzano and “Pescina Centro”. A Wilcoxon signed-rank test [61] was used to investigate the pre/post difference. The complex “Pescina Centro” of Comprehensive Institute “Fontamara” had 14 students that resulted poor comprehenders at the pre evaluation and only 6 (8.82%) at the post evaluation. The difference is statistically significant ($p<0.0001$). The school “Mazzini-Fermi” of Avezzano had 15 students (5.91%) that resulted poor comprehenders at the beginning of the intervention, and only 2 (0.79%) at the end of the intervention. The difference

was statistically significant (p=0.0234). For the whole experimental group, 29 students (9.01%) were poor-comprehenders at the test and only 8 (2.50%) at the re-test. The difference was statistically significant (p<0.0001). So, as for the second objective, the analysis showed that TERENCE stimulation plan significantly improved comprehension in the experimental group and in the single schools that made up of the experimental group.

(ii) Investigate a pre/post difference of the experimental group and the single schools with respect to a control group. As for the second objective we included the control group in the analysis and we used an analysis of variance for repeated measures test [61]. In detail, we used the group as the between factor (experimental vs control group) and time (pre vs post) as the within factor. With respect to comprehension score, the control group was more homogeneous to "Pescina Centro" than to Avezzano, since the average of the comprehension variable, at the pre-test (i.e., 7.67), is closer to "Pescina Centro" (7.61) than to "Avezzano" (8.56). A one-way ANOVA confirmed that only "Pescina Centro" was not different to "Roccasinibalda". Consequently, we focus on the comparison only of "Pescina Centro" with the control group. The analysis showed that the improvement in reading comprehension in "Pescina Centro" vs "Roccasinibalda" is statistically significant (p<0.0001). So, as for the second objective, TERENCE improves reading comprehension also in comparison with a control group.

(iii) Investigate whether a different effect can be identified in poor-comprehenders that in good-comprehenders. As for the third objective, we considered the average values of the comprehension variable for poor comprehenders and for good comprehenders, as in the pre- and the post-tests in the total experimental group. The analysis showed that, despite the larger increase in the poor comprehension group than in the good comprehension group (0.56 vs 0.06), such a difference was not statistically significant (p=0.3665). So, in summary, TERENCE can be used by both poor and good comprehenders.

4.3 The relationship between psycho-pedagogical effectiveness and usability evaluations of the TERENCE system

This section describes the results that came out from the merge of usability and effectiveness data during the large-scale evaluation in the complex of "Pescina Centro" (this school was the only one, among all recruited in Italy, who gave the authorisation to combine the usability data with the psycho-pedagogical data). As for usability, we investigated the following variables: (i) number of stories read, (ii) number of episodes read, (iii) average reading time, i.e., the total amount of reading time (in seconds) divided by the number of the episodes read, (iv) number of smart games played, (v) average playing time, i.e., the total amount of time (in seconds) spent in smart games divided by the total number of the smart games played by a learner, (vi) difficulty level (1, 2 3, or 4) of the story assigned to each child by the system according to his/her level of comprehension at the end of intervention, (vii) precision in the smart games, which indicates the number of the smart games correctly resolved divided by the number of the smart games played by a learner. As

for psycho-pedagogical effectiveness, the average comprehension score at the re-test (CE), after TERENCE stimulation plan, was analysed. To examine whether usability data may predict psycho-pedagogical results, a multivariate linear regression was used [61]. The analysis indicated an association between the average comprehension score at the re-test and the measure of the precision in the smart games ($p=0.015$). So, precision in smart games is linearly related to the average comprehension score at the re-test. Even though, we cannot state that the prediction is reliable, since the R^2 value is quite low ($R^2=0.1773$).

The results coming out from the merge of usability and effectiveness data showed that precision in smart games might predict comprehension at the end of the intervention, but that this prediction is not reliable. This result may be explained by the fact that the sample was not large enough. Another possible explanation is that – as qualitative data analysis for usability testing showed – some children with comprehension problems read more than once the same story, learned the solutions of the smart games, replied them in the successive interactions with TERENCE, without actually having comprehended the story, only to quickly move to the relaxing games. For them, therefore, we may have experienced high precision values, without an actual improvement in comprehension. It is also possible that – simply – the precision is not an indicator enough sensible and/or specific to predict comprehension problems. Further studies are needed regarding the combination of usability and psycho-pedagogical data in a larger sample, so to verify whether TERENCE can be used also as a system for detecting reading comprehension problems.

5 Discussion and conclusions

In this paper we presented some results of TERENCE, a multidisciplinary project aimed at designing the first ALS specifically conceived to support 7-11 poor text comprehenders. The TERENCE solution, shaped around the concept of repeated interaction experience and consistent with consolidated pedagogical approaches built on question-based games, is based on a visual interaction environment where children read stories and play smart and relaxing games. In particular, differently from previous publications discussing aspects related to other aspects of the system (its architecture, the psycho-pedagogical issues and the interaction enviroment), this paper presented the findings of four rounds of evaluations aimed at assessing the interaction environment and the psycho-pedagogical effectiveness. Such findings were reported as evidences come out during evalutations and are furher discussed in the following as a contribute to the discourse going on among researchers about design and evaluation of children-oriented applications.

Ideally, one should be able to turn evaluation evidences into specific guidelines for children-oriented guidelines, also according to the design science approach [62] that underlines the dual role of theories in the design: not only do they constitute the ground of an artifact construction, but they should also be the outcome of the design process. Knowledge and understanding of a problem domain and its solution should hence be achieved by the building and the application of artifacts. However it has to be said that, as discussed also in [63, pp. 361-362], when talking about children the

translation from experimental data to guidelines raises difficulties that designers do not experience when designing for adults, for a number of reasons: first of all, since children are a moving target, rapidly learning and rapidly changing their cognitive, sensory and motor skills, with greater variability for younger children, longitudinal studies should be necessary to understand how children change in their interactions with technology as they get older; furthermore, guidelines may shortly become obsolete since children in one decade tend to have more experiences with ICT devices than children from the previous decade. Anyhow, as discussed also by Hourcade [63], it is crucial that designers report on their findings even when they cannot provide immediate guidelines or recommendations for interaction design, to contribute to the maturation of the field, to open discussions and to avoid possible inconsistencies in guidelines and design principles.

One such case is the comparison between click-move-click and drag&drop techniques, with contrasting results and suggestions in the literature about their use, with reference to both the children age and the distance involved by the task (see, e.g., [64, 65] and [63, pp. 323-325]. In our case, as reported in the findings of Tables 3 and 4, in all user-based evaluations and for all tasks not only we did not detect any problem in the use of drag&drop techniques, but we also realized that the attention required by the use of the technique had the positive effect of making children maintain a correct level of concentration while using the system, with consequent better results on the learning side.

Another interesting observation from our findings is related to the relationship that children engage with avatars: differently from adult-oriented games where avatars are considered mostly as a personification of the player, in the children case the avatar was considered as a person distinct from the child, a kind of "helping character" acting as a surrogate of the teacher (usability evaluation, in fact, showed a high level of stress when children felt to be left alone with the system). For this reason we assigned to the avatar a conversational behaviour, providing instructions, suggestions and rewards in case of correct answers.

A third detected evidence deserving additional comment is the fact the children regularly complained about even minimal incoherencies between the text and the associated illustrations. Considered that according to the dual-code theory [66] both verbal association and visual imagery – processed differently and separately along distinct channels in the human mind – can be used to represent information and to master learned material, a consequent observation is that particular attention has to be paid on the pairing among verbal and pictorial information so that they can re-inforce each other in the learning process.

As a final remark, it has to be underlined that, as many other contemporary applications, learning environments belong to the so-called 3rd paradigm of the HCI, which, differently from 2nd paradigm systems based on the metaphor of "interaction as information communication", views the interaction as a form of "meaning making" [67]. Measures of success cannot be then based (only) on effectiveness and efficiency of information transfer or on the fit between the user and the system; researchers have to investigate what are the politics and the values of the system and how are supported in the design. In our case we then assessed both the usability and the educational value of the system, treating these two aspects – and their evaluation – as tightly integrated. In TERENCE, as to the educational value, the analysis showed that the

stimulation plan significantly improved comprehension in the experimental group and in the single schools that made up of the experimental group, that TERENCE improved reading comprehension also in comparison with a control group, and that improved comprehension both in poor and good comprehenders.

References

1. Brusilovsky P., Millán E. User Models for Adaptive Hypermedia and Adaptive Educational Systems. In: P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.), The Adaptive Web. Methods and Strategies of Web Personalization, pp. 3--53. Springer LNCS 4321 (2007)
2. Santos J., Anido L., Llamas M., Álvarez L., Mikic F. Applying Computational Science Techniques to Support Adaptive Learning. Computational Science. LNCS 2658, pp.1079--1087 (2003)
3. Kobsa A., Fink J. An LDAP-based user modeling server and its evaluation. User Modeling and User-Adapted Interaction, 16(2), pp.129--169, Springer (2006)
4. Jameson A. Modeling both the Context and the User. Personal and Ubiquitous Computing, 5, pp. 29--33 (2001)
5. Henze N., Nejdl W. Adaptivity in the KBS hyperbook system. In 2nd Workshop on Adaptive Systems and User Modeling on the WWW. Workshop held in conjunction with the World Wide Web Conference (WWW8) and the International Conference on User Modeling (1999)
6. Heller J., Levene M., Keenoy K., Albert D., Hockemeyer A. Cognitive aspects of trails: A stochastic model linking navigation behaviour to the learner's cognitive state. In: J. Schoonenboom, M. Levene, J. Heller, K. Keenoy, and M.Turcsányi-Szabó (Eds.), Trails in Education. Technologies that Support Navigational Learning. Rotterdam. Sense Publishers (2007)
7. De Bra P., Smits D., Stash N. The Design of AHA!. Proceedings of the ACM Conference on Hypertext and Hypermedia, 133, Odense, Denmark (2006)
8. Heckmann D., Schwartz T., Brandherm B., Schmitz M., Wilamowitz-Moellendorff M. GUMO - the General User Model Ontology. Proceedings of the 10th International Conference on User Modeling, LNAI 3538, pp. 428--432, Edinburgh, Springer Verlag (2005)
9. Alrifai M., Gennari R., Tifrea O., Vittorini P. The Domain and User Models of the TERENCE Adaptive Learning System. In Proceedings of eb-TEL 2012, LNCS, Springer, (2012)
10. Alrifai M., Gennari R., Vittorini P. Adapting with Evidence: the Adaptive Model and the Stimulation Plan of TERENCE. In Proceedings of eb-TEL 2012, LNCS, Springer (2012)
11. Brusilovsky P. Adaptive Hypermedia: From Intelligent Tutoring Systems to Web-Based Education. In Intelligent Tutoring Systems, pp. 1--7. Springer Berlin Heidelberg (2000)
12. Paramythis A., Loidl-Reisinger S.Adaptive learning environments and e-learning standards. In Second European Conference on e-Learning, pp. 369--379 (2003)
13. Cocea M., Gutierrez-Santos S., Magoulas G. Case based Reasoning Approach to Adaptive Modelling in Exploratory Learning. Innovations in Intelligent Machines – 2. Studies in Computational Intelligence, (376), pp. 167--184. Springer, (2011)
14. Di Mascio T., Gennari R., Melonio A., Tarantino L. Supporting children in mastering temporal relations of stories: the TERENCE learning approach. Special Issue of IJDET on Visual Aspects in Technology Enhanced Learning. International Journal of Distance Education Technologies (in press).
15. Snowling M., Hume C. The science of reading: A handbook. 9. John Wiley & Sons (2008)
16. Gough P., Tunmer W. Decoding, reading, and reading disability. Remedial and Special Education (RASE), 7(1), pp. 6--10 (1986).
17. Nation K., Snowling M. Assessing reading difficulties: The validity and utility of current measures of reading skill. British Journal of Educational Psychology, 67, pp.359--370 (1997)
18. Elwér S., Keenan J., Olson R., Byrne B., Samuelsson S. Longitudinal stability and predictors of poor oral comprehenders and poor decoders. Journal of Experimental Child Psychology, 115(3), pp.497--516 (2013)

19. Shaywitz S. Overcoming dyslexia: A new and complete science-based program for overcoming reading problems at any level. New York, NY: Knopf (2003)
20. Oakhill J., Cain K., Bryant P. The dissociation of word reading and text comprehension: Evidence from component skills. Language and Cognitive Processes, 18, pp.443--468 (2003)
21. Geva E., Massey-Garrison A. A Comparison of the language skills of ELLs and monolinguals who are poor decoders, poor comprehenders, or normal readers. Journal of Learning Disabilities, 46 (5), pp. 387--401 (2013)
22. Vellutino F., Fletcher J., Snowling M., Scanlon D. Specific reading disability (dyslexia): What have we learned in the past four decades? Journal of Child Psychology and Psychiatry, 45, pp.2--40 (2004)
23. Nation K. Children's reading comprehension difficulties. In C. Hulme and M.J. Snowling, The science of reading, pp. 248--26, Oxford: Blackwel (2005)
24. Neale M. Neale Analysis of Reading Ability-Revised. Windsor: NFER-Nelson (1997)
25. Nation K., Snowling, M. Individual differences in contextual facilitation: Evidence from dyslexia and poor reading comprehension. Child development, 69(4), pp. 996--1011 (1998)
26. Cornoldi C., Colpo G. Prove di lettura MT per la scuola elementare [Tests of reading MT for primary school]. Florence, Italy: Organizzazioni Speciali (1998)
27. Cornoldi C., De Beni R., Pazzaglia F. Profiles of reading comprehension difficulties: An analysis of single cases. In C. Cornold and J. Oakhill (Eds), Reading comprehension difficulties: Processes and interventions, pp. 113--136. Mahwah, NJ: Erlbaum (1996)
28. Lyon G., Fletcher J., Barnes M. Learning Disabilities. In E. Mash and R. Barkley, Child Psychopathology. NY: The Guilford Press (2003)
29. Yuill N., Oakhill J. Children's problems in text comprehension: An experimental investigation. Cambridge University Press (1991)
30. Wauters L., Van Bon W., Tellings A. Reading comprehension of Dutch deaf children. Reading and Writing, 19(1), pp. 49--76 (2006)
31. Cain K., Oakhill J., Bryant P. Phonological skills and comprehension failure: A test of the phonological processing deficit hypothesis. Reading and Writing, 13(1-2), pp. 31--56 (2000)
32. Cain K., Oakhill, J. The nature of the relationship between comprehension skill and the ability to tell a story. British Journal of Developmental Psychology, 14(2), pp. 187--201 (1996)
33. Cain K. Making sense of text: skills that support text comprehension and its development. Perspectives on Language and Literacy, 35(2), pp. 11--14 (2009)
34. Cain K., Oakhill J., Lemmon K. Individual differences in the inference of word meanings from context: the influence of reading comprehension, vocabulary knowledge, and memory capacity. Journal of Educational Psychology, 96, pp.671--681 (2004)
35. Oakhill J., Hartt J., Samols D. Levels of comprehension monitoring and working memory in good and poor comprehenders. Reading and Writing: An Interdisciplinary Journal, 18, pp. 657--686 (2005)
36. Cain K., Oakhill J., Barnes M., Bryant P. Comprehension skill, inference making ability and their relation to knowledge. Memory and Cognition, 29, pp. 850--859 (2001)
37. Goldin-Meadow S., Mayberry R. How do profoundly deaf children learn to read? Learning Disabilities Research and Practice, 16(4), pp. 221--228 (2001)
38. Knoors H., Marschark M. Teaching deaf learners psychological and developmental foundations. Oxford University Press, pp. 230--231 (2014)
39. Lewis S. The reading achievements of a group of severely and profoundly hearing-impaired school leavers educated within a natural aural approach. Journal of British Association of Teachers of the Deaf, 20(1), pp.1--7 (1996)
40. Willcutt E., Pennington B. Psychiatric comorbidity in children and adolescents with reading disability. Journal of Child Psychology and Psychiatry, 41(8), pp.1039--1048 (2000)
41. Shaywitz S., Fletcher J., Holahan J., Shneider A.E., Marchione K.E., Stuebing K.K., Francis D.J., Pugh K.R., Shaywitz B.A. Persistence of dyslexia: the Connecticut Longitudinal Study at adolescence. Pediatrics, 104(6), pp.1351--1359 (1999)
42. Norman, D. The design of everyday things. Doubleday, New York (1998).
43. Sackett D.L., Rosenberg W., Gray J.A., Haynes R.B., Richardson W.S. Evidence based medicine: what it is and what it isn't. British Medical Journal, 312(7023), pp 71--72 (1996).

44. Arfé B., Oakhill J., Pianta E. The Text Simplification in TERENCE. In: F. De la Prieta , T. Di Mascio, R. Gennari and P. Vittorini (Eds.), ebuTEL 2013 workshop. Advances in Intelligent Systems and Computing 292. pp. 165--172 (2014)
45. Gennari R., Tonelli S., Vittorini, P. An AI-based Process for Generating Games from Flat Stories. In Proceedings of the 33rd SGAI International Conference on Artificial Intelligence (AI-2013), Cambdrige: UK (2013)
46. Alrifai M., De la Prieta F., Di Mascio T., Gennari R., Melonio A., Vittorini P. The Learners' User Classes in the TERENCE Adaptive Learning System. In Proceedings of International Conference on Advanced Learning Technologies, ICALT 2012, IEEE (2012)
47. Nanjappa A., Grant M.M. Constructing on constructivism: The role of technology. Electronic Journal for the Integration of Technology in Education, 2(1), pp 38--56 (2003)
48. Jong M.S., Lee J., Shang J. Educational use of computer games: Where we are, and what's next. In: R. Huang and J.M. Spector (Eds.). Reshaping Learning, New Frontiers of Educational Research, Springer Berlin Heidelberg, 1, pp 299--320 (2013).
49. Bruckman A., Bandlow A., Forte A. HCI For Kids. In: J. Jacko and A. Sears, Handbook of Human-Computer Interaction, 2nd Ed. NJ: Lawrence Erlbaum Associates (2007).
50. Soloway E., Guzdial M., Hay K.E. Learner-centered design: The challenge for HCI in the 21st century. Interactions, 1(1), pp 36--48 (1994).
51. Nielsen J. Usability laboratories: A 1994 survay. Behaviour and Information Technology 13(1-2), pp 3--8 (1994).
52. Redish J.G., Bias R.G., Bailey R., Molich R., Dumas J., Spool J.M. Usability in practice: formative usability evaluations-evolution and revolution. In CHI'02 extended abstracts on Human factors in computing systems, pp. 885--890. ACM (2002)
53. Gay L.R., Airasian P. Education Research: Competencies for Analysis and Application (6th ed.). Upper Saddle River, New Jersey: Merrill (2000).
54. Bailey R. User interface update. UI Design Newsletter (2001).
55. Di Mascio, T, Gennari, R, Vittorini, P. Expert-Based Evaluation: State of the Art and Results. Deliverable of the TERENCE project – ICT FP7 Programme – ICT-2010-25410 (2012)
56. Di Mascio T. Expert-Based Evaluation: Final Results. Deliverable of the TERENCE project – ICT FP7 Programme – ICT-2010-25410 (2013)
57. Di Mascio T., Gennari R., Vittorini P. Small Scale Evaluation design and results. Deliverable of the TERENCE project – ICT FP7 Programme – ICT-2010-25410 (2012).
58. Di Mascio T., Gennari R., Vittorini P. Large Scale Evaluation Design and Results. Deliverable of the TERENCE project – ICT FP7 Programme – ICT-2010-25410 (2013)
59. Di Giacomo D., Cofini V., Di Mascio T., Cecilia M.R., Fiorenzi D., Gennari R., Vittorini P. The silent reading supported by adaptive learning technology: influence in the children outcomes. Computers in Human Behavior (in press)
60. Cecilia MR, Vittorini P, Cofini V, di Orio F. The Prevalence of Reading Difficulties among Children in Scholar Age. Styles of Communication, 6(1), pp. 18--30 (2014).
61. Riffenburgh H. Statistics in Medicine. Third edition. Academic press (2012)
62. Hevner, A.R., March, S.T., Park, J. and Ram S. Design Science in Information Systems Research. J. of MIS Quarterly, 28(1), pp75—105 (2004)
63. Hourcade, J.P. Interaction design and children. Foundations and Trends in Human–Computer Interaction, 1(4), pp. 277–392 (2007)
64. Joiner, R., Messer, D., Light, P., Littleton, K. It is best to point for young children: A comparison of children's pointing and dragging. Computers in Human Behavior, 14(3), pp. 513–529, (1998)
65. Donker, A., Reitsma, P., Drag-and-drop errors in young children's use of the mouse. Interacting with Computers, 19, pp. 257–266 (2007)
66. Paivio, A. Dual-coding Theory: Retrospect and Current Status. Canadian Journal of Psychology, 45, pp. 255-287 (1991)
67. Harrison, S, Senger, P., Tarar, D, The three paradigms of HCI. In Proceedings of the 2007 SIGCHI conference on Human factors in computing systems, San Jose, CA, ACM Press, New York (2007)