



D5.2 Report on formative evaluation results in Y2



iTalk2Learn
2014-10-31

Deliverable 5.2

Report on formative evaluation results in Y2

31 October 2014



D5.2 Report on formative evaluation results in Y2

Project acronym: iTALK2Learn

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

Work Package: 5

Document title: D5.2-report_on_formative_evaluation_results in Y2

Version: 1.0

Official delivery date: 31 October 2014

Actual publication date: 31 October 2014

Type of document: Report

Nature: Public

Authors: Julia Erdmann (RUB), Michael Wiedmann (RUB), Katharina Loibl (RUB), Nikol Rummel (RUB), Alice Hansen (IOE), Wayne Holmes (IOE), Manolis Mavrikis (IOE), Carlotta Schatten (UHi), Ruth Janning (UHi), Beate Grawemeyer, (BBK), Gerhard Backfried (Sail)

Reviewers: Carlotta Schatten (UHi), Ruth Janning (UHi), Beate Grawemeyer (BBK), Alice Hansen (IOE), Wayne Holmes (IOE), Manolis Mavrikis (IOE)

Version	Date	Sections Affected
0.1	05/09/2014	Initial version (RUB)
0.2	12/09/2014	Including contributions from IOE, UHi; BBK, and SAIL
0.3	21/10/2014	Refinement of all sections after discussion with all partners at the general meeting
0.4	22/10/2014	Review comments from UHi processed
0.5	24/10/2014	Review comments from BBK and IOE processed
1.0	28/10/2014	Final version



D5.2 Report on formative evaluation results in Y2

Executive Summary

The iTALK2Learn project aims to facilitate robust learning in elementary education by creating a platform for intelligent support that combines existing structured tasks with new exploratory tasks, and that provides options for voice interaction. We aim at evaluating the relevant components of the iTALK2Learn platform, namely exploratory learning environment, speech recognition for young learners, and automatic adaptivity concerning sequencing the tasks, switching between exploratory and structured tasks, and support functionalities. The effectiveness and usability of the components are evaluated in iterative design and test cycles. In order not to slow down the development process, we conducted the formative evaluation separately for each of the main developments of the iTALK2Learn project. This deliverable reports on the formative evaluation activities conducted in Y2.

The iTALK2Learn project developed the exploratory learning environment Fractions Lab that facilitates conceptual learning by engaging students in exploratory tasks (see D1.1, D1.2 and D3.2). The design of the user interface and the exploratory tasks has been conducted in close collaboration with teachers and students. Formative evaluation shows that the user interface is well received by both teachers and students who find most of Fractions Lab affordances intuitive. Results of formative trials show that Fractions Lab challenges students' conceptual thinking and helps students to extend the range of fractions representations they work with.

The iTALK2Learn project integrated two existing tutoring environments for structured learning (Maths-Whizz and Fractions Tutor) in the iTALK2Learn platform (see D4.2.1). iTALK2Learn developed a task sequencer based on performance prediction that is compatible with both Maths-Whizz and Fractions Tutor (see D2.2.1). This so-called Vygotsky Policy Sequencer has been shown to effectively support procedural learning with Maths-Whizz and to provide substantial benefits over traditional, curriculum-based sequencers. The adaption of the sequencer to Fractions Tutor is ongoing.

A core component of the iTALK2Learn platform is a speech recognition system for children (see D3.3.1). For training the speech recognition system, extensive speech datasets have been collected in both German and English (see MS5). A first speech recognition model for English has been trained and will be refined through focusing on relevant vocabulary. Speech recognition ties into the iTALK2Learn platform in several ways. It contributes to the prediction of students' performance on learning tasks and their affective states and influences sequencing of tasks and task-independent support. Data features have been identified and feature analysis showed that the developed features can be used for affect recognition. A preliminary machine-learning model has already been trained that can assess student affect.

In addition, iTALK2Learn has developed two support components: task-dependent support for exploratory tasks and task-independent support for both exploratory and structured tasks. A set of rules for task-dependent support for Fractions Lab has been developed that supports problem-solving and that can be delivered automatically by the system. The task-independent support uses the



D5.2 Report on formative evaluation results in Y2

transcribed speech text to provide feedback according to the students' use of mathematics vocabulary in tasks as well as their affective states. In Wizard-of-Oz studies, where humans simulated the computer-based adaptivity, we found that reminding students to use appropriate mathematics vocabulary might help them to think through the problem, reflect and improve their understanding. Also, students responded better to support that was tailored to their affective state. We therefore developed a set of rules for mathematics vocabulary and affect-based support that can be delivered across all different types of tasks implemented in the iTALK2Learn platform.

A further major contribution of iTALK2Learn is the combination of exploratory tasks (i.e., Fractions Lab) and structured tasks (i.e., Maths-Whizz and Fractions Tutor) to foster robust learning. iTALK2Learn developed an intervention model to guide decisions on how to sequence structured and exploratory tasks and when to best switch between the different learning environments (D1.3). On the basis of this, intervention model trials in Germany and the UK are planned as a last step in the formative evaluation trials.

The results of the formative evaluation informed the iterative development process, and form the basis for the summative evaluation in Y3. In the summative evaluation the integration of all components into the unified iTALK2Learn platform and its effectiveness will be evaluated. The final section of this deliverable is dedicated to implications and an update of the summative evaluation plan based on the formative evaluation results. The results of the summative evaluation will be reported in D5.3.



D5.2 Report on formative evaluation results in Y2

Table of Contents

Executive Summary.....	3
Table of Contents.....	5
1. General introduction	8
1.1 Overview of evaluation plan.....	9
2. Results of formative evaluation.....	10
2.1 Fractions Lab and Exploratory Tasks	10
2.2 Automatic adaptivity	15
2.2.1 Speech recognition - Development of the speech recognition system for children.....	16
2.2.2 Sequencing of structured tasks.....	18
2.2.3 Amelioration of performance prediction and sequencing through affect recognition.....	20
2.2.4 Task-dependent support for Fractions Lab	22
2.2.5 Task-independent support based on speech indicators.....	25
2.2.6 Switching between exploratory tasks and structured tasks.....	26
2.3 Overall student perception and feedback.....	29
3. Summative Evaluation.....	30
3.1 Conclusions and lessons learned from the formative evaluation trials.....	30
3.2 Updated summative evaluation plan	34
3.2.1 Participants	35
3.2.2 Research design.....	35
3.2.3 Measures	36
3.2.4 Procedure.....	37
3.2.5 Risks of the summative evaluation.....	37



D5.2 Report on formative evaluation results in Y2

4. Conclusion	39
References	40
Appendix 1.....	42

List of Figures

Figure 1: Architecture of the iTalk2Learn iTalk2Learn platform



D5.2 Report on formative evaluation results in Y2

List of Tables

Table 1: Fractions Lab evaluation studies

Table 2: Wizard-of-Oz studies

Table 3: Overview of evaluation status and contingency plans

Table 4: Conditions in the experiments of the summative evaluation

Table 5: Implementation risks during summative trials and contingency plans

List of Abbreviations

AM Acoustic model

ELE Exploratory Learning Environment

LM Language model

M Month

VPS Vygotsky Policy Sequencer

WER Word error rate

WoZ Wizard of Oz

Y Year



D5.2 Report on formative evaluation results in Y2

1. General introduction

The iTalk2Learn project aims at facilitating robust learning in elementary education. Robust learning includes the acquisition of procedural skills and of conceptual knowledge (Koedinger, Corbett, & Perfetti, 2012). Definitions of procedural and conceptual knowledge, as well as discussion of the interaction between them, are included in D1.1 and in D1.3. The iTalk2Learn project aims to facilitate robust learning in elementary education by creating a platform for intelligent support that combines existing structured tasks with new exploratory tasks, and that provides possibilities for voice interaction. The platform includes a sequencer for structured tasks and implements an intervention model (specified in D1.3) for switching between structured tasks and exploratory tasks. Furthermore, it provides task-dependent support and task-independent support to learners while they are working on specific tasks. The automatic adaptivity is facilitated by a speech recognition system that also enables learners to communicate more naturally with the interface and to reflect on their own thinking. For children, such a system is not yet available, so another strand of the project is the development of a recognition system for children's speech.

Work package 5 (data collection and evaluation) has two main objectives: 1) formative evaluation and 2) summative evaluation. The progress of the consortium's work on the various components of the project and their usability are evaluated by using formative evaluation strategies. The formative evaluation plan (see D5.1) described the envisioned iterative process of developing, implementing, and testing the various components of the iTalk2Learn platform. The results of the formative evaluation inform the design of the summative evaluation. The summative evaluation aims to assess the efficacy of the iTalk2Learn platform to reach its goals of providing an intelligent tutoring system for robust fractions learning with speech support. Two experiments will be conducted in two proven application scenarios and in two European languages (English and German). We will assess students' learning, motivation/engagement and their satisfaction with the system with pre- and post-tests, questionnaires, and interviews. At the end of Y2, we have largely finished the formative evaluation of the exploratory tasks, of speech recognition for young learners, and of automatic adaptivity concerning sequencing and support. While deliverable 5.1 reported on the plans for the formative evaluation, the current deliverable focuses on the results of the formative evaluation that took place in Y2 and their consequences for the summative evaluation.

Figure 1 gives an overview of the architecture of the iTalk2Learn platform and indicates in which section of this deliverable the formative evaluation of the various components are reported. We took the decision not to evaluate the structured tasks within this project because they have previously undergone extensive design and trialling in separate projects and are proven within their respective fields. However, analysis of the structured tasks to identify their suitability to the project and age-appropriateness was carried out during school trials. This is discussed further in D1.2.



D5.2 Report on formative evaluation results in Y2

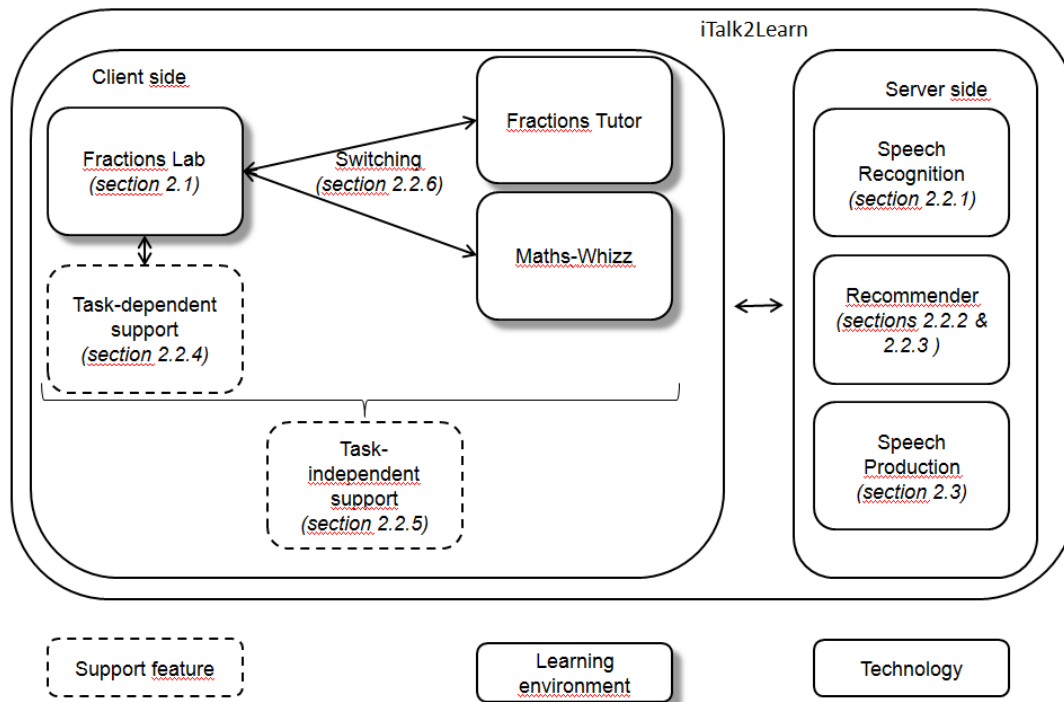


Figure 1: Architecture of the iTalk2Learn platform

1.1 Overview of evaluation plan

The goals of the formative evaluation are to optimally inform the project about the iterative design of the various components of the iTalk2Learn platform in order to improve the design and to advance theoretical principles. We worked on Fractions Lab, speech recognition, and automatic adaptivity (sequencing for structured tasks, task-dependent support for Fractions Lab and task-independent support for the platform, switching between Fractions Lab and Maths-Whizz or Fractions Tutor, respectively) in parallel threads in order not to slow down the overall progress of the project. Paralleling methodologies of educational design research (e.g. Gravemeijer & Cobb, 2006, cf. D5.1.), our evaluation plan includes three phases: preparation for design experimentation, conducting design experimentation, and conducting summative evaluation.

Phase 1: Preparation for design experimentation

The first phase of the evaluation mainly took place in Y1. This phase included literature reviews, analyses of the state-of-the-art, and walk-throughs of preliminary versions of the to-be-developed components of the iTalk2Learn platform with pilot participants. The purpose of phase 1 was to produce a “conjectured local instruction theory” (Gravemeijer & Cobb, 2006, p. 19) that describes how robust fractions learning can be achieved with the iTalk2Learn platform. This local instruction



D5.2 Report on formative evaluation results in Y2

theory has been refined through the second phase of the design experimentation approach in Y2. The steps of phase 1 are reported in detail in D5.1.

Phase 2: Conducting design experimentation

During phase 2, the effectiveness and usability of the developed components of the educational system were tested in iterative trials. The results informed the next developmental stages of the components. For this purpose the components were tested in several iterations of test cycles. The trials that were conducted in this period took place with the same population as the later system users, in schools in Germany and the UK. For most components of the iTalk2Learn project, the second phase is already completed. The description of phase 2 is the focus of this deliverable. We describe the evaluation strategies, the evaluation status, and results for the relevant iTalk2Learn components, that is, exploratory tasks, speech recognition, and automatic adaptivity.

Phase 3: Conducting summative evaluation

The aim of the third phase is to provide “resulting claims that are trustworthy” (Cobb, Confrey, diSessa, Lehrer & Schauble, 2003, p. 13) regarding the efficacy of the iTalk2Learn platform to reach its goals of providing an intelligent tutoring system for robust fractions learning with speech support. In the iTalk2Learn project the third phase will take place in Y3 (see section 3). In the planned summative evaluation we aim to establish these “trustworthy” claims. The results of this third phase will be reflected in the various Y3 deliverables, particularly in the Report on Summative Evaluation (D5.3). In the current deliverable, we assess the readiness of the components for the summative evaluation and provide contingency plans when we identify a corresponding risk. Based on these results, we then describe the updated summative evaluation plan and its risks.

2. Results of formative evaluation

2.1 Fractions Lab and Exploratory Tasks

As mentioned above and in earlier deliverables (e.g., D1.1, D5.1), the iTalk2Learn project aims at fostering robust mathematics learning, which consists of procedural skills and conceptual knowledge. D1.1 discussed that the acquisition of conceptual knowledge can be facilitated by engaging students in exploratory tasks. For these means we developed an exploratory learning environment called Fractions Lab (see D3.2 for design and development of Fractions Lab). In Fractions Lab students work on exploratory tasks, which were also designed as part of the project. The tasks were revised during studies and these revisions are reported upon in D1.2.

Phase 1: preparation for design experimentation

The development and evaluation of Fractions Lab in phase 1 brought together three sources: the



D5.2 Report on formative evaluation results in Y2

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 as well as the use of experts in mathematics to act as critical friends. The preparation for the design experimentation was outlined in D5.1 and was completed in Y1. Y2 has seen the focus on phase 2 involving iterative interventions and design cycles to evaluate a) Fraction Lab's user interface and b) the impact of Fractions Lab and associated tasks on students' conceptual understanding of fractions.

To conduct the trials with the German version of Fractions Lab, it needed to be translated and culturally adapted to the German student's needs. This was done after the main developments in English were completed (after M17).

Phase 2: conducting design experimentation

During Y2 trials to evaluate Fractions Lab and the associated tasks have taken place in the UK in one school with 76 9-11 year old (Year 5 and 6) students and in Germany in two schools with 13 10-12 year old (Year 6) students.

In the UK we worked particularly closely with the fifth grade students (total 37 students) on three occasions over the year (see

Table 1). The students saw themselves as partners in the design and development of Fractions Lab and were keen participants.

Table 1 also includes the trials conducted with sixth grade students in the UK and in Germany.



D5.2 Report on formative evaluation results in Y2

Table 1: Fractions Lab evaluation studies

Month	No. of days	Grade	No. of students ¹	Setting	Objectives / Partners involved	Language
15	2	5	22	1-1 setting	Task trials (IOE with RUB visiting the UK school together)	English
17	3	5	23	Authentic class setting	Task trials, user interface, impact of FL on cognitive development (IOE and BBK)	English
			36	Authentic class setting	Trial of post-test questions (IOE and RUB)	English
			12	Focus group	User interface (IOE and BBK)	English
			4	Small group interview	Fraction representations (IOE)	English
20	2	6	39	Authentic class setting	Task trials, user interface, impact of FL on cognitive development (IOE and BBK)	English
20	5	6	13	1-1 setting	Task trials, user interface, impact of wizard-delivered task-dependent and task-independent support on student behaviour.	German
21	2	5	19	Authentic class setting	Task trials, impact of FL on cognitive development (IOE and BBK)	English

Furthermore, in the UK a series of four hands-on evaluation workshops for teachers were held, attracting 23 teachers who are currently participating in Masters'-level study of primary mathematics education. We selected these teachers because of their specialism in mathematics, their role as practitioners in a variety of schools and range of primary phase classrooms, and their ability to critically evaluate Fractions Lab and its role in the classroom. Further details regarding the method are provided in Hansen, Mavrikis and Geraniou (2014).

On every study with students, the data we gathered included screen recordings, voice recordings, video recordings, written data from student worksheets completed prior to, during or after working with Fractions Lab. This included the piloting of test items to serve as a pre- and post-test that capture procedural and conceptual knowledge for the summative evaluation of the iTalk2Learn platform. The teachers completed questionnaires prior to, during and after the

¹ We worked with some 5th grade students on more than one occasion and others just once during our visits. Total number of 5th grade students who collaborated = 37.



D5.2 Report on formative evaluation results in Y2

workshops and voice recordings were made. We discuss below the impact of our findings from these studies.

User interface

Fractions Lab was iteratively designed and tested by TL and IOE in collaboration with the other project partners. It was designed within a given size as it needs to slot into the iTalk2Learn platform as a component. Furthermore, the entire user interface design has been created with an eye on the possible exploitation on mobile devices (tablets in particular). Hence, every choice in terms of usability (size of buttons, behaviour of menus etc.) has been done by taking that requirement into account. For further detail about Fraction Lab's design and the user interface see D3.4.1.

Overall the user interface has been very well received by students and teachers, both in Germany and the UK (see also section 2.3). They found Fractions Lab's colour scheme and overall design attractive and the layout intuitive. The earlier student experiments and the teacher workshops provided helpful data about enhancing Fractions Lab's user interface. As expected, students typically found Fractions Lab more intuitive to use than the teachers. In brief, teachers are often challenged by educational technology but the workshops and usage of the platform overall and Fractions Lab in particular in the classroom has helped us understand better their requirements and professional development needs in order to use the system in their classrooms as well as recommend to colleagues, parents and students.

Studies earlier in Y2 with students and teachers led to the following indicative enhancements to Fractions Lab:

- Change Equivalence Box to Comparison Box, introducing < and >
- Reduction in the options for how Fractions Lab can add and subtract fractions
- Change in the way the addition and subtraction menus are used to reduce number of mouse clicks
- Double clicking on parts of fractions to 'use' them
- Animations demonstrating addition and subtraction were slowed down to show the underpinning concepts more clearly

For further detail see Study Reports in Appendix 1 and Appendix 2.

The impact of Fractions Lab and associated tasks on students' conceptual understanding of fractions

The most significant part of our evaluation has been related to students' conceptual understanding



D5.2 Report on formative evaluation results in Y2

of fractions because it is a core outcome of the iTalk2Learn platform. D1.1 defines conceptual understanding broadly. In relation to fractions within the iTalk2Learn project, we define it as implicit or explicit understanding about fraction representations, fraction interpretations, fraction types, task types and the fine-grain goals related to those tasks (see D1.2). Our representation of a coherent system of fractions shows the underlying principles and structures of fractions and their interrelated nature, the fraction representations within Fractions Lab, and the core pedagogical considerations. The focus of this type of knowledge lies on understanding why, for example, different mathematical principles refer to each other and on making sense of these connections. Conceptual understanding of equivalent fractions, for example, includes students being able to make connections between fraction representations by understanding what is the same and different within them (Lesh et al., 1983) and show that a fraction represents a number with many names (Wong & Evans, 2007).

The data are rich and at the time of writing continue to be analysed. However, we have interim analysis of the data that informed the BERA paper, "Designing interactive representations for learning fractions" (Hansen, Geraniou & Mavrikis, 2014). The Study Report in Appendix 2 provides more detail, but key findings include the following:

- Although students had a range of representations to draw upon when showing $\frac{1}{4}$ before their Fractions Lab experience, none used number lines or liquid measures. After Fractions Lab, a significant number reported broadening their use and understanding of a range of representations, including number lines and liquid measures.
- The students were asked which representation they preferred to use in Fractions Lab. 35% preferred liquid measures because they found them useful and clearer to use. We found these results very surprising (and promising) after a limited time using Fractions Lab. Our findings add weight to Silver (1983) supposition that representations beyond the area model may help students' fractions understanding by enabling flexibility and the finding of Stein, Smith, Henningsen & Silver (2000) that students can use virtual manipulatives more flexibly.
- When asking the students how the different representations in Fractions Lab supported their fractions learning, it appears that representations may be more effective at supporting students' conceptual understanding of some aspects over others.
- The students self-reported learning more about fraction representations, how partitioning can be used to find and show equivalent fractions, addition and subtraction, finding common denominators and the size of fractions. This was pleasing as it resonates with the design decisions we made.



D5.2 Report on formative evaluation results in Y2

Conclusions

From our findings to date we conclude that students' interaction with Fractions Lab provokes them to think conceptually about fractions using representations that are new as well as familiar to them. They appear to be able to capitalise on their intuition, and sometimes challenge it, discouraging them from simply procedurally calculating an answer. We have data to demonstrate that some students may benefit from being introduced to a wider range of representations than they are currently exposed to, and that for some students the number line and liquid measures appear to support their fractions knowledge more than the typical area/region representation. Developing virtual manipulatives that enable students to witness what happens dynamically as they create a fraction, partition it to find an equivalent or add/subtract two fractions appears to have the potential to enhance their conceptual understanding of fractions.

Our next step in the analysis is to include the video and voice data we have in order to triangulate the findings. We would like to undertake further work on how the representations, particularly liquid measures, support students' fractions understanding.

2.2 Automatic adaptivity

One major aim of the iTalk2Learn project is the development of automatic adaptivity to provide students with individual sequences of tasks and support that fits their progress and needs. Speech recognition offers an intriguing way to provide automatic adaptivity especially for young children. Speech recognition is therefore built into all three threads of automatic adaptivity developed for the iTalk2Learn platform (although all threads also work without indicators from speech; see section 2.2.1).

(1) The first thread focuses on performance prediction to improve task selection within the existing tutors for structured practice (Maths-Whizz and Fractions Tutor) that we integrated in our iTalk2Learn platform. To improve the existing tutors, we used recommender technology to develop an adaptive task sequencer (Vygotsky Policy Sequencer) which selects the order of the tasks based on the students' learning process (see section 2.2.2). In addition, the sequencing and performance prediction will be ameliorated through affect recognition (see section 2.2.3)

(2) The second thread aims at providing task-dependent and task-independent support to students. With regard to structured practice, the existing tutors Maths-Whizz and Fractions Tutor already provide task-dependent support. For our newly-developed exploratory learning environment, Fraction Lab, we developed task-dependent support. Task-dependent support is given based on students' actions using the iTalk2Learn platform and their performance. For Fractions Lab, this support was tested in Wizard-of-Oz studies (see section 2.2.4). We also developed task-independent support for the exploratory tasks as well as for the structured tasks. This task-independent support responds to students' affect by using speech indicators and taking students



D5.2 Report on formative evaluation results in Y2

utterances as well as other screen and mouse activity into account (see section 2.2.5).

(3) Finally, the project seeks to innovate on how to best switch between structured practice and exploratory activities. The question is when students should switch from structured to exploratory tasks and vice versa. The theoretical intervention model in D1.3 that the switching in the formative evaluation is based on, takes into account both actual and predicted student performance within the structured practice or exploratory activities, as well as other speech indicators. Plans on this last formative evaluation step are reported in section 2.2.6.

2.2.1 Speech recognition - Development of the speech recognition system for children

As stated above, speech recognition is a prerequisite for the development of the automatic adaptivity for the italk2learn platform. In D3.1, we discussed that existing speech recognition developed for adults does not achieve the necessary accuracy on recognizing young children's speech. Therefore SAIL is developing a speech recognition system for young learners. Automatic speech recognition requires acoustic models (AM) and language models (LM) for the target domain and speakers of iTalk2Learn. The baseline for speech recognition is represented by SAIL's standard models trained on adult speaker voices and textual material from the domain of broadcast-news. As iTalk2Learn addresses young learners, the AM built from adult voices cannot be used for transcription purposes. Rather, specific models – reflecting the acoustic qualities of the speakers as well as of the recording environment – had to be created. This process entails in phase 1 the collection of appropriate corpora (acoustic data + transcripts according to specific guidelines) and in phase 2 subsequent training of AM. For a detailed description on the development evaluation of the speech recognition model for children see D3.3.1.

Phase 1: Preparation for design experimentation

In this phase, work on automatic speech recognition focussed on the creation of suitable corpora to train acoustic models (AM) and language models (LM). Phase 1 consisted of collecting speech data of the same population as the later system user, i.e., students in Germany and the UK. The data were collected through several trials in the UK and in Germany by IOE Whizz and RUB. In Germany speech data of 251 students were collected. Of these students 138 (49 hours of speech recordings) worked in a 1-on-1 setting and 113 students (83 hours) in a classroom setting (at schools in an effort to resemble the envisaged setting where iTalk2Learn will be used.). The total amount of speech recording time in Germany was about 132 hours. In the UK the total amount of speech recording time was approximately 58 hours of 178 students were collected. In the UK speech data were collected in all of the conducted evaluation studies reported in this deliverable (for details on the settings and objectives see Table 1 and Table 2). The numbers presented here refer to the duration of the sessions with the students. Obviously, the actual speaking time is lower. Typically, only a certain percentage of these recordings contain actual audio to be used for model training – the current estimate for this lies between 30% and 60% of all data collected (for more details see



D5.2 Report on formative evaluation results in Y2

D3.3.1). The speech corpora in English and German contain speech collected during problem-solving scenarios (with the current state of the platform) 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.). In addition, several corpora targeting similar domains were acquired and can now be used for AM training.

Phase 2: Conducting design experimentation

The AM need to be trained for English and German and are being trained in an iterative manner with increasing amounts of data (and improved performance). The LM, on the other hand, need to reflect the target domain of Math tutoring, in particular the fractions domain. Thus, specific terminology and utterances typically uttered by students when interacting with the system form the basis of this process. In addition, vocabulary and speech patterns reflecting the affective state are taken into account. While the former serve as input to both, task-dependent as well as task-independent support, the latter serves as a basis for the detection of a student's affective state. Following the initial creation of models for English, the word-error-rate (WER) has been determined on a held-out test set. As described in D3.3.1 numbers range above 50%, which is not unusual given the complexity of the task and the amount of training material available at the time of model building. The performance of subsequent models incorporating more training material is expected to improve substantially. However, we also believe that a key aspect to consider when evaluating is the 'cost' of any errors. Measuring the WER has to be extended by measuring of precision and recall of key-terminology as these terms form the basis for upstream processing, e.g., the detection of proper usage of mathematical terms. These measures are expected to yield a better view onto the performance of speech recognition in the context of student tutoring. In addition, based on our experiences with the audio recordings in the formative trials, we have developed guidelines for the audio recording setup in future trials in order to guarantee a sufficient quality of audio recordings for the recognition (see also section 3.1).

Conclusions

Not all of the recorded speech data have yet been included in the language models. Thus, next steps will include training of the acoustic models for English and German including all collected and transcribed data. Regarding vocabulary, the addition of further phrases and words typically uttered by student learners will help to create a more focused vocabulary and the training of the language models will be adjusted to the exact use. Lastly an evaluation of all models using precision and recall on key-terminology will take place. Details on the plan for next steps regarding model training and evaluation are reported in D3.3.1



D5.2 Report on formative evaluation results in Y2

2.2.2 Sequencing of structured tasks

One of the components of automatic adaptivity is performance prediction. This is one source of information for the intervention model that determines which tasks learners receive. To test performance prediction, we have developed an adaptive content sequencer for the structured practice component (Maths-Whizz in the UK and Fractions Tutor in Germany). This so-called Vygotsky Policy Sequencer is a proof of concept of performance prediction and is described in detail in D2.2.1. In this section, we present the formative evaluation results of the newly developed sequencer. We evaluated in our experiment the Vygotsky Policy Sequencer in comparison with the Maths-Whizz sequencer. The Maths-Whizz-sequencer is a rule-based sequencer that was refined for years by experts following the national curriculum. The purposes of the trial with the students were twofold: Firstly, we wanted to show that it is possible to sequence tasks just considering students' score on previous tasks. Secondly, we wanted to evaluate sequencer performances in comparison with the current Maths-Whizz sequencer.

Phase 1: Preparation for design experimentation

The Vygotsky Policy Sequencer is composed of a performance predictor and a score-based policy inspired by the concept of Zone of Proximal Development. In order to develop it and to evaluate it UHi designed and implemented a simulated environment, as described in D2.2.1 and by Schatten, & Schmidt-Thieme, (2014), to perform simulated online experiments as proof of concept.

In order to perform the online evaluation UHi in collaboration with Whizz needed to integrate the working prototype of the Vygotsky Policy Sequencer into the Maths-Whizz platform. The first integration step consisted of a feasibility study of integrating the Vygotsky Policy Sequencer into Maths-Whizz and of using the dataset generated by Maths-Whizz for performance prediction. This feasibility study is described in D2.2.1 and led to the Schatten, Janning, Mavrikis, and Schmidt-Thieme (2014) publication. Thanks to the study, it was possible to create a first performance prediction model based on Maths-Whizz data. In the second integration step the Vygotsky Policy Sequencer was modified to achieve real-time performance. This involved the implementation of an online update for Matrix Factorization algorithms adapted from (Rendle & Schmidt-Thieme, 2008) and of a lightweight interface with the Maths-Whizz platform (Schatten, Witsuba, Schmidt-Thieme & Gutierrez-Santos, 2014).

The prototype of the Vygotsky Policy Sequencer can be used also for sequencing tasks in the German structured practice component, the Fractions Tutor. To train the sequencer, we needed to: collect new data because (1) available historic data of students working with the English Fractions Tutor can only be used in a very limited way because Fractions Tutor was translated and adapted to the German students' needs. These modifications may influence the fit between a possible model developed with historic data and data collected with the modified Fractions Tutor (2) the Fractions Tutor (unlike Maths-Whizz) does not present the tasks in many different orders. As discussed in



D5.2 Report on formative evaluation results in Y2

D2.1, availability of data from different orders of the tasks is mandatory to develop a sequencer model based on Reinforcement Learning technology.

Phase 2: Conducting design experimentation

Schools agreed to let 8-9 years old children interact with Math-Whizz during school hours. The 98 students who took part on the study, were also able to practice at home and use all other related features of Math-Whizz, e.g., spend coins gained for passing tasks for decorating their virtual room. The system randomly assigned the students to two groups - one practicing with the Maths-Whizz sequencers and one with the Vygotsky Policy Sequencer. In order to answer our research questions UHi, IOE and Whizz chose and analysed the following success indicators: learning gains based on post-test scores and log-file data, user experience based on a five-item questionnaire, and integration performance.

The log file data showed that both groups worked on approximately 2000 tasks. To test the accuracy of the sequencers in performance prediction, we compared log file data from the sequencers with post-test performance. This log file data includes information on how well students performed tasks and what their predicted performance is. In other words, we checked whether the sequencers could predict performance on the post-test. While the Vygotsky Policy Sequencer assessment and post-test performance were almost equal, the performance of the Maths-Whizz group was underestimated by the Maths-Whizz sequencer. The Vygotsky Policy Sequencer thus has better user modelling and, over time, should be better in adapting to the knowledge acquisition rate of the students.

Students reported a better user experience when working with the Vygotsky Policy Sequencer version: Maths-Whizz was more fun, less repetitive, easier to understand, and exercises were easier compared to working with the existing sequencer. Both versions of Maths-Whizz were seen as equally helpful.

Students working with our new Vygotsky Policy Sequencer performed just as well as students using the existing Maths-Whizz sequencer on the post-test assessments. Demonstrating differences in performance of two sequencers is known to be difficult since it requires a large number of students working with the sequencers over extensive time periods. As such, after discussions also with our advisory board and taking into account the exploitation plans of the project, we consider the trial of the Vygotsky Policy Sequencer to be successful: it promotes learning equally well as the existing Maths-Whizz sequencer while providing significant advantages over the existing sequencer that will become visible over time:

- Simple integration in not ad-hoc constructed systems.
- Having comparable response time as the existing rule-based system with 30 students



D5.2 Report on formative evaluation results in Y2

interacting at the same time.

- Achieving the same post-test results with almost no curriculum authoring effort.
- Possessing more accurate user modelling

To collect the required log file data for training the Vygotsky Policy Sequencer for Fractions Tutor, a study with 113 students in Germany (age 10-12/grade 6) was conducted by RUB with local support by UHi. The trial was able to demonstrate technical feasibility of implementing the iTalk2Learn platform in a local network. Particularly in schools where internet bandwidth is limited, this has proven to be a promising approach to handle server tasks. Additionally the trial served the purpose of speech data collection reported in section 2.2.1. To produce the required variations in task sequencer, the students were randomly assigned to one of three groups: interleaved, blocked, and mixed sequences of task types. For a detailed description of the Fractions Tutor and the task-types see D1.2. The children were also asked to give feedback on their experience with the platform (for results, see D7.3.1). The collected German log data are currently being analysed in order to train the Vygotsky Policy Sequencer.

Conclusions

The experiment results are promising considering the coherence between the different selected success indicators. Because the Vygotsky Policy Sequencer possesses better user modelling, we believe that a longer period of time would have shown differences also in the post-test. Consequently, we are looking forward to a bigger experiment for the summative evaluation.

2.2.3 Amelioration of performance prediction and sequencing through affect recognition

As described in deliverable D3.4.1, we aim at ameliorating the performance prediction and task sequencing for the iTalk2Learn platform by interpreting the behaviour of students interacting with the system. More explicitly, we aim at recognising the emotions and affect of the students by extracting appropriate features from students' speech input and distinguishing between different affects by means of machine learning methods. By recognizing, for instance, when students are under- or over-challenged with tasks, the sequencer can better select appropriate next tasks for students.

Phase 1: Preparation for design experimentation

To develop a machine learning model for affect detection, we first had to collect a dataset with speech input from students solving appropriate tasks and manually assign labels for the appropriate affect to that data. This was a labour-intensive process that was achieved in close collaboration of RUB and UHi. Subsequently, UHi and SAIL developed and analysed appropriate features from the speech data that can be used to classify the behaviour of students by means of machine learning models. Two different kinds of features are used: linguistic features and acoustic



D5.2 Report on formative evaluation results in Y2

features (i.e., disfluencies like pauses and fillers). The feature development and analysis are described in detail in deliverable D3.4.1. The next step was to select and train an appropriate machine learning classification model to be able to automatically map the features, extracted from the collected data, to the manually assigned labels. The feature development and analysis as well as a first training of a state-of-the-art classification model took place in phase 2 and are hence summarized in the next section (see also D3.4.1 for more details).

Phase 2: Conducting design experimentation

The evaluation of the proposed data features showed that these features are able to describe student affects. This feature analysis was done by statistical methods, more explicitly by mapping feature values to the affect labels and applying a (multivariate) linear regression for different single features and feature combinations. The regression showed that there are feature combinations that are able to describe students affects (based on p-values and R^2 -value; for more details see D3.4.1).

We chose a support vector machine as a state-of-the art machine-learning classification model for affect recognition according to the literature and trained it with the collected dataset. As usual in machine learning the evaluation of the classification with the trained model was done using a k-fold cross validation. For a k-fold cross validation, the data is split into k sets and one conducts k experiments. In each experiment one of the k sets is the data for the test of the model and the other (k-1) sets are used for training the model. The classification performance of the model, or more precisely the classification test error, is expressed by the ratio of wrongly classified examples to all examples. However, the classification performance of the first trained model was not yet satisfactory enough, hence as a next step UHi plans to create a new model for improving the classification performance.

The above mentioned results of the feature analysis and of the training of the preliminary affect recognition model are reported in detail in deliverable D3.4.1 and are published in the proceedings of the EDM 2014 conference and of the EC-TEL 2014 conference (Janning, Schatten & Schmidt-Thieme., 2014a; Janning, Schatten & Schmidt-Thieme et al., 2014b). The further development of an improved model for affect recognition will be reported in D3.4.2 and submitted to further conferences and journals.

Conclusions

We were able to show by a feature analysis that the developed features can be used for affect recognition. However, the classification performance with a state-of-the-art machine learning model applied to those features extracted from a real dataset is not yet satisfactory enough. To increase the affect recognition performance, we will investigate how to expand or change the currently used, state-of-the-art classification model. After developing an appropriate ameliorated model, we have to conduct a further experiment for measuring the classification performance and comparing it to the former measured classification performance to show that the new model



D5.2 Report on formative evaluation results in Y2

performs better than the old one.

2.2.4 Task-dependent support for Fractions Lab

The design-based formative evaluation of Fractions Lab (discussed in section 2.1) also served as an opportunity to conduct Wizard-of-Oz (WoZ) studies to inform the further development and refinement of the task-dependent support for Fractions Lab. In WoZ studies, humans (the wizards) simulate the adaptivity. Actions of the students on the computer are relayed to the wizards. The wizards then respond to these actions following a detailed script that also forms the basis of the adaptive computer component. All wizard action is again relayed through the computer system so that students do not notice that a human is providing the adaptive support, not the computer. At this developmental stage of Fractions Lab the tasks as well as the wizard-delivered task-dependent and task-independent support were read out loud by the system.

Regarding structured tasks, the tutors Maths-Whizz and Fractions Tutor already provide hints depending on the student's progress (for description of Maths-Whizz and Fractions Tutor see D1.1). We left the hint functionalities of the existing tutors intact. For the exploratory tasks, IOE and BBK in collaboration with RUB developed 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). As discussed in detail in D1.3, the aim of the task-dependent support is to provide personalized feedback during interaction with Fractions Lab, to help the student to deal with and learn from errors that they make while responding to the tasks. We developed three types of feedback: 'instruction', 'problem solving', and 'reflective' feedback.

Phase 1: Preparation for design experimentation

The development of adaptive task-dependent support is highly interlinked with the development of the exploratory tasks described in section 2.1. As we did for the other 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 facilitates the design of the task-dependent support to be provided in Fractions Lab.

Fractions Lab tasks and task-dependent and task-independent support were developed and tested first in English. We then conducted multiple iterations of formative evaluations in the UK (reported below) before adapting Fractions Lab and its support functionalities to German students' needs. The first evaluation study in Germany was conducted by the end of M20. First results are provided for children's perception of Fractions Lab (see section 2.1 and 2.3) and support functionalities (see 2.3). Further data are currently still being analysed. The results will be reported in D5.3.



D5.2 Report on formative evaluation results in Y2

Phase 2: conducting design experimentation

In parallel to the formative evaluation of Fractions Lab, we conducted WoZ studies on task-dependent support in a classroom equipped with computers. In total 51 students took part in the WoZ studies conducted in March and July (see Table 2). Support was provided by the wizards using a script to decide when and what type of feedback ('instruction', 'problem solving', or 'reflective') should be provided to the students, based on their interaction with the learning environment, their performance and what they said. For details of this WoZ methodology see Study Reports in Appendix 3 and the corresponding conference paper (Mavrikis, Grawemeyer, Hansen & Gutierrez-Santos, 2014). In a second step, we are implementing automatic delivery of the support. So far, this implementation has been done for one particular task. Our studies show that there was no difference between the wizard-delivered support and the automatically delivered support. This was true for: what support was given, how students reacted to and how they perceived the support given.



D5.2 Report on formative evaluation results in Y2

Table 2: Wizard-of-Oz studies (run in parallel with the Fractions Lab studies outlined in 2.1)

Month	No. of days	Grade	No. of students	Setting	Objectives	Partners involved	Language
16	3	5	12	Authentic class setting	Impact of wizard-delivered task-dependent and task-independent support on student behaviour.	IOE and BBK	English
19	2	6	5	Authentic class setting	Impact of wizard-delivered task-dependent and task-independent support on student behaviour.	IOE and BBK	English
20	5	6	13	One-to-one setting	Task trials, user interface, Impact of wizard-delivered task-dependent and task-independent support on student behaviour.	RUB	German
20	2	5	21	Authentic class setting	Test the implementation automatically Impact of wizard-delivered and automatically delivered task-dependent and task-independent support on student behaviour.	IOE and BBK	English

A key outcome of all WoZ studies was the value of task-dependent support to a student's progress through Fractions Lab. Students were frequently observed to modify their behaviour, almost always in a positive way, and usually to successfully continue the task. Based on these outcomes, the WoZ script was further systematised.

We also investigated whether there was an effect of the modality of the task-dependent support feedback (whether the presentation was 'high interruptive' or 'low interruptive') on a student's affective state. As described in the Study Report (Appendix 5), the results show that, for example, when students are frustrated, high interruptive feedback is more effective than low interruptive



D5.2 Report on formative evaluation results in Y2

feedback. Accordingly, the iTALK2Learn system was amended to tailor the modality of the task-dependent support feedback to the student's affective state, with the aim of further enhancing the learning experience.

Conclusions

In our test and design cycles we developed and evaluated task-dependent support for Fractions Lab. We showed that this support fosters the problem-solving process and enhances students' perception of Fractions Lab. Furthermore, we showed for one example task that the task-dependent support can be delivered automatically by the iTALK2Learn platform, thus not requiring a human tutor. The automatic support was as effective as the human/wizard provided support with regard to both the problem-solving process and students' perception of Fractions Lab. As a next step, the task-dependent support will be developed further for additional tasks.

2.2.5 Task-independent support based on speech indicators

The aim of the task-independent support is to enable natural interaction with the platform through speech. Task-independent feedback is provided for both structured and unstructured tasks according to speech: (1) mathematics vocabulary and (2) affective states. It builds on advanced behavioural interaction interpretation (i.e., speech detection) and the direct interaction with the system especially within Fractions Lab (e.g. mouse interaction with the Fractions Lab objects).

Phase 1: Preparation for design experimentation

In order to derive speech indicators we needed to record what students utter during interaction with the system. For this purpose we conducted several studies in the UK and in Germany (see section 2.2.1). From the collected speech recordings, vocabulary lists were compiled for a) possible utterances regarding perceived task difficulty and related affect such as boredom or frustration, and b) relevant mathematics terminology. The lists are integrated in the LM and AM of the speech recognition system at the time of this writing. The speech recognition system is then going to be trained to detect the words from these lists in children's utterances. In this regard, we had 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. Another aim of these studies was to find out how students react to the task-independent support, particularly in reliance on the students affective state.

Phase 2: Conducting design experimentation

The WoZ studies described above also served the purpose to inform the design of the task-independent support. We were particularly interested in the following questions: (1) Will students be able to use mathematics vocabulary if prompted to do so? (2) Is there an effect of different affective state types upon reaction towards feedback? (3) Which feedback types were most



D5.2 Report on formative evaluation results in Y2

successful given a particular affective state?

In order to address these questions the WoZ studies investigated the use of mathematics vocabulary and the effect of affective states on different feedback types at different stages of the task. In order to investigate whether task-independent support can amend the support of traditional problem-solving feedback, task-dependent support was also included here.

The following different feedback types were provided to students: AFFECT - affect boosts, TALK ALOUD - talking aloud, TALK MATHEMATICS - using particular domain specific mathematics vocabulary, PROBLEM SOLVING – problem solving feedback, REFLECTION – reflective prompts, and OTHER – non-learning specific support. We analysed the affective states that occurred while the different types of feedback were given and whether the students reacted to the feedback.

Overall, students reacted well to requests to talk aloud, reflect, and talk mathematics. In particular, when students were in a negative affective state, such as frustration or confusion, those speech related requests were more effective than, for example, problem solving support (for details on the analyses see Study Report in appendix 3) and high interruptive feedback was more effective than low interruptive feedback (please see appendix 5).

Conclusions

Our results indicate that traditional problem-solving feedback is only able to support students to some extent. When confused, students may have found the problem-solving feedback too interruptive, as it might have suggested switching to a new strategy for answering the task. Additionally, when frustrated student's motivation might be low and also there might be increased cognitive load. Providing problem-solving feedback when students are frustrated does not seem to be a very effective strategy.

In contrast, asking students to talk aloud when confused or frustrated might help them to express their problems, which might move them out of their negative affective state. Additionally, the results imply that reflecting on one's own strategy of solving a task is motivating, even when confused or frustrated. We noticed that it may also have helped students to identify misconceptions or may have led them to new ideas about how to solve the learning task. Reminding students to use specific mathematics vocabulary might help them to think through the problem and resolve their confusion.

2.2.6 Switching between exploratory tasks and structured tasks

A final formative evaluation step informs the design of the intervention model that is used for *switching* between exploratory tasks in Fractions Lab, and structured tasks in Maths-Whizz in the UK, or Fractions Tutor in Germany, respectively. To promote robust fractions learning, iTALK2Learn switches between these two task types but also sequences tasks within each task type.



D5.2 Report on formative evaluation results in Y2

As discussed in D1.3, the aim of this strand of work is to ensure the student engages with learning activities (exploratory tasks and/or structured tasks) that are most appropriate for their current affective and cognitive state – specifically, their current conceptual and procedural knowledge of fractions. In this context, the notion of ‘most appropriate’ should take into account (1) the individual student’s achievements in the system so far, (2) their affective state, (3) what might be most pedagogically appropriate.

For this reason, sequencing and switching represents a touch point for the project’s multiple strands – specifically: student response modelling and performance prediction (BBK, UHi), affect recognition based on speech and actions in the system (BBK, SAIL, UHI), and pedagogy (IOE, RUB). Prerequisites, therefore, for evaluating this strand of work were the formative evaluation of the newly-developed components of iTalk2Learn, namely Fractions Lab, the Vygotsky Policy sequencer, speech recognition, task-independent and task-dependent support, as well as the further refinement of the intervention model that prescribes strategies for switching (see D1.3).

Phase 1: Preparation for design experimentation

The preparation for conducting formative evaluation trials started with the formative evaluations conducted over the summer. The evaluated first versions of the separate components have been integrated into the iTalk2Learn platform. This integration is described in detail in D4.1 and D4.2.1. This prepared the ground for the technical implementation of the switching between the Fractions Lab and the structured practice environment. Concurrently, the intervention model was being iteratively revised as part of D1.3 and in parallel, the design-based formative evaluation of Fractions Lab (discussed in Section 2.1) as well as the sequencing studies (see section 2.2.2) acted as an opportunity to also inform our understanding with respect to switching and sequencing and feed back to D1.3.

Phase 2: Conducting design experimentation

The iterative test and design cycles of the formative evaluations already described in section 2.2.4 & 2.2.5 were also used to develop our approach to sequencing and switching as it is currently being implemented in the system (see the intervention model described in D1.3).

An important insight resulting from already conducted WoZ trials was which factors determined whether students were under challenged’, ‘appropriately challenged’, or ‘over challenged’. We found that the following factors were important in deciding a student’s level of challenge:

1. the type of task, and fine- and coarse-grain goals it is designed to address;
2. the student’s response to the task;
3. the student’s affective state;
4. the amount and type of feedback delivered to the student by task-dependent support.



D5.2 Report on formative evaluation results in Y2

Following this, the selection of the subsequent task also relied on the perceived potential of a task (on behalf of the humans taking part in the WOZ study) in helping the student to challenge a particular misconception or consolidate their newly acquired concepts.

We have formalized these empirical findings in D1.3 as a 'student needs analysis' that, based on these factors, will determine how tasks are assigned to students and thus influence the strategies for sequencing and switching.

Regarding sequencing, the iteratively designed intervention model, simply put, suggests if the student appears 'under challenged', the system will provide a Fractions Lab task that the student is likely to find more challenging; and if they are 'over challenged', the system should sequence to a less challenging Fractions Lab task. In both outcomes, the aim is to keep the student in their 'zone of proximal development' (see D1.3 for more details)

Regarding switching, if the student seems 'appropriately challenged' by an exploratory task, the system should switch to a structured task with the aim of giving them an opportunity to consolidate what they have explored and learned in the Fractions Lab, by means of structured practice. Our empirical findings, however, also suggested that the decision to switch from the structured tasks to exploratory tasks can also depend on other, more pragmatic, factors. For example, the structured tasks take typically much less time to complete than the Fractions Lab tasks, which suggests that both time and number of tasks completed should also inform the decision whether or not to switch.

This strategy for switching between exploratory and structured tasks prioritises conceptual learning over procedural learning. As evident from the formative evaluations (and complemented by our theoretical understanding from early versions of D1.3), there was little point in the student undertaking practice if they don't understand the concepts they are being asked to practise. Instead, they may need the opportunity to engage with a less-challenging fractions task that relates to the concept. Similarly, if the student was previously under challenged, time spent practising those tasks would not be an efficient use of the student's available time and energies.

The next step in Y3 will be to test the efficacy of this switching strategy in the unified platform where the now evaluated components are being integrated. Two formative trials, one in UK and one in Germany, will be conducted with the iTalk2Learn platform integrating the now evaluated components. They will again take place in a WoZ-setting (for details on the methodology see Appendix 3) and will be conducted with UK and German fifth and sixth graders.

Conclusions

Based on the formative evaluation of the newly-developed components of the iTalk2Learn platform and the refinement of the intervention model, switching between Fractions Lab and the structured practice environment can now be evaluated in more detail. Since this will be the first evaluation of



D5.2 Report on formative evaluation results in Y2

the integrated platform, this is an important step on the way to the summative evaluation trials.

The data produced by the upcoming switching trials will continue informing the development of a rule-based system that is based on the intervention model and takes into account the outcome of the other evaluated components (performance prediction, speech recognition and affect detection).

2.3 Overall student perception and feedback

In every field trial we undertook in both UK and Germany, students have generally been enthusiastic about all aspects of iTalk2Learn (with some caveats of course and constructive comments that have helped us improve both the content and the overall student experience). In particular, in both countries students themselves reported that being encouraged to talk by the system helps their thinking and some appreciate that it can help them reflect on their learning. As such we recognize that the potential impact of iTalk2Learn is also beyond the direct claims it can make on learning gains (cf. Oliver, 2011).

Therefore, for the sequence studies reported in Section 2.2.2, the evaluation studies of Fractions Lab (section 2.1), and the WoZ studies (section 2.2.4 and 2.2.5) undertaken in the UK and in Germany, we designed child-friendly questionnaires. Based on positive success in the literature (Read et al. 2008) and our past experience, we decided to employ the visual analogue scale from the Fun Toolkit, albeit not to measure the construct of fun but to gauge students' perception on various constructs related to their interaction with the system (see 'Student Experience' questionnaires in Appendix 7).

The overall feedback from the students is already very positive. Students felt very good after working with Fractions Lab and found Fractions Lab very helpful. The students found the hints mostly supportive and they liked that the support as well as the tasks were read out loud by the system. Analysis showed that there were significant correlations between question 7 ("How much did the feedback get in your way?") and questions 1 ("Now that you have finished the session, how do you feel?"), 3 ("How helpful was Fractions Lab?") and 6 ("Was the feedback helpful?"). In other words, a key suggestion of the formative assessment is the hypothesis that if students are to find the system and its feedback helpful, and are to feel positive following their experience, the feedback must be appropriate and not interruptive.

In addition to surveys we also interviewed student focus groups after their WoZ experience to gain richer data about their impressions of the platform. The students were keen to inform future iterations and as such saw themselves as participants in the design process. They gave us helpful comments related to the speech production (e.g. suggesting that one voice was sarcastic and recommended it was changed, which we subsequently did) and how feedback should be offered (e.g. a pop-up message was deemed to be a "surprise" or a "shock"). They also offered suggestions related to the content and timing of feedback that we were able to implement in later iterations.



D5.2 Report on formative evaluation results in Y2

Conclusion

We plan to continue using these perception metrics as they provide useful insights into where to focus attention and complement our qualitative and quantitative research on the effectiveness iTALK2Learn overall.

3. Summative Evaluation

Following the design trials that are reported in this deliverable, the next phase of the evaluation in Y3 will focus on the summative evaluation of the developed components. In this phase, the parallel developments of the project (i.e., exploratory tasks, speech recognition, and automatic adaptivity) will be evaluated together. All components are combined in a unifying platform (see D.4.1 and D4.2.1 for more information) and they will interplay as described in the intervention model (see D1.3). We will now briefly reiterate the structure of the iTALK2Learn platform to ease understanding of the following discussion of the summative evaluation plan. The platform allows integrating existing tutors for structured practise (i.e., Maths-Whizz or Fractions Tutor) and combining them with Fractions Lab. The platform includes an intervention model for sequencing the structured tasks (realized by the Vygotsky Policy Sequencer) and for switching between structured tasks and exploratory tasks. Furthermore, it provides task-independent support and task-dependent support to learners while they are working on specific tasks. All but one of the adaptive components of the platform work with and without speech indicators (i.e., sequencing structured tasks, switching between structured- and exploratory tasks, and task-independent support). The exception is task-dependent support which relies entirely on students' actions in the system (i.e., making errors), and is therefore not based on speech indicators. Task-dependent support had already been embedded in the existing Tutors for structured tasks and has now additionally been developed for Fractions Lab for the exploratory tasks. We will now first briefly summarize the conclusions that we drew from results of the formative evaluation trials with regard to progress and usability of the components. In this, we will assess the readiness of the components for the summative evaluation and provide contingency plans when we identify a corresponding risk. Based on these results, we will then describe the updated summative evaluation plan and its risks.

3.1 Conclusions and lessons learned from the formative evaluation trials

First of all, we found our approach to evaluate the iTALK2Learn platform productive and successful. The iTALK2Learn platform is a complex learning environment that combines many separate components in one platform to facilitate robust learning. Waiting for the integration of all components before conducting formative evaluations would have meant a considerable delay in developing the platform. Our approach thus was to evaluate and redesign each component in parallel. This required constant communication between the consortium partners working on the



D5.2 Report on formative evaluation results in Y2

respective components. The valuable lessons that were thus shared facilitated the iterative design and evaluation cycles of all components.

Fractions Lab was developed specifically for the iTalk2Learn platform. The formative evaluation trials of the user interface have shown that both students and teachers respond very positively to Fractions Lab. Based on these trials, we have already gained additional insight into which types of exploratory tasks and representations are most productive for stimulating conceptual thinking about fractions. This is a significant achievement for the project because the findings add to the mathematics education literature related to fractions elementary learning and teaching. The summative evaluation will be able to evaluate the iTalk2Learn platform including a fully-functioning version of Fractions Lab.

Speech recognition is a central component of the iTalk2Learn platform. Detecting affect in speech and using it for decision making can ameliorate performance prediction and sequencing models. Affect features have already been identified successfully and a first machine-learning model for detecting these features in speech has been trained. Additional development is still needed to improve affect classification performance. Considerable progress has also been made for word detection in both English and German. Because sufficient accuracy in word recognition has been difficult to attain so far, additional training of the acoustic and language models is still needed. Performance should be much improved by training the acoustic models with additional data and by further focusing the language model on the exact terminology that will be used while working with iTalk2Learn. Specifically, these are utterances concerning task difficulty and the use of mathematical language. These adjustments should lead to sufficient precision and recall on key terminology. An important prerequisite for accurate speech detection is also the quality of audio recordings. In the usual noisy classroom settings, it is especially important to keep noise at a minimum, place microphones carefully, and check recording settings. To prevent quality issues in the future, we have developed guidelines based on our experiences in the formative trials that will be followed in the summative evaluation. Based on these additional steps, we expect that the summative evaluation should be able to evaluate the benefits of speech recognition for automatic adaptivity. The iTalk2Learn platform has also been designed in such a way that should speech recognition not work, or be limited to specific functions of speech recognition such as word detection or affect detection, we will still have a functioning platform. The summative evaluation will then not be able to assess the (full) benefits of speech detection on adaptivity, but it will still be able to assess, for instance, the benefits of combining structured- and exploratory tasks for robust fractions learning.

The Vygotsky Policy Sequencer, that adapts tasks to students within structured practice, has shown great promise in the formative trials as a proof of concept for performance prediction. For Maths-Whizz, trials showed that the Vygotsky Policy Sequencer is at least as good as the existing, curriculum-based Maths-Whizz sequencer. Within the short time frame of the study, it already produced comparable learning gains and was perceived just as helpful as the Maths-Whizz



D5.2 Report on formative evaluation results in Y2

sequencer. The Vygotsky Policy Sequencer also produced a better user experience—for example exercises were seen as less repetitive and more fun. Differences in learning gains and additional advantages of the ease of implementing the Vygotsky Policy Sequencer are expected to be seen when studies run over longer time. For Fractions Tutor, training and evaluation of the sequencer is still ongoing. The summative evaluation will be able to evaluate a full version of the iTALK2Learn platform that includes the Vygotsky Policy Sequencer.

Regarding task-dependent support, much progress has been made on developing a set of rules for learners based on their interaction with the Fractions Lab. The efficacy of the task-dependent support rules has been shown already in WoZ studies and the automatic delivery of this support has been piloted as well. Much work has also been put into further refinement of the support rules, taking into account related experience with providing feedback to students and additional design drivers arising from the literature and the designers' pedagogical approaches. The summative evaluation will be able to evaluate a full version of the iTALK2Learn platform that includes task-dependent support. As a contingency measure, the already tested script can be used for providing task-dependent support by wizards.

Task-independent support based on affect detection has been successfully tested. We determined in WoZ studies which feedback type is most effective given a particular affective state. The rules derived from these studies can now be implemented in the automated system. Task-independent support will be included in the iTALK2Learn platform that is evaluated in the summative evaluation. As a contingency measure, affect detection can also be simulated by wizards or be reduced to interpreting students actions in the system.

Some initial experience with switching comes from studies in the UK where it was already possible to use both Fractions Lab and Maths-Whizz. Based on these studies, we have developed a switching strategy informed by a students' needs analysis. We also started developing a better appreciation of the amount of time it takes students to work through Fractions Lab and Maths-Whizz or Fractions Tutor, respectively. This not only informs D1.3 but helps us plan subsequent studies. For example, one implication is that we might require students to work over several periods of time with the system. These encouraging results will be built upon in the next step on the way to the summative evaluation. The iTALK2Learn platform in the summative evaluation trials will automatically switch between structured- and exploratory tasks. As a contingency measure, the sequence of tasks can be fixed.

Finally, the formative trials have shown technical challenges, as can be expected when newly developed components run for the first time under live conditions and are integrated into a unified platform. To ensure that the iTALK2Learn platform will run smoothly in the summative trials, we will conduct extensive local testing beforehand. For example, in Germany, we will again set up local area networks to implement the iTALK2Learn platform independently from internet access and possible bandwidth issues, like we successfully did in the formative trial conducted in Hildesheim.



D5.2 Report on formative evaluation results in Y2

As a contingency measure, we have set aside time and resources to fix any technical problems that may arise. BBK will collaborate with RUB and IOE, respectively, to resolve technical problems on site. The local testing will also include students working with the system so that the platform will be tested under realistic conditions. This process should ensure that we have a working system by the time the main summative trials start. An additional contingency measure can be to limit the number of students working simultaneously with the system to reduce demands on the system.

Table 3 summarizes the status of each component regarding the summative evaluation. It also provides an overview of contingency plans in the form of efforts that are still needed to get a component ready for the summative evaluation and consequences for the summative evaluation should these efforts be delayed. This contingency plan is elaborated in the Project Periodic Report

Table 3: Overview of evaluation status and contingency plans.

Component	Status	Contingency plans	
		Efforts to reach completion	Consequence for summative evaluation if efforts are delayed
Fractions Lab	Working	n/a	n/a
Word recognition	First models trained Error rate of word detection unsatisfactory	Add key terminology and continue model training Assess precision and recall of key terminology	Only include affect detection OR word recognition is simulated by wizards
Amelioration of performance prediction and sequencing through affect recognition	First model trained Error rate of affect classification unsatisfactory	Find and train alternative model	Use Vygotsky Policy Sequencer without amelioration through affect recognition OR Affect recognition is simulated by wizards
Vygotsky Policy Sequencer for Maths-Whizz	Working	n/a	n/a
Vygotsky Policy Sequencer for Fractions Tutor	Training ongoing	Continue training Formative evaluation in German switching study	Baseline version without Vygotsky Policy Sequencer



D5.2 Report on formative evaluation results in Y2

Task-dependent support	Manual delivery working Automatic delivery for one task implemented	Implement automatic delivery	Delay of summative evaluation OR task-dependent support is simulated by wizards
Task-independent support (mainly based on speech indicators)	Manual delivery working	Complete word and affect recognition to implement automatic delivery	Rely on student actions in the system without speech indicators
Switching	First study with manual switching successful Additional studies are being conducted	Complete formative evaluation trials	Delay of summative evaluation OR fixed sequence of Fractions Lab and Maths-Whizz, or Fractions Tutor, tasks
Integration	First tests successful	Further tests are needed	Delay of summative evaluation OR Smaller number of students at a time (as many as can be accommodated in the IT suite at one time)

3.2 Updated summative evaluation plan

The formative evaluation has shown that all components for the iTalk2Learn platform are either ready or are nearing completion and will be ready in time for the summative trials planned for M29/30 in UK and in Germany. Usability of each component is high and we have generally received positive feedback from teachers and students, which is promising for the summative evaluation and also with regard to long-term dissemination of the platform. For the components that are only nearing completion, we are aware of the risks and developed a contingency plan (see

Table 3 and Project Periodic Report). In addition, we have designed the conditions of the summative trials so that even if, in a worst case, specific components were not working, we would still be able to make meaningful assessments of the working iTalk2Learn platform. This will be detailed in the following update of the summative evaluation plan originally reported in D5.1.

The summative evaluation aims to assess the pedagogical outcomes of the project in combination with the technological feasibility of the whole iTalk2Learn platform. In particular, we will empirically investigate two hypotheses:



D5.2 Report on formative evaluation results in Y2

- 1) Combining structured practice and exploratory tasks promotes robust learning.
- 2) An adaptive system with indicators from speech enhances learning more than an adaptive system without speech indicators.

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), in two settings (controlled setting and realistic setting), and using different tutorial systems for the structured tasks (Fractions Tutor and Maths-Whizz). We will now detail participants, research design, measures on which outcomes are assessed, and procedure of the experiments. We will conclude with an assessment of the risks involved in the summative evaluation and respective contingency plans.

3.2.1 Participants

For the experiments, we will recruit children from 5th grade classes. We have chosen 5th grade because:

- the arrangements of the National Curriculum and the curriculum in North Rhine-Westphalia are such that 9-10 year old students in UK and 11-12 year old students in North Rhine-Westphalia are at an appropriate stage of readiness for the mathematics content we focus on in the project. Formal fractions instruction starts at the beginning of 6th grade so that the last months of 5th grade are an ideal time to introduce fractions with iTalk2Learn.
- during our formative studies we found the 5th grade students benefited from the iTalk2Learn platform more than the older students, who appeared to have more established existing procedural methods they drew upon, making assessing their conceptual gains difficult in the post-test.

In UK, we will recruit primary schools via the existing school networks of IOE and Whizz. In Germany, classes will be recruited with the help of the School Lab of the Ruhr-Universität Bochum (RUB, 2014). The RUB School Lab is an extracurricular 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 School Lab has many school contacts that will be activated to recruit students. Studies will preferably be conducted with whole classes. As a reward for participation, teachers will receive a voucher to purchase learning materials for the classes.

3.2.2 Research design

To test our hypotheses, we will compare multiple versions of the iTalk2Learn platform as displayed in



D5.2 Report on formative evaluation results in Y2

Table 4. Although the participating students will be working in close proximity to one another, the system makes it possible to allocate students within the same class and session to different versions of the iTalk2Learn platform. Accordingly, we will aim to conduct each session with whole classes of students (or as many as can be accommodated in the IT suite at one time) but assign each student randomly to one of the following conditions.



D5.2 Report on formative evaluation results in Y2

Table 4: Conditions in the experiments of the summative evaluation

Components of iTALK2Learn utilized	Experimental conditions			
	Full version with speech	Full version without speech	Version without Fractions Lab	Baseline version
Vygotsky Policy Sequencer for structured tasks ²	Yes, with speech indicators.	Yes, without speech indicators.	Yes, with speech indicators.	No.
Switching between structured and exploratory tasks ³ .	Yes, with speech indicators.	Yes, without speech indicators.	No.	No.
Task-independent support	Yes, with speech indicators.	Yes, without speech indicators.	Yes, with speech indicators.	No.

Two conditions are based on the full versions of the iTALK2Learn platform; one with and the other without speech. The full versions will be implemented in both the UK and Germany, and will work with the implemented intervention model of D1.3 and with structured tasks from either, Maths-Whizz in the UK or Fractions Tutor in Germany. We then have two control conditions, one using a baseline version and the other a version of iTALK2Learn without Fractions Lab. The version without Fractions Lab serves to evaluate the contribution of this component within iTALK2Learn (hypothesis 1). To this purpose, the condition will include only the newly developed Vygotsky Policy Sequencer, and automatic adaptivity based on speech indicators, but not Fractions Lab. The baseline version represents practice as-is: learning within either of the two existing tutorial systems, without the new sequencer, without speech detection, and without Fractions Lab. By comparing the baseline version to the full versions, we can provide further tests of hypothesis 1 that a combination of procedural and conceptual learning promotes robust learning of fractions. The added benefit of adaptivity based on speech recognition (hypothesis 2) will be tested by comparing the two full versions of the iTALK2Learn platform to each other.

3.2.3 Measures

We will measure effectiveness for learning by pre- to post-test gains. We have developed and piloted test items during the formative trials that measure differences in understanding that are targeted by our learning platform. In addition, we will compare the conditions with regard to students' satisfaction with the respective version of the system and their motivation/engagement, both measured by means of questionnaires. In previous projects as well as in the formative trials of this project, we have employed a 5-point Likert scale to evaluate important constructs including

² The structured tasks include task-dependent support as provided by Whizz or Fractions Tutor.

³ The exploratory tasks include task-dependent support, which has been developed as described in 2.3.3.



D5.2 Report on formative evaluation results in Y2

perceived helpfulness or repetitiveness, comprehension, and affect (see section 2.3). This scale is appropriate for children of the age group we work with in iTalk2Learn because it uses a visual analogue scale that employs pictorial representations that children can relate to (e.g., the Fun Toolkit in Read, 2008). We will also conduct interviews and focus groups with selected students to gain a deeper understanding of the platform's effects on learning and motivation.

3.2.4 Procedure

In UK, IOE will evaluate the iTalk2Learn platform in a realistic setting. Teachers will supervise the learning process and will provide the students not only with technical support, but also individual support as and when required. However, they will be asked to prioritise any process and technical issues, leaving learning-related support on an as-needed basis. Depending on the school's approach, learning time may be extended beyond class by providing students the opportunity to continue work online from home. The aim of the experiment in the UK is to ensure ecological validity of the study, that is, that the sessions replicate as far as possible what might happen in a typical IT suite in a typical school when students are using computers to study mathematics.

In Germany, RUB will evaluate the iTalk2Learn platform in a controlled setting. Researchers will supervise the learning process and only provide technical support, no learning-related support. Interaction with the students will be standardized. Teachers will only observe. Learning time will be strictly controlled and limited to the experimenter-led session. The aim of the experiment in Germany is to ensure internal validity of the study, that is, that the sessions replicate as far as possible what might happen in a laboratory setting where possible confounding variables are strictly controlled.

In both experiments, students will complete a pre-test that is centred on their conceptual and procedural understanding of fractions, based on the instrument piloted in the UK and in Germany by IOE and RUB (see Sections 2.1 and 2.2.4). They will then engage with the iTalk2Learn platform in one session of up to 90 minutes in Germany, and in three sessions of up to 45 minutes, one session per day, in the UK. After the learning phase with the iTalk2Learn platform, they will complete a post-test that is directly comparable to the pre-test (i.e., covering the same assessment areas and same levels as the pre-test).

3.2.5 Risks of the summative evaluation

In addition to the contingency plans with regard to the single components of the platform presented in Table 3, we are considering additional risks inherent in conducting the summative evaluation experiments.

Table 5 gives an overview of these risks, what we have already done to address them, what we will continue to do to mitigate these risks, and what consequences for the summative evaluation would arise from risks that cannot be mitigated in time. We plan to implement all four conditions, but



D5.2 Report on formative evaluation results in Y2

there is a risk we will not be able to recruit enough participants and thus would lack statistical power for detecting meaningful differences between all four conditions. In this case, we would only implement one control condition in each of the two summative experiments. This could be done, for instance, by using the baseline version in UK and the version without Fractions Lab in Germany, thus reducing the overall number of conditions to three. Should sample sizes not be sufficient to include three conditions, we would compare the version without Fractions Lab to the full version (see also contingency plan detailed in Project Periodic Report). In addition, there are some risks concerning the local implementation of iTalk2Learn in schools, such as low audio quality, limited internet bandwidth, and potential lack of computer labs. As described in

Table 5, we are already taking measures in order not to let these risks delay the summative evaluation. We consider the probability of a delay due to these risks to be low.

Table 5: Implementation risks during summative trials and contingency plans.

Risk	Status	Contingency plans	
		<i>Efforts to reach completion</i>	<i>Consequence for summative evaluation if efforts are delayed</i>
Low number of schools and students volunteer for summative trials	Recruitment efforts have started Experimental plan is designed with optional conditions	Continue recruitment efforts	Experimental plan reduced to two or three conditions
Audio recording quality is low	Guidelines for audio quality are specified in writing	Test audio settings in participating schools at start of summative trials	Speech detection accuracy suffers (but overall student experience might not have major impact)
Internet bandwidth is limited	Implementation of iTalk2Learn in a local area network has been successfully tested	Additional testing needed in participating schools at start of summative trials	Delay of summative evaluation OR smaller number of students at a time (as many as can be accommodated in the IT suite at one time)
Schools lack computer labs	Laptops are reserved (Germany)	Check availability of computers in participating schools	Delay of summative evaluation



D5.2 Report on formative evaluation results in Y2

4. Conclusion

The formative evaluation of each component in parallel has been a successful approach to evaluating the complexity of a platform such as iTalk2Learn. For each component, we have learned valuable lessons that were exchanged between consortium partners and that benefited the development and evaluation of the other components. Progress in developing and evaluating the components so far has been good. One important final step on the way to the summative evaluation is the switching trial of the integrated, working platform. After deliverables and reporting period are completed, we can move on to realizing the plans for the summative evaluation. The summative evaluation will show how well the components work together, and how well the goal of iTalk2Learn of providing a platform for robust fractions learning with adaptive, speech-enhanced support will be met by the final product. The results of the summative evaluation will be reported in D5.3.



D5.2 Report on formative evaluation results in Y2

References

- Cobb, P., Confrey, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiment in educational research. *Educational Researcher*, 32(1), 9-13. doi:10.3102/0013189X032001009
- 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.
- Hansen, A., Geraniou, E., & Mavrikis, M. (2014, September). *Designing interactive representations for learning fractions*. Paper presented at the British Educational Research Association Conference, London.
- Hansen, A., Mavrikis, M., & Geraniou, E. (2014). *Professional development through cooperative inquiry: Improving teachers' technological pedagogical content knowledge of fractions*. Manuscript submitted for publication.
- Janning, R., Schatten, C., & Schmidt-Thieme, L. (2014a). Multimodal affect recognition for adaptive intelligent tutoring systems. *Extended Proceedings of the 7th International Conference on Educational Data Mining*, 171-178.
- Janning, R., Schatten, C., & Schmidt-Thieme, L. (2014b). Feature analysis for affect recognition supporting task sequencing in adaptive intelligent tutoring systems. In C. Rensing, S. de Freitas, T. Ley, & P. Muñoz-Merino (Eds.), *Lecture notes in computer science. Open learning and teaching in educational communities* (pp. 179-192). Springer International Publishing. doi:10.1007/978-3-319-11200-8_14
- 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. doi:10.1111/j.1551-6709.2012.01245.x
- Mavrikis, M., Grawemeyer, B., Hansen, A., & Gutierrez-Santos, S. (2014). Exploring the Potential of Speech Recognition to Support Problem Solving and Reflection. In C. Rensing, S. de Freitas, T. Ley, & P. Muñoz-Merino (Eds.), *Lecture Notes in Computer Science. Open Learning and Teaching in Educational Communities* (pp. 263-276). Springer International Publishing. doi:10.1007/978-3-319-11200-8_20
- Oliver, M. (2011). Technological determinism in educational technology research: some alternative ways of thinking about the relationship between learning and technology. *Journal of Computer Assisted Learning*, 27 (5), 373-384.
- Rendle, S., & Schmidt-Thieme, L. (2008). Online-updating regularized kernel matrix factorization models for large-scale recommender systems. In F. L. Kastensmidt & G. Bronevetsky (Eds.), *Proceedings of the 2008 Workshop on Radiation Effects and Fault Tolerance in Nanometer Technologies* (pp. 251-258). New York, N.Y.: ACM Press.
- Schatten C., & Schmidt-Thieme, L. (2014, April). *Adaptive content sequencing without domain information*. Paper presented at the 6th International Conference on Computer Supported Education, Barcelona.
- Schatten, C., Wistuba, M., Schmidt-Thieme, L., & Gutierrez-Santos, S. (2014). Minimal Invasive Integration of Learning Analytics Services in Intelligent Tutoring Systems. In IEEE Computer



D5.2 Report on formative evaluation results in Y2

- Society (Ed.), *14th International Conference on Advanced Learning Technologies (ICALT)* (pp. 746–748).
- Schatten, C., Janning, R., Mavrikis, M., & Schmidt-Thieme, L. (2014). Matrix factorization feasibility for sequencing and adaptive support in intelligent tutoring systems. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), *Proceedings of the Seventh International Conference on Educational Data Mining* (pp. 385–386).
- Silver, E. A. (1983). Probing young adults' thinking about rational numbers. *Focus on Learning Problems in Mathematics*, 5, 105-117.
- Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). *Implementing standards-based mathematics instruction: A casebook for professional development. Ways of knowing in science series*. New York: Teachers College Press.
- van Joolingen, Wouter 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. *Learning in Innovative Learning Environments*, 21(4), 671–688. doi:10.1016/j.chb.2004.10.039
- Wong, M., & Evans, D. (2007). Students' conceptual understanding of equivalent fractions. In J. Watson & K. Beswick (Eds.), *Mathematics: essential research, essential practice. Proceedings of the 30th annual conference of the Mathematics Education Research Group of Australasia* (pp. 824–833). Adelaide, S.A.: MERGA.



D5.2 Report on formative evaluation results in Y2

Appendix 1

Study Report - Fraction Lab's User Interface

Date(s): Month 17

Participants:

23 Mathematics Specialist Teachers

Aim(s) of study:

To receive feedback from teachers on the Fractions Lab user interface

Method:

The teachers were invited to attend an optional 50-minute professional development workshop during a scheduled Mathematics Specialist Teacher day.

The focus was to provide teachers with the opportunity to familiarise themselves with Fractions Lab and offer them the opportunity to design tasks they could use in their classrooms and with colleagues. During the session we briefly introduced the project and Fractions Lab.

The teachers were given a brief introduction to Fractions Lab and then an opportunity to explore Fractions Lab with the remit to think about the tasks they could plan for their pupils.

The session concluded with a group discussion about Fractions Lab where we took feedback on the design to feed into the next iteration.

Results/findings:

Indicative teacher comments from the workshops about the user interface:

- *Nice ICT resource for children to explore*
- *It would be useful to combine fractions together - drag and drop*
- *I found it tricky to learn to use the interface within this short session and so I would need to go back to consider the use of Fractions Lab in teaching*
- *When changing the denominator the different images were quite confusing. It would be useful to have a 'clear page' option.*
- *There are a few technical glitches with the partitioning where the lines do not show.*
- *The blank representation needs to be able to go straight to the bin.*
- *Sometimes the partition button hid behind the representation buttons.*

Results from teacher questionnaires about using Fractions Lab again:

- 100% of the teachers expressed interest in Fractions Lab, stating they were at least likely to use it in their own teaching

22 of the 23 teachers (96%) reported they were at least quite likely to recommend it to colleagues. The teacher who did not feel she could recommend qualified this by explaining that she might recommend it



D5.2 Report on formative evaluation results in Y2

once she had seen Fractions Lab as a final product.

Conclusion:

The teachers received Fractions Lab positively. They could see its value in supporting students' conceptual understanding of fractions and there was a general willingness to use Fractions Lab in their teaching.

There is some work to undertake to make Fractions Lab as user-friendly for teachers as possible, but this is to be expected at this stage of development. A number of changes were submitted to Testaluna. These included a small number of bug fixes (e.g. some lines do not show when partitioning rectangles), some improvements to user interface (e.g. being able to drag and drop a blank representation into the bin; double clicking on a fraction to add/subtract it), and some enhancements to conceptual support (e.g. changes to the addition/subtraction animation). Testaluna implemented the changes in the next version of Fractions Lab.



D5.2 Report on formative evaluation results in Y2

Appendix 2

Study Report - The impact of Fractions Lab and associated tasks on students' conceptual understanding of fractions

Date(s): M20

Participants:

36 Year 6 students (10-11 years old)

Aim(s) of study:

To identify how the design decisions in Fractions Lab impact on students' conceptual understanding of fractions

Method:

During the study each student worked with Fractions Lab for a duration of 15 - 30 minutes in an authentic classroom setting (computer suite).

The cohort was given a pre-test the day before their Fractions Lab experience and post-test the day after. We were interested in the range of representations the students drew upon to represent fractions so the pre-test began by asking the Year 6 students to show $\frac{1}{4}$ as many ways as they could and then to brainstorm what they knew about fractions. The post-test involved returning their earlier work to them and asking them to amend/add/delete anything they wished to so that we could identify if Fractions Lab had changed their thinking about fraction representations.

The students were also given a 'reflection about my learning' questionnaire that involved questions about the impact Fractions Lab had on their understanding about fractions. They were also asked to think about how the different representations helped their fractions learning and to identify their preferred representation and why. We used this to identify how their experience using Fractions Lab may have changed their thinking about fractions immediately after using the program.

Results/findings:

Selected data from "Show $\frac{1}{4}$ in as many ways as you can":

	Number of students writing or drawing representation prior to Fractions Lab	Number of students adding representation after Fractions Lab	Total number of students using the representation
Equivalent fraction(s) listed	89%	3%	92%
Circle(s) draw	89%	11%	100%
Square(s) drawn	34%	6%	40%
Vertical rectangle drawn	34%	20%	54%
Horizontal rectangle drawn	9%	20%	29%
Partitioned rectangle drawn	11%	29%	40%
Jug		89%	89%
Number line		77%	77%



D5.2 Report on formative evaluation results in Y2

Triangle		6%	6%
Irregular figures		6%	6%
Nothing added		3%	3%

"What is your preferred representation and why?"

Preferred representation	Indicative comments
Rectangles (49%)	- Because I'm used to rectangles when I'm being taught [sic] - Because it helped me to understand partitioning best
Liquid measures (34%)	- It helps me understand more and it lays it out more understandably - You can put a jug on a jug to see if a fraction is equal - Because you can read more easier and the jug is bigger so it's easier to make out
Number line (11%)	- I found it much clearer than the others - It really helped me to understand
No preference (3%)	- I had seen all the representations before

"What have you learned about fractions using Fractions Lab?"

Indicative comments about fractions representations	Indicative comments about equivalence	Indicative comments about addition and subtraction	Indicative comments about fraction size
- They can be represented in many ways - The different ways to show fractions - I learnt that you can do fractions with a jug - That fractions can be anything from water jugs to number lines	- Equivalence. How to find them. By going up in the multiples - That when you find an equivalent fraction it is not always times by 2 - Seeing the fractions, I can see it is not equivalent	- $1/3 + 2/6$ doesn't equal $3/9$ - How to add fractions, changing the denominator to match - Which fractions are the same because it colours in the rectangle - You can multiply the fractions and it will multiply the picture for it too	- Taught me a lot. How to tell if it is smaller or bigger. I made $1/3$ and $1/5$ and the $1/3$ section was bigger. I didn't know that before - If the denominator is bigger than the other denominator that fraction is actually smaller

Conclusions:

It is unsurprising that rectangles remained the most popular **representation** due to area models being the predominant representations used by the teachers, curricula materials (Alajami, 2012; Pantziara & Philippou, 2012) and the students all identifying an area model (the circle) in their pre-tests. However, we were surprised by the number of students stating they preferred number lines (11%) or liquid measures (34%) because of the relatively short time they had on Fractions Lab. Our findings add weight to Silver's (1988) supposition that representations beyond the area model may help students' fractions understanding by enabling flexibility and Steen *et al.*'s (2006) finding that students can use virtual manipulatives more flexibly. This is an area worth exploring further, both for the students' conceptual understanding of fraction and in our next iteration of Fractions Lab.

Many of the students used the **partitioning tool**. They observed patterns in how the fraction symbol was changing and drew conclusions using multiplicative reasoning. This supports Olive & Lobato's (2007) premise that a student can establish a relationship between a part and a whole by partitioning, changing the denominator and numerator while leaving the original whole intact. Furthermore, the *process* of constructing (Mariotti, 1997) the fractions showed how particular fraction instantiations supported students in considering fraction equivalence.



D5.2 Report on formative evaluation results in Y2

The students made situated abstractions about **adding and subtracting fractions** and how the fractions required the same denominator. Some students explained how they made equivalent fractions to enable an addition or subtraction to take place. By using the Partitioning tool to change fractions so they share the same denominator before they are added or subtracted shows how the tools "entered the students' thoughts, actions and language" (Noss and Hoyles, 1996:59). Furthermore, the students were able to choose the extent to which they used them (Clarebout, Elen, Johnson & Shaw, 2002).

A number of students made statements about the **size of fractions**, particularly how the denominator increasing made the proportion of the fraction decrease. We were particularly surprised by this finding. We had made an assumption that these students would have already been aware of the relative size of fractions. This appears to be as a direct result of the students being able to see a dynamic virtual manipulative rather than a static fraction being drawn by the teacher or on a worksheet, which enabled students to undertake "profoundly different" (Nardi, 1998) actions that enabled them to "act, perceive and reason beyond [their] natural limits" (Nunes, 1997:30).



D5.2 Report on formative evaluation results in Y2

Appendix 3

Study Report - WOZ in M17-M20 for task-dependent support in Fractions Lab

Date(s): March 2014 (M17) (continued iteratively through to M20)

Participants: In total 12 students took part in the WOZ study. Data pre-processing errors we were able to analyse the interaction of only 10 students. Year-5 (9 to 10-year-olds).

Aim(s) of study:

- To investigate the provision of task-dependent support (TDS) that has the aim of providing personalised feedback during interaction with the ELE, to help the student to deal with and learn from errors that they made while responding to the tasks.
- To identify effective problem solving support.

To identify when and how students could be encouraged to verbally reflect on their learning

Method: Ecologically valid WOZ in a classroom.

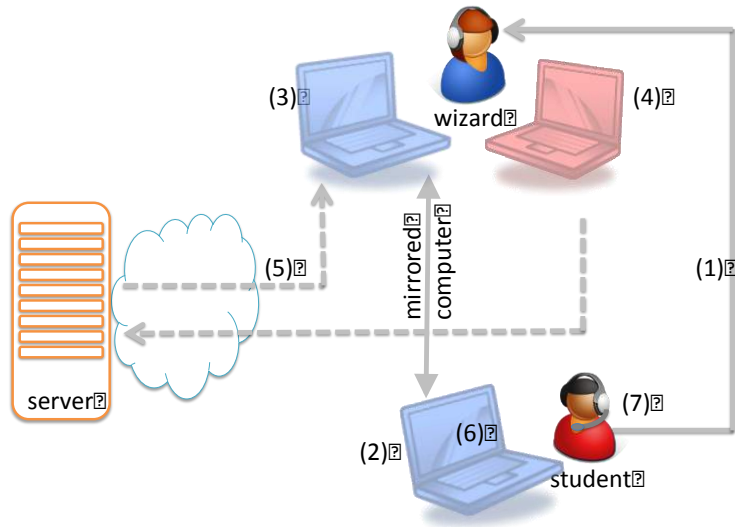
A WoZ study undertaken in parallel to the formative evaluations of the ELE. More details of the study are reported in Mavrikis *et al.* (2014)⁴.

In a classroom equipped with computers, two computer were set up to allow human facilitators (wizards) to listen to students thinking-aloud while having access to their interaction with the environment. The setup is shown below.

⁴ Mavrikis, M., Grawemeyer, B., Hansen, A., Gutiérrez-Santos, S. (2014) Exploring the Potential of Speech Recognition to Support Problem Solving and Reflection - Wizards Go to School in the Elementary Maths Classroom. In proceedings of EC-TEL 2014 (pp 263-276) Also available online <http://www.italk2learn.eu/wp-content/uploads/2014/07/ECTEL-WOZ.pdf> (DOI: 10.1007/978-3-319-11200-8_20)



D5.2 Report on formative evaluation results in Y2



Each student speaks on a headset (mic) that is connected to the wizard's headphones (1). The student interacts with a console (i.e., keyboard, mouse, screen) that is connected to a laptop on the wizard's side (2,3) so as the latter can witness their interaction. The wizard can send messages and change the task sequence (4) by using specially designed wizard tools. These messages arrive to a server and subsequently to the mirrored laptop) (5) where they can be seen (6) and heard (7) by the student.

The wizards provided support using a script and following an iterative methodology that deliberately limited their communication capacity (Mavrikis, Gutierrez-Santos, 2010⁵) in order to simulate the actual system. Whenever the student was observed by the wizard to be having difficulties or to be making errors in the iTalk2Learn system (ELE and structured task), an appropriate TDS or TIS statement was copied from the pre-written WoZ script, pasted into the system and spoken out loud to the student by the automatic speech production system.

The wizard was not physically near enough to the students to observe them directly, and therefore observed them by indirect mediated means: the student's voice was heard by using microphones and headsets, and their screen was observed on a second screen that mirrored the first.

Also, the wizard did not have direct access to the students' screens (so e.g. could not point to anything on their screens for emphasis), could not see the students' faces (for facial cues), and could not communicate to students by using body language, only by means of the facilities provided by the wizard-of-oz tools, which resemble those of the final system.

⁵ Mavrikis, M & Gutierrez-Santos, S 2010, 'Not all wizards are from Oz: Iterative design of intelligent learning environments by communication capacity tapering' *Computers and Education*, vol 54, no. 3, pp. 641-651. doi:10.1016/j.compedu.2009.08.033



D5.2 Report on formative evaluation results in Y2

Those tools included a script for the wizard to decide when and what type of feedback should be provided to the students, based on their interaction with the learning environment, and their performance.

Based on the script, problem solving support was provided, when the student was struggling with the task. Reflective prompts were provided if the wizard decided that the student would either benefit from reflecting while they were performing the task, or when they finished the task at the end of the exercise.

Results/findings:

74 messages were provided to students (40 problem solving and 34 reflective prompts). Table 1 shows that students mainly reacted to problem solving or reflective prompts (85%).

	Reacted		
Feedback type	1	0	Total
Problem solving	33	7	40
Reflection	30	4	34
Total	63	11	74

Table 1: Student reaction towards the different feedback types

Conclusion:

The results show that students mainly reacted to the task-dependent feedback. There was no difference in reaction to problem solving and reflection feedback. This implies that students were open to both traditional problem solving support, and also with less traditional reflective feedback, where they had to verbalise their reflection.

An independent t-test revealed no significant difference between the feedback types and whether students reacted ($t(72)=-.68, p>.05$).



D5.2 Report on formative evaluation results in Y2

Appendix 4

Study Report – WoZ Formative Evaluation of task-dependent support

Date(s): M20 2014

Participants: 17 students - Year 5 students (10-11 years old)

Aim(s) of study:

Comparing the effectiveness of automatic task-dependent support to support provided by a human (wizard).

Method:

Ecological valid user study (including WoZ) in a classroom.

A user study, where children used Fractions Lab to perform different tasks. In one of those tasks support was provided automatically by the task-dependent support. On all other tasks, a human facilitator (wizard) provided support based on the child's interaction with Fractions Lab. More Details on the method are provided in the previous Study Report (Appendix 3).

On tasks, where the automatic support was not available, the wizard observed the student by indirect mediated means: the student's voice was heard by using microphones and headsets, and their screen was observed on a second screen that mirrored the first.

Results/findings:

153 messages were provided to students. Out of those 75 messages were provided by the automatic task-dependent support. Table 1 shows the different feedback provided to students (automatic and WoZ) and a positive affect that occurred after the feedback was provided.

	Automatic	WoZ	Total
Problem solving	58 (29 positive.)	45 (27 positive)	103
Reflection	17 (9 positive)	33 (21 positive)	50
Total	75 (38 positive)	78 (48 positive)	153

Table 1: Messages provided and affect that occurred afterwards (positive).

An independent t-test revealed no significant difference between the automatic and the wizarded feedback on positive affect that occurred afterward the feedback was provided ($t(151)=1.35$, $p>.05$).

Conclusion:

Statistical analysis of the user study showed that there was no significant difference between the students' response (positive affect) to TDS feedback delivered automatically and TDS feedback delivered by the human wizard in the other tasks.



D5.2 Report on formative evaluation results in Y2

Appendix 5

Study Report – WoZ Presentation of feedback for task-dependent support¹

Date(s): M20 2014

Participants: 17 students - Year 5 students (10-11 years old)

Aim(s) of study:

To investigate if there is an effect of different presentations of feedback (high or low interruptive) on the affective state of a student, following provision of the feedback

Method:

Ecological valid WoZ in a classroom.

Details of the method are provided in the previous Study Report (Appendix 3).

Participants were randomly assigned to two groups (8 participants in the high and 9 participants in the low interruptive feedback group). In both groups feedback was provided to students based on their speech and on their interaction with the learning environment. This feedback included for example, problem solving support, reflective prompts, and affect boosts. Based on which group the student was assigned to, they either received this feedback in a 'high interruptive' way as a pop-up window or in a 'low interruptive' way through an indication that feedback is available which they could access by clicking on a glowing light bulb button.

Participants in the low interruptive group were able to ignore the feedback provided, by not clicking on the highlighted light bulb. In contrast, participants in the high interruptive group had to dismiss the pop-up window before they could proceed with the task.

Participants in both groups performed a range of tasks with Fractions Lab where different feedback messages were sent by the wizards in the presentation format assigned to the group, for around 15 minutes.

Results/findings:

In total 306 messages were sent to 17 students (153 high interruptive and 153 low interruptive messages).

The raw video data was analysed independently by two researchers who categorised the affective states of students while the feedback messages were provided. The results of the categorisations were compared against each other. There was a match of 76%. Where there was a mismatch, the categorisations were re-analysed and agreed upon between the researchers.

Only three out of the five affective states were detected during this study (enjoyment, confusion, and frustration). This might be because the sessions only lasted 15 minutes and students did not get bored during this short time. Also, surprise does not seem to occur frequently. In order to investigate whether



D5.2 Report on formative evaluation results in Y2

there was an effect of the modality of the feedback presentation on the learning experience, we looked at whether a student's affective state was enhanced, stayed the same or worsened.

We apply chi-square tests to investigate statistical significant differences between the groups, as the data is categorical. When students were enjoying their activity there was no significant association between the groups on whether their affective state stayed the same or worsened after feedback was sent ($X^2(1)=.22$, $p>.05$). Students in the high interruptive group mainly stayed within the same enjoyment state (85%). Their affective state worsened in 15% of cases. Similarly, the low interruptive group stayed mainly in the same affective state (82%), and worsened in 18% of cases.

However, when students were confused there was a significant association of the group on whether their affective state improved, stayed the same, or worsened, $X^2(2)=7.52$, $p<.05$. Here, within the high interruptive group students' affective state was enhanced in 41% of cases, stayed the same in 58%, and worsened in only 1% of cases. In contrast, in the low interruptive group, the affective state was enhanced in 33% of cases, stayed the same in 55%, and worsened in 12%.

Additionally, When students were frustrated there was also a significant effect of the group on whether their affective state improved, or stayed the same, $X^2(1)=4.43$ $p<.05$. Here, in the high interruptive group, there was an enhancement of students' their affective state in 71% of cases. For the other 29%, affective state remained the same. In contrast, the low interruptive group affective state was enhanced in only 23% of cases, and stayed in the same for 77%.

Further, Within the low interruptive group there was a significant association between the different affective states and whether or not students clicked on the light bulb to view the feedback ($X^2(2)=13.12$, $p<.05$). When students were enjoying their activity they clicked on the light bulb in 71% of cases, when confused in 81%, but when frustrated in only 31% of cases.

Additionally, When students were confused within the low interruptive group, there was a significant association between clicking on the light bulb and when the affective state enhanced, stayed the same, or worsened ($X^2(2)=11.26$ $p<.05$). Here, when students viewed the feedback they enhanced their affective state in 41%, stayed the same in 53%, and worsened in 6% of cases. When students did not view the feedback they stayed the same in 64%, and worsened in 37%.

Further, When students were frustrated within the low interruptive group, there was a significant association between message viewed and if the affective state got enhanced, or stayed the same ($X^2(1)=8.78$ $p<.05$). When students viewed the feedback they enhanced their affective state in 75%, and stayed the same in 25% of cases. When students did not view the feedback they stayed the same in 100% of cases.

Conclusion:

When students were enjoying their activity high and low interruptive feedback were both effective.

When students were confused the results show that they welcomed feedback. However, when students ignored the feedback available in the low interruptive group, this resulted in a significantly worsened affective state. The reason why students ignored the feedback might have been that their motivation at



D5.2 Report on formative evaluation results in Y2

this point was low. In order to enhance the learning experience when students are confused, high interruptive feedback should be provided.

Within the low interruptive group, frustration was associated with not viewing the feedback, but when viewed it was associated with an enhanced affective state. This indicates that when students were frustrated they ignored the low interruptive feedback. Frustration can increase cognitive load, which might explain why students did not react to the highlighted light bulb, as they might not have realised that help was available. Therefore, the presentation of the feedback should be highly visible and interruptive when students are frustrated as it is otherwise likely to be ignored.



D5.2 Report on formative evaluation results in Y2

Appendix 6

Study Report - WoZ for TIS – Reaction to Feedback ¹

Date(s): M16 and M20 2014

Participants: 27 children - Year-5 (9 to 10-year old)

Aim(s) of study:

- Is there an effect of different affective state types upon reaction towards feedback?
- What feedback was most successful given a particular affective state?
-

Method:

Ecologically valid WoZ in a classroom (for a description of the WoZ method see Appendix 1).

The types of feedback provided by the wizard ranged across affect boosts, talk aloud prompts, talk mathematics reminders, problem solving, reflective prompts, and other task-independent support

Results/findings:

In total 434 messages were sent to 27 students. The raw video data was analysed independently by two researchers who categorised the affective states at the time the feedback messages were provided, and noted whether there was a reaction. The results of those categorisations were compared against each other. There was a match of 76% of categorisations. Where there was a mismatch, the categorisations were re-analysed and agreed upon between the researchers.

Table 1 shows the different types of messages sent to students, the affective states that occurred at the time the feedback was given; and whether they reacted to the feedback.

Feedback type	Affective state					total
	enjoyment	boredom	confusion	frustration	surprise	
AFFECT	45 (15)	2 (2)	25 (11)	6 (4)	0 (0)	78 (32)
TALK ALOUD	39 (27)	0 (0)	44 (27)	2 (2)	0 (0)	85 (56)
TALK MATHEMATICS	5 (3)	0 (0)	5 (4)	0 (0)	0 (0)	10 (7)
PROBLEM SOLVING	48 (36)	1 (1)	81 (48)	11 (2)	0 (0)	141 (87)
REFLECTION	32 (30)	2 (2)	42 (32)	7 (5)	1 (1)	84 (70)
OTHER	13 (8)	2 (2)	20 (15)	1 (1)	0 (0)	36 (26)
Total	182 (119)	7 (7)	217 (137)	27 (14)	1 (1)	434 (278)

Table 1. Feedback types, including affective state that occurred; and whether there was a reaction after the feedback was provided, in (brackets).

A two-factorial ANOVA revealed no main effect of feedback type ($F(5, 412)=1.24, p>.5$) or affective state ($F(4, 412)=1.13, p>.05$) on reaction towards feedback. However, there was a significant interaction **between** feedback type and affective state ($F(4, 412)=1.80, p<.05$) on reaction towards feedback.



D5.2 Report on formative evaluation results in Y2

When students enjoyed the activity, there was a significant effect of feedback type on reaction ($F(5,176)=8.14, p<.01$). Here students responded very well to reflective prompts (94%). The least reaction occurred if students were provided with an affect boost (33%).

When students were confused, there was no significant effect of feedback type on reaction ($F(5,211)=1.91, p>.05$). Students reacted to all feedback types similarly.

When students were frustrated, there was a significant effect of feedback type on reaction ($F(4,122)=2.93, p<.05$): frustrated students did not respond well to problem solving feedback (18%).

Conclusion:

Is there an effect of different affective states upon reaction towards feedback?

The results show that across the different affective states students mainly reacted to feedback positively.

Students mostly reacted to feedback received when they were enjoying their activity. This is an interesting finding, as in theory feedback would interrupt their learning flow. Here, it appears that student motivation was high and they did not mind being interrupted. Students particularly reacted positively to feedback to reflect.

Also in most cases where students were confused, they reacted to the feedback. This implies that students welcome feedback that could help them to get out of their confused state. Thus, in designing feedback for learning environments students should be provided with feedback that enables them to overcome their confusion, such as task-dependent problem solving feedback, or feedback to reflect on their learning, which might help them to identify and overcome misconceptions.

In contrast, when students were frustrated, they reacted to feedback in only 52%. This indicates that frustration can reduce motivation and may also increase cognitive load. Here feedback that might help to decrease the frustration, such as reflecting on the difficulty of the learning task, might help to motivate the student.

If students were bored, any type of feedback was reacted to. This suggests that students may welcome a distraction from their learning and react to feedback if they are bored. As boredom indicates a reduction in learning, the feedback provided to students when they are bored should aim to motivate and support the student to continue with the learning task.

Which interventions were most successful given a particular affective state?

The results indicate that for certain affective states, different feedback types are more effective than others.

Providing affect boosts were most effective when students were bored (100%) and when students were



D5.2 Report on formative evaluation results in Y2

frustrated (67%). The focus group confirmed that students liked the encouragement, and that it helped with their motivation to continue to work on the particular learning task especially if they were frustrated. In contrast, students only reacted to affect boosts in 44% of the cases when they were confused or 33% when they were enjoying their activity. This might indicate that students found the feedback too interruptive.

It is interesting to see that although students who were frustrated reacted to feedback in only 52% of the cases, they responded well to talk aloud prompts 100%. Here, students might have found it helpful to talk about their problems in performing the learning task. When students were enjoying their activity they responded in 69% of cases to talk aloud prompts. Here, students might have found the feedback inviting them to talk aloud too interruptive. This was similar when students were confused (64%). Providing prompts to talk mathematics was very effective if students were confused (80%), which might indicate that reminding students to use a-specific mathematics vocabulary might help them to think through the problem and resolve their confusion. In contrast when students were enjoying the activity they only reacted in 60% of the cases to talk mathematics prompts. Again, students might have found this feedback to interruptive when they were enjoying they activity.

The highest number of reactions to problem solving feedback was given by students who were enjoying their activity; 75%. Here, although students enjoyed their activity they seem to be open to receiving support in performing their learning task. However, in only 59% of the cases was problem solving feedback reacted to while students were confused. This might be because students were trying to resolve their problem to solve the task and might have found the feedback too interruptive, as it might have suggested switching to a new strategy for answering the task. The number drops further when students were frustrated (18%). Here, students' motivation might be low when frustrated and also there might be increased cognitive load. Providing problem solving feedback when students are frustrated does not seem to be a very effective strategy.

Prompts to reflect were very effective across the affective state types. When students enjoyed their activity they reacted in 94% of the cases to reflective prompts, when they were confused in 74% and even if they were frustrated in 71% of the cases. This implies that reflecting on one's own strategy of solving a task is motivating even if confused or frustrated. We noticed that it may also help students to identify misconceptions or lead to new ideas on how to solve the learning task.

Providing non-learning related prompts were also effective across the affective types (boredom and frustration 100%; confusion 75%), except if students were enjoying their activity (61%). This might imply that students did not want to be interrupted and wanted to continue with their activity.

Implications for designing feedback that is responsive to the affect state

The results show that certain types of feedback are more effective than others based on the student's affective state. This was particularly noticeable when students were enjoying their activity or when they were frustrated.

When students were enjoying their activity their motivation was high and they reacted to feedback posi-



D5.2 Report on formative evaluation results in Y2

tively across the different feedback types, except for affect boosts. This might imply that if students are interrupted with feedback that does not relate directly to their particular learning goal then this feedback might be ignored.

When students are frustrated then their motivation is low and their cognitive load might be increased. The provision of problem solving feedback was not very effective and students did not follow the advice given. Here the cognitive load might have been too high for the student to follow the problem solving advice. In contrast if the feedback enables the student to talk about their frustration then this might reduce the cognitive load and might enhance the affective state of the student.

The difference of the effect of the feedback types on students' reactions in other affective states (such as confusion or boredom) was not that high. Here students seem to welcome the support provided and followed the advice.




























D5.2 Report on formative evaluation results in Y2

Appendix 7

iTalk2Learn: Student experience questionnaire (VPS study)

Name: Student ID:

Was Maths-Whizz fun?	 Not fun at all	 Not much fun	 It was OK	 A little bit	 Great fun
Were the exercises repetitive?	 Very repetitive	 Repetitive	 It was OK	 A little bit	 Not at all!
Were the exercises easy?	 Very difficult	 A little bit difficult	 They were OK	 Easy	 Very easy!
Was the system helpful?	 Never	 Not much	 It was OK	 A little bit	 Very helpful.
Was the system easy to understand?	 Very difficult	 A little bit difficult	 They were OK	 Easy	 Very easy!



D5.2 Report on formative evaluation results in Y2

	Not at all	No	It was OK	Yes	Very easy!
--	------------	----	-----------	-----	------------


























Any other comments:



D5.2 Report on formative evaluation results in Y2











iTalk2Learn: Student experience questionnaire (WoZ study)

Please read each question and tick the face on the same row that is the best answer.

Now that you have finished the session, how do you feel?	 Very unhappy	 Unhappy	 OK	 Happy	 Very happy
How much fun was Fractions Lab?	 Not fun at all	 Not much fun	 It was OK	 A little bit	 Great fun
How helpful was Fractions Lab?	 Not helpful at all	 Not much	 It was OK	 A little bit helpful	 Very helpful
What did you think of the feedback (the messages shown on the screen)?	 Not very useful	 Not much	 It was OK	 A little bit useful	 Very useful
Was the feedback easy to understand?	 Not at all	 No	 It was OK	 Yes	 Very easy!



D5.2 Report on formative evaluation results in Y2

Was the feedback helpful?	 Not at all	 No	 It was OK	 Yes	 Very helpful!
How much did the feedback get in your way?	 It was always in my way	 It was a little bit in my way	 It was OK	 It was not much in my way	 It was never in my way

If you have any other comments about Fractions Lab (for example, what you liked and what you didn't like), please write them on the other side of this paper.






































D5.2 Report on formative evaluation results in Y2

iTalk2Learn: Student experience questionnaire (WoZ study in Germany)






Name _____ Klasse _____ Schule _____

In dieser Tabelle findest du einige Fragen dazu, wie sehr dir das Programm geholfen und gefallen hat. Kreuze bitte bei jeder Frage den Smiley an, der für dich persönlich am besten zutrifft.

Hat dir die Übung gefallen?	ja, sehr      nein, gar nicht
Hat es dir gefallen, dass du mit dem Computer sprechen konntest?	ja, sehr      nein, gar nicht
Fandest du die Hilfen, die du vom Programm bekommen hast, hilfreich?	ja, sehr      nein, gar nicht
Konntest du die Hilfen gut verstehen?	ja, sehr      nein, gar nicht
Hast du die Hilfen immer gelesen?	ja, sehr      nein, gar nicht
Findest du, dass du oft die gleichen Hilfen bekommen hast?	ja, sehr      nein, gar nicht
Hast du die Aufgaben immer gut verstanden?	ja, sehr      nein, gar nicht



D5.2 Report on formative evaluation results in Y2

<p>Hast du dich in dem Lernprogramm gut zurecht gefunden?</p>	<p>ja, sehr      nein, gar nicht</p>
---	--