iTalk2Learn
2013-10-31

Deliverable 5.1

Report on evaluation plan

31 October 2013

Project acronym: iTalk2Learn

Project full title: Talk, Tutor, Explore, Learn: Intelligent Tutoring and Exploration for Robust Learning

Work Package: **5**

Document title: D5.1-report_on_evaluation_plan

Version: **1.0**

Official delivery date: **31.10.2013**

Actual publication date:

Type of document: Report

Nature: Public

Authors: Katharina Loibl (RUB), Nikol Rummel (RUB), Manolis Mavrikis (IOE), Alice Hansen (IOE), Carlotta Schatten (UHi), Lars Schmidt-Thieme (UHi), Norbert Pfannerer (Sail), Gemma Solomons (Whizz), and Richard Marett (Whizz)

Internal reviewers: Beate Grawemeyer and Sergio Gutierrez-Santos (BBK), Carlotta Schatten and Lars Schmidt-Thieme (UHi)

Version	Date	Sections Affected
0.1	30/07/2013	First draft (RUB)
0.2	09/09/2013	Exploratory tasks (IOE)
		Summative evaluation (IOE)
		Sequencing for structured tasks (UHi)
0.3	20/09/2013	Speech recognition (Sail)
		Automatic switching between structured and exploratory tasks (RUB)
0.4	30/09/2013	Refinement of all sections after discussion with all partners at the general meeting
0.5	02/10/2013	Exploratory tasks (IOE)
		Speech recognition (Sail)
1.0	29/10/2013	Final version

Executive Summary

This deliverable reports on the iTalk2Learn evaluation plans. We first define our evaluation goals. Subsequently we describe two interlinked plans: the formative evaluation plan and the summative evaluation plan. The section on the formative evaluation plan describes the iterative design and test process involved in developing adaptive sequencing and support mechanisms, exploratory learning activities, and infant speech recognition. The summative evaluation plan defines appropriate methodologies to test the new technology in two experiments, using quantitative measures.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AM	Acoustic model
ELE	Exploratory Learning Environment
HCI	Human-Computer-Interaction
LM	Language model
LVF	Logotron Visual Fractions
M	Month
n	number
OOV	out of vocabulary rate
SIG	Special Interest Group
WER	Word error rate
WOZ	Wizard of Oz
Y	Year

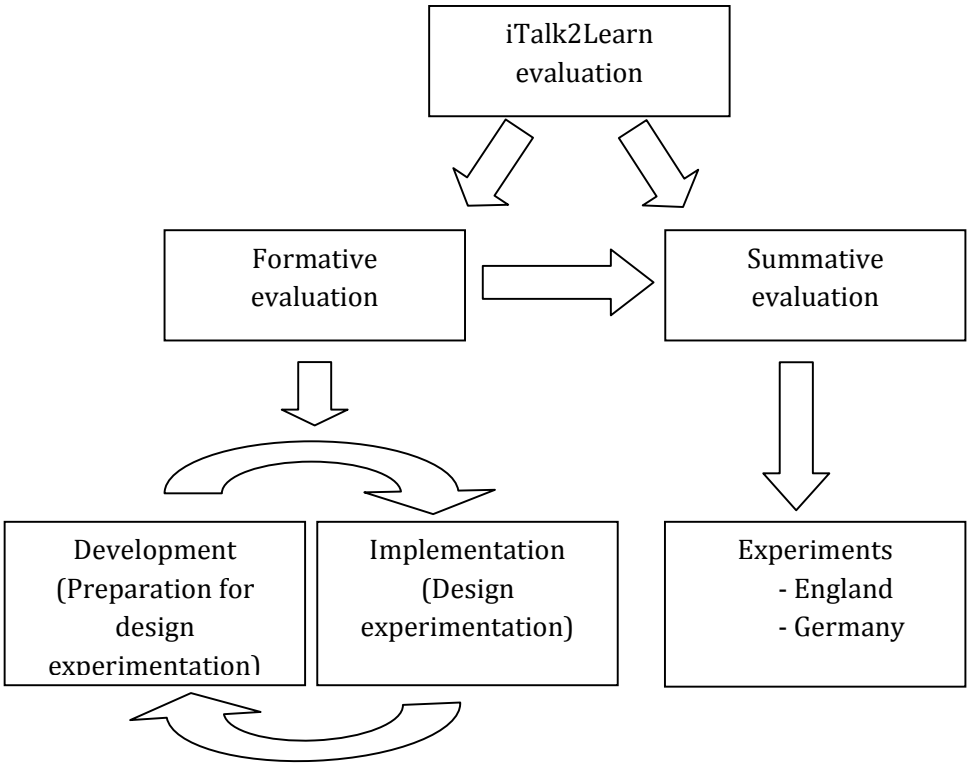
1. General Introduction

The iTalk2Learn project aims to facilitate robust learning in elementary education. Robust learning includes the acquisition of procedural skills and of conceptual knowledge (Koedinger, Corbett, & Perfetti, 2012). Research has shown that procedural skills can be best supported by structured practice (Rittle-Johnson, Siegler, & Alibali, 2001). In contrast, students acquire conceptual knowledge through experiencing more open-ended tasks, such as exploratory tasks (Ainsworth & Loizou, 2003; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Lewis, 1988; VanLehn, 1999; also see D1.1). Definitions of procedural and conceptual knowledge, as well as the interaction between them, are included in D1.1 and will be extended in D1.3.

Against this background, the iTalk2Learn project aims to facilitate robust learning in elementary education by creating a platform for intelligent support that combines existing structured learning tasks with new exploratory learning tasks, and that provides options for voice interaction. Intelligent components will adaptively sequence the learning tasks and provide adaptive support for students' as they interact with them. In order to enable learners to communicate more naturally with the interface and to reflect on their thinking, another strand of the project is the development of speech recognition for children.

We aim at evaluating all relevant components of the iTalk2Learn platform, namely exploratory learning tasks, speech recognition for young learners, and automatic adaptivity concerning sequencing and support. The components will be evaluated in iterative design and test cycles. The progress and outcome of the project will be evaluated by using formative and summative evaluation strategies as illustrated in Figure 1. The formative evaluation plan (see chapter 2.) describes the foreseen iterative process of developing, implementing, and testing the various components of the iTalk2Learn platform. The results of the formative evaluation will inform the design of the summative evaluation. The summative evaluation (see chapter 3.) will evaluate the pedagogical and technological outcomes of the project in two experiments that will be conducted in two proven application scenarios, in two European languages (English and German), and with quantitative learning measures. In the following we will describe the formative evaluation plan and the summative evaluation plan in more detail.

Figure 1: iTalk2Learn evaluation plan



2. Formative evaluation

The iTalk2Learn project aims to facilitate robust learning by creating a platform enriched with intelligent support that combines existing structured learning tasks with new exploratory learning tasks in elementary education and enables young learners to communicate more naturally with the interface.

Working on all three components i.e., exploratory learning tasks, automatic adaptivity (both sequencing and support), and speech recognition at once would lead to very high complexity and slow-down the progress of the project as the progress of one component is dependent upon the progress of the other components. We therefore work on these three interrelated components in parallel threads. In consequence, the formative evaluation plan foresees an evaluation of the work on these components separately. The main goal of the formative evaluation is to optimally inform the project about the current state of development of the various components of the iTalk2Learn platform. A second goal of the formative evaluation is to advance theoretical principles. As argued in DiSessa and Cobb (2004), 'grand' theories of learning (such as Piaget's theory) are not always precise enough to inform instructional design decisions. Similarly, other theoretical perspectives (referred to as 'orienting frameworks', DiSessa & Cobb, 2004) such as constructivism may provide general principles for conceptualising instructional design but lack the prescriptive power required to develop specific designs. Similarly, Self (1999) observes that such theories of learning or frameworks are often not adequate for facilitating the implementation of computational support and advocates for 'a mixture of

theory and empiricism' to inform the design of intelligent systems. This is exactly what we attempt in the formative evaluation phase of the project.

Following both goals (informing the project of the current state and advancing theoretical principles), we apply methodologies from the fields of Human-Computer-Interaction (HCI) and Educational Design Research. The common component between both methodologies lies in their iterative approaches: In the field of HCI, developing (educational) technology iteratively, means to repeatedly implement early versions of the developed technology to derive further specification for the improvement of the technology as well as to collaborate with the end users early in the design process (Preim, 1999). This user-centred approach builds on Nielsen's model of usability engineering (Nielsen, 1989) and aims at a high level of usability. The Educational Design Research methodology (also referred to as Design-Based Research) also involves several design phases and early collaboration with the end users (Cobb, Confrey, diSessa, Lehrer, & Schauble, 2003). The aim of Educational Design Research cycles is to bring about educational improvement through investigating instructional designs that are often significantly different to the typical forms of education. As a result, the researcher is more likely to identify relevant factors that contribute to the emergence of a 'new theory', and the products of design experiments are often innovative (Cobb et al., 2003). Edelson (2002) explains that the emergence of a new theoretical principles occurs as "design research explicitly exploits the design process as an opportunity to advance the researchers' understanding of teaching, learning, and educational systems" (p.107). He identifies three types of developments that result from design research: domain theories, design frameworks, and design methodologies. Educational Design Research includes a phase of *preparation* for design experimentation, a phase of conducting *design experimentation* and a phase of conducting *retrospective analyses* and their iteration.

Phase 1: Preparation for design experimentation

The purpose of phase 1 is to produce a "conjectured local instruction theory" (Gravemeijer & Cobb, 2007, p. 19), which will be refined through the second phase of the design experimentation approach. The first phase includes literature reviews, analyses of the state-of-the-art, and walk-throughs of preliminary versions of the to-be-developed components of the of the iTalk2Learn platform.

For most components of the iTalk2Learn project, the first phase already took place or is in progress at the time of this writing. Therefore, the steps within this phase are described in more detail than the steps for the remaining phases.

Phase 2: Conducting design experimentation

During phase 2, the effectiveness and usability of the developed educational system (in our case the iTalk2Learn platform) is piloted in iterative trials. Through this experimentation period involving iterative intervention and design cycles, theoretical principles are developed through the design and redesign of an educational system. The development of these principles takes place within a "learning ecology" (Cobb et al., 2003, p. 9) that takes the complex system of the specific educational setting and context into account. In iTalk2Learn, the learning ecology is set within a number of primary schools. The classroom is a dynamic environment in which the various aspects must be seen as continually acting and re-acting with each other. This dynamism negates the possibility of interpreting the classroom as a

situation where just a collection of activities or a list of separate factors influences learning and introduce the need to take into account a broader set of influencing factors to the new forms of learning.

The trials that will be conducted in this experimentation period will take place in the UK (with a subsample of Whizz users) and in Germany (with students recruited with the help of the RUB Schülerlabor).

Phase 3: Conducting retrospective analyses

The aim of this phase is to provide "resulting claims that are trustworthy" (Cobb et al., 2003, p. 13). It is essential that the outcomes of the retrospective analysis have been developed systematically through all levels and types of data from all iterations.

In the iTalk2Learn project the third phase will take place in the summative evaluation (see chapter 3). The summative evaluation will allow us to establish "trustworthy" claims. The third phase also allows for triangulation between all partners, producing valid and reliable findings. The results of this third phase will be reflected in the various Y2/Y3 deliverables and in the summative evaluation (D5.3). It will, however, not be part of the deliverable reporting on the formative evaluation (D5.2

In the remainder of this chapter we describe our formative evaluation plan in line with the Educational Design Research methodology and structured according to the above mentioned phases: Preparation for design experimentation (phase 1) and conducting design experimentation (phase 2). We describe the steps that we undertake in the respective phases for all components, that is, exploratory tasks, speech recognition, and automatic adaptivity. These components will be combined in the iTalk2Learn platform.

2.1 Exploratory learning environment and tasks

As mentioned above, the iTalk2Learn project aims at fostering robust learning, which consists of procedural skills and conceptual knowledge. In D1.1, we discussed that the acquisition of conceptual knowledge can be facilitated by engaging students in exploratory learning tasks. For this mean we develop an exploratory learning environment (ELE). In the ELE students work on exploratory tasks, which are designed by iTalk2Learn.

Phase 1: Preparation for design experimentation

In this phase iTalk2Learn is using a bootstrapping process for the development and evaluation of the ELE that brings together three sources: the literature, students' cognitive walk-throughs using paper-based tasks or tasks from related existing state-of-the-art software, and the partners' own design knowledge and expertise. This phase is nearing completion as of the writing of this deliverable. More specifically, up to now we have undertaken the following steps within phase 1: We have conducted literature reviews related to students' conceptual development in fractions and the pedagogy of fractions to inform the ELE and exploratory task designs. School-based trials have been an important aspect of this phase, working with students to analyse existing software, early iterations of the designed ELE and tasks, and gaining insight into students' understanding of fractions. We have also made use of

experts in mathematics education to act as critical friends throughout phase 1. More details are provided in Table 1.

Table 1: Development of exploratory learning environment (phase 1)

Month(s)	Tasks	Objective / Brief commentary
M4 – M6	Literature review on students' conceptual development on fraction, i.e. their interpretations of fractions and the representations of fractions.	The literature review fed into the Design Drivers to inform the design of the ELE in Phase One of the methodology (see D3.2 for a review of and explanation about the design drivers). It also informed the design of the tasks (more detail is provided in D1.1 and D1.2) that will be tested in Phase Two.
M6	Content analysis and observation of existing fractions software in use: Whizz, Fractions Tutor, and Logotron Visual Fractions (LVF)[1]	We met with four 10-11 year old children to compare currently existing activities and identify how the ELE will be integrated in these products to complement and build upon the present content.
M6	Presentation of initial ELE design ideas to the IOE Mathematics Education Special Interest Group (SIG)	Discussions with the IOE Mathematics Education SIG formed part of the initial bootstrapping process for Phase One, based on the design drivers (from the literature review) and our own design assumptions.
M6	Application to IOE Ethics Committee to undertake research in schools	Following the British Education Research Association's professional code of ethics, the application included a summary of the planned research, an overview of the participants, data collection and storage, and information for participants' parents. The application was approved.
M7	Trial and review of existing state-of-the-art software: Logotron Visual Fractions (LVF), Whizz	The trial with four 10-11 year old students fed into the Design Conjectures (reported in D3.2) and contributed to the ELE and tasks designs.

[1] Logotron Visual Fractions (see http://www.r-e-m.co.uk/logo/?Titleno=26562) is an existing product that utilizes a number of fraction representations. In this regard it is the closest to what we wish to achieve with models in the ELE. Using it at an early stage supported our understanding of how students use fractions models and how the ELE could effectively build on existing software.

	and Gizmo	
M7 – M8	Cognitive walk-throughs	Four students (10-11 years old) worked with bespoke-designed virtual and paper-based tasks with a teacher. Interaction and engagement with state-of-the-art software and conceptual understanding of fractions was analysed to support the design of the ELE and exploratory tasks.
M7 - M8	Design of ELE	Design document produced (see Appendix 1 of D3.2) for Testaluna as basis for the development of the prototype.
M9	First prototype of ELE trialled in school	Four 10-11 year old students' interactions with the prototype using fractions addition problems provided data for designing exploratory tasks and the first iteration of the ELE product for Phase Two.
M9 – M10	Designs of tasks produced	These were based on literature reviews and analysis of students' engagement with the existing state-of the-art software.
M10	Revision of ELE design based on trial	Communicated with Testaluna to develop the first iteration of the ELE product for Phase Two.
M11 – M12	Continue design of tasks under the assumptions of the design of the ELE	

Phase 2: Conducting design experimentation

The design experiment phase will begin now and continue throughout Y2 of the project. Phase 2 will involve iterative intervention and design cycles. Our initial intentions moving forward include trialing paper-based tasks with students (with different math ability levels). The findings will inform the final stage of designing the exploratory tasks. As soon as the next iteration of the ELE is available, the first exploratory tasks using the ELE will be trialled. In these trials we will use a Wizard of Oz (WOZ) method to simulate the intelligent components that are being designed in parallel. In WOZ studies, the design decisions are evaluated through a process where a human reacts to students' actions via the computer based on a script while students assume that the reactions come from the system. In this way WOZ studies allow testing the impact of certain design features before those features are technically implemented (Mavrikis & Gutierrez-Santos, 2010). Further iterative trials with increasing numbers of students will be conducted throughout the year as other features are integrated (such as speech recognition, support mechanisms) until the point at which the ELE is working most effectively for the project aims and the summative evaluation.

2.2 Speech recognition

In D3.1 we discussed that existing speech recognition developed for adults does not achieve the necessary accuracy on recognizing young children's speech. For the purpose of the evaluation of speech recognition for young learners, we are following the general methodology as outlined above.

Phase 1: Preparation for design experimentation

The first step towards developing a speech recognition system for young learners is to collect speech data of the same population as the later system user, i.e. students in Germany and UK. We collect data to feed into the statistical models for training and testing the speech recognition module (T3.3), and the behavioural data mining task (T3.4). The data collection is currently in progress. These speech corpora in English and German contain speech collected during problem-solving scenarios and transcripts representing the utterances as well as non-speech events which occurred during the recording (e.g. coughing, background-noises, filler words, hesitations, etc.). The collection process for both speech and interaction corpora is taking place in a staged manner, so as to be able to adjust collection processes to experiences made on initial sub-portions of the corpus. Data collection takes place at schools in an effort to resemble the envisaged setting where iTalk2Learn will be used.

Phase 2: Conducting design experimentation

The speech corpus collected in phase 1 will be used to train the models of the speech recognition system. The speech recognition system uses two statistical models, the acoustic model (AM), which models how the different sounds of the language are represented in the audio, and the language model (LM), which models probabilities of occurrences of words and sequences of words. The AM is trained from the collected recordings and transcripts; the LM is trained from a corpus of text, which will contain the transcripts of the speech data and additional text, e.g. from observing and writing down what the students say, handcrafting of possible phrases, or text grabbed from online forums or chat-rooms.

The training process makes use of a number of parameters that need to be set accordingly, e.g. how much the probability of certain keywords should be boosted to improve their recognition. After a model has been built, it needs to be evaluated to assess its performance. In an iterative manner the parameters will then be adjusted and the process repeated.

To evaluate one of these models (AM or LM), a portion of the data, maybe 10%, is set aside and not used for training. The resulting model is then tested on the 10% of the data that was set aside. In case that only little training data is available, and to achieve the best possible performance of the whole system it is advisable to use all of it in the training process, the above described testing process can be repeated with different sets of 10% of the data, and once an adequate set of parameters has been found, train a final model using all training material.

The LM can be built much more rapidly and easier than the AM, and we envision to train and test multiple LMs iteratively, and for different contexts. To evaluate a language model, we will compute two measures of the LM, the so-called perplexity and the out of vocabulary rate (OOV) on the held out data.

The purpose of the LM is to assign probabilities to the next possible word, dependent on the partial utterance up to a given point in time. In other words, the LM tries to predict the next word. We measure how well this prediction fits the test data, which is called "perplexity".

An important part of the LM is the vocabulary. The vocabulary determines what can be recognized - any word NOT part of the vocabulary cannot be recognized. Therefore any word, that will be said by the children, and which is not in the vocabulary, will lead to at least one recognition error, possibly more, because the LM will lose its context and has a lesser chance to correctly predict the following words. But we might want to use not all of the words from the training corpus, to reduce the chance that words are acoustically similar and generate higher probability of errors. Additionally, depending on the origin of the corpus, not all words will be relevant, e.g. there might be typos if the text was grabbed from online forums or chat rooms (online forums or chat rooms may be possible sources for the text corpus that we need to look at in more details). The design of the vocabulary also has to take into account the indicators needed by the automatic adaptivity, e.g. fillers or hesitations have to be part of the vocabulary. The OOV-rate measures how well the vocabulary fits the test data, specifically how many words of test set were missed by the selection of the vocabulary.

The AM cannot easily be tested alone, because in the final system it is always combined with a LM. To evaluate its quality, log files of the training process will be analyzed, which show whether the training process converges to the training data. Then a combined system containing an AM and an LM will be tested. The standard indicator is the word error rate (WER), which is the percentage of mis-recognized words in the test set. For this aim (i.e., testing the speech recognition system) we transcribe the students' utterances and compare these transcripts to the outcome of the speech recognition system. In our setup, where we envision to use speech recognition, for instance, to assess the children's use of the correct mathematical terminology, and are therefore interested in these keywords only, the usual measures are precision and recall. These are calculated from the two possible types of error, an uttered keyword was not recognized, which is called a false negative; or a keyword was recognized when it was not said, which is called a false positive. Depending on the usage scenario one of these types of errors may be more severe than the other; the training parameters can then be adjusted to find the optimal operating point. Precision is the percentage of correctly recognized keywords among all recognized keywords, whereas recall is the percentage of correctly recognized keywords among all uttered keywords.

Speech recognition will be included in the development of automatic adaptivity as described below.

2.3 Automatic adaptivity

As described in the introduction, the iTalk2Learn project aims to support the acquisition of robust knowledge, which includes procedural skills and conceptual knowledge (Koedinger et al., 2012). Procedural skills can be fostered by structured practice; conceptual knowledge by learning with exploratory tasks. Thus, we aim to develop a platform including both structured practice and exploratory activities by building on and enhancing existing tutors and developing new components (i.e., the ELE as described above). The platform will further facilitate natural interaction by speech recognition.

The development of automatic adaptivity for the iTalk2Learn platform will focus on three threads:

1) Concerning structured practice we integrate existing tutors (Whizz and Fractions Tutor; however, Fractions Tutor must first be translated and adapted to the needs of German students) in our iTalk2Learn platform. To improve these tutors we use recommender technology to develop an adaptive content sequencer which selects the order of tasks in dependence of students learning process (see 2.3.1).

2) In addition, we focus on how to best combine structured practice and exploratory activities in order to reach our goal of fostering robust learning (see 2.3.2). In other words, when should students switch from structured practice to exploratory tasks, and vice versa.

3) Finally, we aim at guiding students during the whole learning process (see Table 2 and 2.3.3). With regard to structured tasks the existing tutors already provide task dependent hints to students. For our newly-developed exploratory learning tasks we aim to also develop task-dependent support. In addition, we plan to develop task-independent support that takes students' utterances into account (for instance, we envisage that the system could prompt students to use mathematical terminology).

All threads will include work with and without indicators from speech.

Table 2: Task-dependent and task-independent support

	Task-dependent support	**Task-independent support**
Structured Tasks	As provided by Math-Whizz and Fractions Tutor	To be implemented by iTalk2Learn (for instance, encouragement to use mathematical terminology).
Exploratory Tasks	To be implemented by iTalk2Learn: Hints and feedback during exploratory tasks	

2.3.1 Sequencing for structured tasks

As indicated above, the first component of automatic adaptivity aims at developing an adaptive content sequencer for the iTalk2Learn platform.

Phase 1: Preparation for design experimentation

To reach the goal of an adaptive content sequencer we build a probabilistic model that selects tasks based on recordings of learner-tutor interactions. Each new observation (i.e., interaction with the tutor) can be utilized to shape the model to select tasks that maximize our goal of fostering robust learning. Realizing adaptive sequencing is not easy, because the quality of the sequencer needs to be evaluated online (i.e., while students are interacting with the system).

One possible approach would be to record a dataset that explores the relationship between sequence of the content and learning (i.e. exploring the majority of possible element combination of a sequence). This would mean that all students (high achievers and low achievers) would have to solve very difficult tasks as well as easy tasks in a variety of combinations. This process could easily frustrate the young learners of our target population (Chi Min et al., 2011). We will therefore initially implement the sequencer exploiting performance prediction methods, by building a model using historic data. With historic data we refer to log files of the structured tutors that are already available. A similar approach was used by Koedinger and colleagues (2011). However, we will need to collect new data from students actually using the system with the developed sequencer to test its quality. This will be part of the second phase (i.e., conducting the design experiment).

At the beginning, the adaptive sequencing will consider performance measurements (for instance, success and error rate, time needed etc.). We will conduct an analysis of historic Whizz data in order to evaluate which log files can be used to create an efficient student model for performance prediction. Corrupted data (e.g., a negative number for age) will be removed and log data of students "gaming the system" (i.e., students not really learning but engaging for example in trial and error behavior to finish the assigned tasks) will be discussed. The analysis of Whizz data also involves evaluating possibilities of creating metadata that could give us more information about the actual condition of the user. Possible information in the metadata could include the error frequency, the tendency to ask help etc. This information will be utilized to bias the performance predictor in order to obtain a more precise prediction. The same procedure will be undertaken for Fractions Tutor; however, in this case we will already need to collect some new data in this first phase because historic data of the Fractions Tutor can only be used in a very limited way for two reasons: 1) We will translate the Fractions Tutor and adapt it the German students' needs. These modifications may influence the fit between the model developed with historic data and data collected with the modified Fractions Tutor. 2) The Fractions Tutor does not present the tasks in many different orders. Data from different orders of the tasks is needed to develop the model.

Finally, we will integrate information gained by the speech recognition in the model to adapt the sequence accordingly. In the initial phase, we evaluate existing information from state-of-the-art methods and Sail's previous experience in speech analysis. For example, if we can detect utterances by the students indicating that they perceive the task as being too easy, we will bias the task selection model in a way, that students subsequently receive a more difficult task. The model will be built with the collected and preprocessed iTalk2Learn data.

Phase 2: Conducting design experimentation

In a second step the developed sequencer will be implemented in the iTalk2Learn platform. As already said the model will be built with the preprocessed data and tested with some elementary students recruited with the help of the Whizz user data base in England and the RUB Schülerlabor in Germany. In this test phase we will collect the interactions in log files to evaluate students learning process and report differences between the ones using the adaptive version and the ones using the original versions of Whizz or Fractions Tutor.

We will further evaluate whether recognizing students' speech provides useful information for sequencing the tasks. In this regard we aim at testing whether the use of speech indicators allows us to model the knowledge acquisition of the student. We expect that by including speech indicators into the model of the sequencer we will be able to provide an improved trajectory of contents to the students.

Requirements and state-of-the-art for adaptive intelligence will be reported in D2.1. A first and more advanced prototype of the adaptive sequencer will be delivered and commented in D2.2.1 and D2.2.2 respectively. The outcomes of the evaluation including students' speech will be reported in D3.4.

2.3.2 Automatic switching between structured and exploratory tasks

In order to know how to combine structured and exploratory task we evolve an intervention model. The theory-based intervention model will be described in D1.3. It will serve as framework for answering two main questions:

1. Should students initially start the learning sequence with exploratory learning tasks or structured practice?
2. When should students switch between the two different types of tasks (i.e. from structured practice to exploratory learning tasks, or vice versa)?

Phase 1: Preparation for design experimentation

With regard to the first question we conduct a literature review focusing on instructional approaches which with different sequences of structured and exploratory learning tasks (e.g., Van Merriënboer, Clark & de Croock, 2002; Reigeluth, Merrill, Wilson & Spiller, 1980). In particular, we focus on a current scientific debate about timing of exploratory tasks, structured tasks, and instruction (e.g., Kapur, 2008; Kapur & Bielaczyc, 2012; Kirschner, Sweller & Clark, 2006; Schwartz & Bransford, 1998). More detail will be provided in D1.3.

In order to answer the second question we need to determine when switching between structured practice and exploratory tasks is useful. With regard to this goal, we aim at identifying specific learning indicators, which show, for instance, that the respective student has already developed (enough) procedural or conceptual knowledge and hence might not need to repeatedly engage in the same type of tasks. In such a situation switching to the other type of task may provide students with further learning opportunities. For identifying such learning indicators (e.g. decrease of time spent on a single task for procedural knowledge gains), we are reviewing literature from the field of HCI in general and from the field of educational technology in particular. In addition to learning indicators that can be identified in students' performance on a task, we envisage that the system could also adapt its recommendation for the next tasks based on utterances that inform about students' perception on tasks (difficult, boring etc).

Building on the (theoretical) intervention model, we will develop methods to automatically switch between structured and exploratory tasks as the learner progresses within the iTalk2Learn platform. No state-of-the-art literature is available for machine learning that could be applied to this task. Thus,

accomplishing this task requires a lot of innovative work by several partners. In particular, UHi, RUB, and IOE will collaborate strongly to develop a method for automatic switching.

Phase 2: Conducting design experimentation

In order to underline our theoretical considerations we intend to test different sequences with small samples of students. These tests will provide further indicators to refine our intervention model.

Furthermore, we will test the automatic switching between structured tasks and exploratory tasks in iterative design cycles with a small set of students. The first trials will evaluate switching based on indicators derived from students' actions in the system. Later trials may additionally include indicators derived from speech. We propose to evaluate automatic switching by measuring the number of successfully solved tasks and the required time. In addition, we will implement a posttest. The descriptive statistics of these measurements will provide feedback for the refinement of switching between structured tasks and exploratory tasks.

2.3.3 Support (Hints and Feedback)

As indicated earlier (see Table 2), support (i.e., hints and feedback) can be given at two levels: task-dependent and task-independent.

Task-dependent support:

Phase 1: Preparation for design experimentation

Regarding structured tasks, the tutors (i.e., Whizz and Fractions Tutor) already provide hints depending on the student's progress. As already mentioned, we will leave the hint functionalities of the existing tutors intact.

For the exploratory learning tasks, we will develop task-dependent support because research on guided discovery learning has shown that support is a prerequisite for learning in these settings (e.g. van Joolingen, de Jong, Lazonder, Savelsbergh, & Manlove, 2005; also see D1.1). Thus, the development of adaptive support is highly interlinked with the development of the exploratory learning tasks described in 2.1. Similar to the above components, we derive indicators for when to provide what kind of support from a literature review and from our early observations with students. The development of the exploratory tasks that has been described above also helps toward deriving initial information that facilitate the design of the task-dependent support to be provided in the ELE.

Phase 2: Conducting design experimentation

With respect to the task-dependent support within the exploratory learning tasks, we will undertake WOZ studies. These studies act as both design and evaluation studies. Evaluating the performance of ELE support is a difficult question by itself; traditional strategies, like comparing a guided version to an unguided version of the exploratory learning tasks is not a valid approach because it is well-understood that an unguided version will not result in any productive learning, that is, in any event support will be better than no support. IOE and BBK have devised a methodology to evaluate the performance of

intelligent support in exploratory environments (exemplified in Mavrikis et al., 2012) and it will be used in the context of iTalk2Learn. The methodology relies on various metrics that allow identifying design problems and measuring students' perception of the intelligent support at various implementation stages. For example, we have employed the metric of 'relevance' as a measure of how many support interventions made by the system were relevant for the student i.e. where appropriate to the situation as judged by experts (our only available gold standard). We will adapt this methodology according to the requirements of this project.

In addition, we will aim to collect information on children's opinion on the exploratory environment in general but also gauge their perception of the intelligent support in particular, in order to help us make design decisions such as the actual messages, their appearance and the general approach we are taking.

Task-independent support:

Phase 1: Preparation for design experimentation

Task-independently (i.e., for structured tasks and exploratory tasks) we will provide further support building on advanced behavioral interaction interpretation (i.e., speech). For instance, we aim at deriving indicators for perceived task difficulty (e.g., students stating that the task is easy or difficult) and other affective factors (e.g. boredom or frustration) that can be used as additional information to adapt the task-independent support. We further envisage detecting students' use of mathematic terminology and provide feedback accordingly.

In order to derive speech indicators, we first record what students utter when working with the system. Additionally, we compile vocabulary lists for a) possible utterances regarding perceived task difficulty and b) relevant mathematic terminology. Both lists will be integrated in the LM and the AM of the speech recognition system. The speech recognition system will be trained to detect the words from these lists. In this regard, we have to run several training and testing cycles of the speech recognition system to find the optimal balance between boosting these words in order to ease their detection without increasing the probability of a false detection too much.

Phase 2: Conducting design experimentation

With respect to the task-independent support, we will test the developed support (e.g., feedback regarding the use of mathematics terminology, affective aspects, and perceived difficulty) with a small sample size in iterative design cycles. We will first conduct WOZ studies to test the general effect of the developed support (without relying on the system), both with and without including speech indicators. Afterwards the design experimentation will include tests with the "real" speech recognition system as described above.

3. Summative evaluation

In the summative evaluation the parallel developments of the projects (i.e., exploratory tasks, automatic adaptivity, and speech recognition) will be brought together. All components are combined in a unifying

platform. This platform will allow integrating existing Tutors for structured practise (i.e., Whizz or Fractions Tutor) and combining them with the developed ELE for exploratory tasks. The platform will include an automatic sequencer for sequencing the structured tasks and for switching between structured tasks and exploratory tasks. Furthermore, it will provide task-independent support. All adaptive components of the platform (i.e., sequencing structured tasks, switching between structured and exploratory tasks, and task-independent support) will work with and without speech indicators. For more information on the unifying platform see D4.1. The task-dependent support features described above are embedded in the existing Tutors for structured tasks and will be embedded in the ELE for the exploratory tasks. These features rely on students' actions in the system (i.e., they are not based on speech indicators).

As described in chapter 2, our formative evaluation plan includes phase 1 (Preparation for design experimentation and phase 2 (Conducting design experimentation) of the Educational Design Research methodology. The third phase (Retrospective analysis) of the Educational Design Research methodology will be conducted within the summative evaluation. This phase aims to evaluate the pedagogical and technological outcomes of the project in order to derive claims that are trustworthy. In particular, we will investigate two hypotheses: 1) Combining structured practice and exploratory tasks promotes robust learning. 2) Indicators from speech will enhance the automatic adaptivity. To test these hypotheses, we compare multiple versions of the unified platform as displayed in Table 3. Students will be randomly assigned to one of the conditions (i.e., the will work with one of the versions).

Table 3: Conditions in the experiments of the summative evaluation

	Full version with speech	**Full version without speech**	**Version without ELE**
Sequencing structured tasks[2]	Yes, with speech indicators.	Yes, without speech indicators.	Yes, with or without speech indicators.
Switching between structured and exploratory tasks[3].	Yes, with speech indicators.	Yes, without speech indicators.	No.
Task-independent support	Yes, with speech indicators.	Yes, without speech indicators.	Yes, with or without speech indicators.

Comparing the version without ELE and the full version without speech to the full version with speech allows testing the two mentioned hypotheses, which are described in more detail below:

[2] The structured tasks include task-dependent support as provided by Whizz or Fractions Tutor.
[3] The exploratory tasks include task-dependent support, which will be developed as described in 2.3.3.

Hypothesis 1: Robust learning requires procedural skills and conceptual knowledge. Switching between structured tasks and exploratory tasks is required to foster both types of knowledge. Thus, a system including exploratory learning task (i.e., a full version) will foster robust learning in comparison to systems with structured tasks only (i.e. a version without ELE). To test hypothesis 1, we will implement a control condition without the newly developed exploratory learning environment (i.e., version without ELE). We will make sure that this condition only differs with regard to the type of task, not with regard to the content that can be learnt with the system.

Hypothesis 2: Speech recognition allows identifying learning indicators that cannot be measured otherwise. Including these indicators to the adaptivity of the system (i.e., the full version with speech) will foster learning in comparison to an adaptive system without speech-based learning indicators (i.e. the full version without speech). To test hypothesis 2 we will implement a control condition that does not use speech recognition to adapt sequencing, switching, and task-dependent support (i.e., full version without speech). In this condition, the adaptivity will be based only on students' keyboard entries.

As mentioned in the DOW, the pedagogical and technological outcomes of the project will be evaluated in two application scenarios in two European languages with quantitative learning measures. One experiment will take place in a controlled setting (i.e., lab study) in Germany using content from the Fraction Tutor for the structured tasks. The other experiment will take place in the UK using the platform online (i.e., in a more open and less controlled setting) using content of Whizz for the structured tasks. In both scenarios, Whizz or Fractions Tutor will be combined with the developed ELE for the full versions (see Table 3).

The experiments will examine the general effectiveness of the adaptive intelligent support for robust learning and test the aforementioned hypotheses. In both experiments, we will test for learning outcomes. In order to be able to compare the results of both experiments we unify the metrics as follows: We choose system inherent learning indicators that can be logged by both systems, Whizz and Fraction Tutors as well as the developed ELE. Moreover the post-test will constitute of the same, yet translated, items measuring procedural skills, and conceptual knowledge. Items testing for procedural skills will present problems isomorphic to the ones students worked on with the system. Thus, these items will require students to apply the procedures they learnt during the interaction with the system. Items testing for conceptual knowledge will ask for explanation to identify students understanding of the learnt concepts, and ask students to adopt the learnt procedure to unfamiliar problems. The post-test will be piloted with students from our target group.

In addition to student learning, we will implement surveys measuring students' satisfaction with the system and their motivation/engagement. We will rely on metrics appropriate for students in this age. In previous projects, we have employed a 5-point Likert scales appropriate for children to evaluate important constructs including helpfulness, repetitiveness comprehension and affect (see Mavrikis et al. 2012) using a visual analogue scale that employs pictorial representations that children can relate to (eg. the Fun Toolkit in Read, MacFarlane, & Casey, 2002)..

3.1. Experiment in Germany

As already indicated above, RUB will evaluate the iTalk2Learn platform in a controlled setting in Germany. For the experiment, we will recruit children from elementary schools (n = 30 per condition) with the help of the "Schülerlabor" of the Ruhr-Universität Bochum. The RUB Schülerlabor is an extra-curricular location where classes can spend a day working on a specific well-prepared topic that goes beyond the school curriculum. Due to these activities, the RUB Schülerlabor has many school contacts that will be activated to recruit students. Students will work with different versions of the system (cf. Table 3) and we will collect their interactions in log files as well as their learning outcomes measured by a (paper-based) post-test.

3.2. Experiment in England

The experiment in England will be conducted online through the iTalk2learn platform. We will advertise through emails to IOE contacts and Whizz customer base (that fit our target population). We will further rely on previous agreements with two English schools and other contacts that we are making as the project progresses. This way we hope for a substantial number of students (around 40 per condition). Note that this experiment may take place in a less controlled setting (e.g. students may be asked to interact with the platform in a longer period of time from home) so the results will be qualified accordingly.

It is worth stating that the implementation of a speech recognition system in an uncontrolled setting is a risky task due to the complexities and uncertainties behind the speech recognition software working in varied, uncontrolled contexts. Thus, the experiment will provide insights into the ecological validity of our approach.

4. Conclusion

To conclude, we foresee to evaluate the iTalk2Learn progress and outcomes in two ways. First, we evaluate the process as described in the formative evaluation plan (chapter 2) using Educational Design Research methodologies. During the formative evaluation, the components of the iTalk2Learn platform (i.e. exploratory tasks, automatic adaptivity, and speech recognition) will be developed and evaluated in an iterative process of *design experimentation*. We will work on the components in parallel to ensure that the project progresses in time. The results of the formative evaluation will be reported in D5.2. These results will inform the project about progress and the usability of the developed components. Moreover, the results of the formative evaluation will allow us to derive specific hypotheses for the experiments conducted as part of the summative evaluation.

The summative evaluation will test the hypotheses described above (that will possibly be modified and specified dependent on the results of the formative evaluation using quantitative methods with larger sample size. To test the generalizability of our outcomes and the applicability of the iTalk2Learn platform to different educational settings, the summative experiments will be conducted in two

European countries (Germany and UK), using different tutorial systems for the structured tasks (Fractions Tutor and Whizz) in two settings (controlled laboratory setting and open online or classroom setting). The results of the summative evaluation will be reported in D5.3.

References

Ainsworth, S., & Loizou, A. T. H. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*, 669-681.

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145-182.

Chi, M., VanLehn, K, Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical tactics. *User Modeling and User Adapted Instruction(UMUAI), 21*, 137-180.

Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiment in Educational Research. *Educational Researcher, 32*(1), 9-13.

diSessa, A., & Cobb, P. (2004). Ontological Innovation and the Role of Theory in Design Experiments. *The Journal of the Learning Sciences, 13*(1), 77-103.

Edelson, D. C. (2002). Design Research: What we learn when we engage in design. *Journal of the Learning Sciences, 11*(1), 105-121.

Gravemeijer, K. & Cobb, P. (2006). Design research from a learning design perspective, In: J. van den Akker, K. Gravemeijer, S. McKenney & N. Nieveen (Eds.) *Educational Design Research* (pp. 17–51). London: Routledge.

Kapur, M. (2008). Productive failure. *Cognition and Instruction, 26*(3), 379-424.

Kapur, M. & Bielaczyc, K. (2012): Designing for Productive Failure, *Journal of the Learning Sciences, 21*(1), 45-83.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist, 41*(2), 75-86.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798.

Koedinger, K.R., Pavlik Jr., P.I., Stamper, J.C., Nixon, T., & Ritter, S. (2011). Avoiding problem selection thrashing with conjunctive knowledge tracing. In *Proceedings of the Fourth International Conference on Educational Data Mining* (pp. 91-100).

Lewis, C. H. (1988). Why and how to learn why: Analysis-based generalization of procedures. *Cognitive Science, 12*, 211-256.

Mavrikis, M., & Gutierrez-Santos, S. (2010) Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering. *Computers & Education, 54*(3), 641-651.

Mavrikis, M., Gutierrez-Santos, S., Geraniou, E., & Noss, R. (2012).Design requirements, student perception indicators and validation metrics for intelligent exploratory learning environments. In *Personal and Ubiquitous Computing*, Available online at http://dx.doi.org/10.1007/s00779-012-0524-3

Nielsen, J. (1989). Usability Engineering at a Discount. In G. Salvendy & M.J. Smith (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge-Based Systems* (pp. 214-221). Amsterdam: Elvisier Science.

Preim, B. (1999). *Entwicklung interaktiver Systeme. Grundlagen, Fallbeispiele und innovative Anwendungsfelder [Development of interactive systems. Basic principles, case studies, and innovative fields of application].* Berlin Heidelberg: Springer-Verlag.

Read, J., MacFarlane, S., & Casey, C. (2002), Endurability, engagement and expectations: Measuring children's fun. In *Proceedings of Interaction Design and Children* (pp. 189–198), Eindhoven, Shaker Publishing.

Reigeluth, C.M., Merrill, M.D., Wilson, B.G., & Spiller, R.T. (1980). The Elaboration Theory of Instruction: A model for sequencing and synthesizing instruction. *Instructional Science, 9*(3), 195-219.

Rittle-Johnson, B., Siegler, R., & Alibali, M. (2001). Developing conceptual understanding and procedural skill in mathematics: *An iterative process. Journal of Educational Psychology, 93*(2), 346-362.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction, 16,* 475–522.

Self, J. (1999). The defining characteristics of intelligent tutoring systems research: ITSs care, precisely. *International Journal of Artificial Intelligence in Education, 10,* 350-364

van Joolingen, W. R., de Jong, T., Lazonder, A. W., Savelsbergh, E. R., & Manlove, S. (2005). Co-Lab: research and development of an online learning environment for collaborative scientific discovery learning. *Computers in Human Behavior, 21*(4), 671–688.

VanLehn, K. (1999). Rule learning events in the acquisition of a complex skill: An evaluation of Cascade. Journal of the Learning Sciences, 8(1), 71-125.

Van Merriënboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for complex learning: The 4C/ID-model. *Educational Technology Research and Development, 50,* 39-64.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication, 48*(9), 1162-1181.

Worsley, M., & Blikstein, P. (2011). What's an expert? Using learning analytics to identify emergent markers of expertise through automated speech, sentiment and sketch analysis. In *Proceedings of the Fourth International Conference on Educational Data Mining* (pp. 235-240).