iTalk2Learn
2015-10-31

Deliverable 5.3

Report on Summative Evaluation

31 October 2015

Project acronym: iTalk2Learn

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

Work Package: **5**

Document title: D5.3-report_on_summative_evaluation

Version: **1.0**

Official delivery date: **31 October 2015**

Actual publication date: 31 October 2015

Type of document: Report

Nature: Public

Authors: Michael Wiedmann (RUB), Claudia Mazziotti (RUB), Nikol Rummel (RUB), Wayne Holmes (IOE), Alice Hansen (IOE), Manolis Mavrikis (IOE), Carlotta Schatten (UHi), Beate Grawemeyer, (BBK), Gerhard Backfried (Sail)

Reviewers: Sergio Gutierrez-Santos (BBK), Carlotta Schatten (UHi), Lars Schmidt-Thieme (UHi)

Version	Date	Sections Affected
0.1	05/09/2015	Draft of introduction and method section
0.2	25/09/2015	Revision of introduction
0.3	02/10/2015	Draft of result section
0.4	16/10/2015	Included sections on formative trials from UHi, BBK, IOE, and Sail. Revised section on risk mitigation.
0.5	26/10/2015	Review comments from BBK and IOE processed
1.0	28/10/2015	Final version

Executive Summary

The iTalk2Learn project aims to facilitate robust learning in elementary education by creating a platform for intelligent support that combines structured tasks from existing tutoring environments (Math-Whizz and Fractions Tutor) with exploratory tasks from a newly developed learning environment (Fractions Lab), and that provides an interface for voice interaction. The summative evaluation tested how well the project achieved these aims. This deliverable reports on the formative evaluation activities conducted in Year 3 in final preparation for the summative evaluation, and presents the results of the summative evaluation conducted both in Germany and the United Kingdom.

Intelligent learning environments present a unique opportunity to create scalable solutions to educational challenges. To be effective, they require careful development and fine-tuning. The formative evaluation trials in Year 3 focused on ensuring that the iTalk2Learn platform was ready for the summative evaluation. Trials exploited data from Year 2 and new data which was collected with 52 students in the United Kingdom and Germany. Results showed that the individual components of the learning platform achieved their purpose and could be optimized with slight adaptations. After integrating individual components into the learning platform, extensive testing identified where the complexity of the system required further fine-tuning. After this iterative cycle of testing and bug fixing, the integrated platform was piloted in classroom settings with 129 students in Germany and the United Kingdom to ensure the readiness of the platform in the ecological context of the summative evaluation. The experiences collected in the pilots fed into a final round of platform modifications before the start of the summative evaluation.

The promise of educational technologies requires careful evaluation. The iTalk2Learn project investigated two hypotheses on how an intelligent learning environment can foster robust learning. These hypotheses focussed on central innovations of the iTalk2learn platform: the combination of different learning environments providing exploratory or structured tasks for conceptual and procedural learning, and the use of speech to tailor support more closely to young learners whose written-language interaction capabilities are still limited. The summative evaluation tested these hypotheses in a quasi-experimental design in two educational contexts (United Kingdom and Germany), using two different tutoring environments for structured learning (Maths-Whizz and Fractions Tutor). In the United Kingdom, a total of 184 students from three schools, and in Germany, a total of 233 students from six schools participated in the summative evaluation.

The results of the summative evaluation clearly demonstrate that the combination of structured and exploratory tasks promotes learning more than structured tasks alone. This result was replicated in both educational contexts which underlines the effectiveness of iTalk2Learn to foster robust fractions knowledge in students. The results also showed that the role of speech is more complex than previously thought: in the United Kingdom, the version with speech adaptivity fostered learning more than the version without speech adaptivity, albeit not significantly so. In Germany, the version without speech adaptivity fostered learning more than the version with speech adaptivity. One possible explanation is that the benefits of speech depend on the prior knowledge of students: they may be more pronounced

for students with low prior knowledge. This and other explanations will be investigated in further analyses.

The iTalk2Learn project has demonstrated how educational technologies can meet educational challenges through deep analyses of learning content and student misconceptions, careful design of pedagogical interventions, and collaboration with stakeholders in the design of the iTalk2Learn platform. The summative evaluation has shown the efficacy of the platform for fostering robust knowledge of fractions. The collected data also allow the investigation of many additional research questions, for example how accurate affect detection was and how this relates to learning, or the role of representations in shaping students' thinking-in-change. The platform can support mathematics instruction in classrooms and serve as a testbed for future technological and theoretical developments.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
ELE	Exploratory Learning Environment
FL	Fractions Lab
FT	Fractions Tutor
ITS	Intelligent Tutoring System
M	Month
MF	Matrix Factorization

PTDC	Perceived Task Difficulty Classifier
SNA	Student Needs Analysis
TDS	Task-dependent support
TIS	Task-independent support
VPS	Vygotsky Policy Sequencer
WoZ	Wizard of Oz
Y	Year

1. Introduction

Students often struggle with learning fractions and the richness fractions afford with respect to different representations and interpretations (Charalambous & Pitta-Pantazi, 2007; Martin et al., 2015). Perhaps because of these challenges in learning, Siegler et al. (2012) found that elementary students' knowledge of fractions and division at 10 years of age is a uniquely accurate predictor of their attainment in algebra and overall maths performance five or six years later. It therefore seems vital to support students in mastering these challenges and acquiring robust knowledge of fractions. Robust knowledge is knowledge that is retained over the long-term and that transfers from the learning situation to other situations that differ from the learning situation (Koedinger, Corbett, & Perfetti, 2012). For knowledge to be robust, it requires a combination of two types of knowledge, procedural knowledge and conceptual knowledge (e.g., Rittle-Johnson, Siegler, & Alibali, 2001; for our working definitions of these terms, see D1.1). Educational technology has been shown to foster procedural knowledge (with intelligent tutoring systems, ITSs, e.g. Koedinger & Corbett, 2006; VanLehn, 2006). There are also learning environments that focus on supporting conceptual knowledge acquisition (exploratory learning environments, ELEs, e.g. Mavrikis, Gutiérrez-Santos, Geraniou, & Noss, 2013; Noss et al., 2012). But how can educational technology foster both types of knowledge and therefore support learners in acquiring robust knowledge of fractions? Moreover, fractions are typically taught early in the school curriculum. Young learners have often not yet mastered basic literacy skills, so their ability to interact with educational technology based on written language is limited. How can educational technology provide a more accessible interface for these learners?

The iTalk2Learn project has investigated these questions and developed an adaptive system that supports students in acquiring robust knowledge about fractions. The system combines conceptually-oriented exploratory learning tasks and procedurally-oriented structured practice tasks. Geared towards young learners, it uses state-of-the art speech recognition and production to provide a speech-based interface in addition to traditional written language.

This deliverable reports on the formative and the summative evaluation of the iTalk2Learn platform conducted in Y3 of the project. The next two sections of this introduction briefly discuss the theoretical background of the research questions addressed by the summative evaluation. Section 0 describes the final steps of formative evaluation of the platform that were conducted in Y3 of the project. Section 3 addresses challenges that posed a risk for the summative evaluation, describes how we mitigated them and evaluates what this means for the project vision. Section 4 presents a detailed account of the summative evaluation which investigated the effect of combining exploratory and structured tasks and the effect of speech adaptivity. Section 5 presents a study in which the summative evaluation was extended to gain a more detailed understanding of students' thinking-in-change. Data collected in this study are still being analysed, but a study report in the appendix presents initial findings. Finally, section 6 discusses our findings and provides a glimpse into ongoing and future analyses that can be performed with the rich datasets we collected, as well as future opportunities for continuing research on promoting robust fractions knowledge with ITSs.

1.1 Combining Exploratory and Structured Tasks to Promote Learning

As mentioned above, for knowledge to be robust, it requires a combination of two types of knowledge, procedural knowledge and conceptual knowledge (e.g., Rittle-Johnson et al., 2001; for our working definitions of these terms, see D1.1). Both types of knowledge develop over the same period of time (e.g., LeFevre et al., 2006). They develop iteratively: increases in one type of knowledge lead to gains in the other type of knowledge, which in turn lead to increases in the first type of knowledge (cf. Rittle-Johnson et al., 2001). However, the development of the two types of knowledge are thought to rely on different types of learning activities and therefore require different kinds of instructional support (e.g., in the form of computer-based learning environments; Koedinger et al., 2012).

In this context, two different types of computer-based learning environments have shown great success in fostering mathematics knowledge. ITSs are suited particularly well for the development of procedural knowledge. ITSs offer students efficient instructional support for practicing problem-solving procedures because students solve problems step-by-step, and receive immediate feedback. This way they can automatize the problem-solving procedure bit by bit (e.g., Anderson & Lebiere, 1998). A common criticism of these learning environments is that they focus too much on drill and practice while neglecting conceptual knowledge construction. ELEs, on the other hand, are suited particularly well for the development of conceptual knowledge as students can for example manipulate representations, make their own experiences and discover the underlying concepts. A common criticism of these learning environments is that they do not provide enough guidance to students (Kirschner, Sweller, & Clark, 2006). D1.1 and D3.2 describe how by providing students with exploratory learning tasks and by encouraging reflection and self-explanation, students can be supported to abstract information, construct schemata, and hence develop conceptual knowledge (e.g., Koedinger et al., 2012).

Given these limitations of existing learning environments, iTalk2Learn presents a major innovation. Prior work in the learning sciences and educational technology has focused on fostering *either* procedural knowledge with structured tasks (within ITSs) *or* conceptual knowledge with exploratory tasks (within ELEs). iTalk2Learn is the first learning environment that combines both exploratory and structured tasks for robust learning of fractions. The newly-developed ELE Fractions Lab provides rich opportunities for exploring fractions using a diverse set of representations of fractions. It is paired with Fractions Tutor (in Germany) and Maths-Whizz (in UK), two established ITSs that focus on structured practice. D1.3 presented the pedagogical model which describes how to combine both types of learning tasks and thus specifies how students are supported in acquiring robust fractions knowledge while working with the iTalk2Learn platform.

As will be discussed in Section 3, the summative evaluation investigated whether the combination of exploratory and structured tasks implemented within iTalk2Learn promotes students' learning more than structured tasks alone. Table 1 presents an overview of the platform configuration for experimental conditions. Details are discussed in section 4.1.1 and 4.1.5.

Table 1
Experimental conditions of the summative evaluation

Platform components	Experimental conditions C1 (Full Platform)	C2 (No Speech)	C3 (No ELE)
Exploratory tasks (Fractions Lab, TDS)	+	+	-
Structured tasks (Maths-Whizz/Fractions Tutor)	+	+	+
Speech (TIS, speech production)	+	-	+

1.2 Utilising Speech to Promote Learning

Speech can help to promote learning in at least four interrelated ways: (1) speech provides a natural interface that allows learners who have not mastered written language yet to interact more easily with educational technology, (2) prompting students to verbalize their thinking helps elaborate their knowledge, (3) what students say can be used to infer their cognitive state so they can be provided with cognitive support (e.g., feedback), and (4) what students say and how they say it can be used to infer their affective state so they can be provided with affective support which in turn promotes knowledge acquisition.

In terms of the first two aspects, research in mathematics education and cognitive science highlights the important role that spoken language plays in learning in general and mathematics in particular. For example, translations between the written number symbols and the other modes of representation including spoken symbols help promote learning (Verschaffel & Corte, 1996). This translation process reflects students' representational flexibility, a key component of conceptual knowledge (e.g, Lesh, 1999). Other research (Mercer & Sams, 2006; Rajala, Hilppö, & Lipponen, 2012; Teasley, 1995; Zakin, 2007) has shown that, when students are encouraged to put their thoughts into words and to give self-explanations about the target principle, there are various benefits for learning. In addition, related research suggests that spontaneous self-explanation is more frequent in spoken rather than in typed interactions (e.g., Hausmann & Chi, 2002) that are typical of digital learning systems, and that audio feedback is beneficial both to task performance and to learning (e.g., Fiorella, Vogel-Walcutt, & Schatz, 2012).

In terms of the latter two aspects, speech can promote learning in an indirect way. Inferring cognitive and affective states from students' speech gives additional information about the students' learning progress and thus helps to create a more precise student model (i.e. accurate representation of students' knowledge). This in turn makes the system's adaptive support more effective. Existing systems rely on recording students' actions within the learning environment and analysing text-based interaction to form and update their student model. However, text-based interfaces are hard to use by young children who are still perfecting their reading and writing skills. This means that the input children can provide for student models through these modalities is limited. To the best of our knowledge, the role and effectiveness of voice user interfaces for elementary mathematics learning

has not been investigated yet. In the sibling field of elementary reading skills, results from experiments with LISTEN, one of the few intelligent tutors utilising speech recognition and production targeted at improving reading comprehension, suggest not only that the addition of a natural user interface is plausible but also that it has positive learning gains (Mostow & Aist, 2001).

The iTalk2Learn platform takes advantage of the affordances of speech for learning by utilising innovative technology for speech production and recognition. Students are encouraged to talk to the system because it also talks to them, and speech recognition in turn is used to form a more accurate model of students' cognitive and affective state. With these models, the system can provide not only cognitive, but also affective support. As will be discussed in section 4, the summative evaluation investigated whether this adaptive system fosters fractions knowledge more when utilising speech than a version without speech (also see Table 1).

2. Formative Evaluation

The formative evaluation conducted over Y2 and Y3 followed the layered evaluation approach described by Paramythis, Weibelzahl, and Masthoff (2010). This approach is related to design-based research and is particularly suited to the evaluation of adaptive learning environments since they are necessarily complex. Each trial focused on the contribution of one component. In Y3, the formative trials focused on ensuring that individual platform components were ready for the summative evaluation. Once components were integrated into the platform, a trial also studied the interaction of components. Substantive testing of the integrated platform ensued before we conducted a pilot study for the summative evaluation to evaluate procedures and materials. The summative evaluation described in section 3 presents the culmination of the layered evaluation approach as it evaluates the complete system. This section briefly discusses each of the formative trials conducted in Y3 (for an overview, see Table 2).

Table 2
Overview of formative evaluation trials in Y3

Trial focus	Aim	Data source
Task-independent support (TIS)	Train Bayesian network	26 students from Y2 Wizard-of-Oz trials
Intervention model	Test sequencing and switching	2 students in laboratory study
Fractions Lab	Trial the final exploratory fraction addition and subtraction tasks IOE had designed	50 students in classroom study

Trial focus	Aim	Data source
	Provide final feedback to TL regarding the user interface of Fractions Lab	
Vygotsky Policy Sequencer	Train sequencer for Fractions Tutor	88 students from Y2 classroom study
Speech recognition	Improve speech recognition performance	251 students from Y2 classroom study
Integrated platform	Identify and fix bugs	155 bugs reported by consortium members
UK summative evaluation pilot	ensure the readiness of platform trial summative evaluation procedures and measures	27 students in classroom study
Germany summative evaluation pilot	Test different versions of platform for experimental conditions Stress test local area network installation	102 students in classroom study

2.1 Task-Independent Support

The aim of this first formative evaluation trial was to train Bayesian networks for task-independent support (TIS).

TIS provides support based on mathematical vocabulary and affective states (see D2.2.2). Affect and vocabulary detection and selection of adaptive support are based on Bayesian networks. To train these networks, data were collected in a Wizard of Oz trial at a suburban primary school in Y2 (see D5.2). In Y3, data was analysed from 26 Year-5 (9 to 10-year old) students that had participated in the Y2 trial. In the Y2 trial, wizards followed a script with pre-determined messages to send messages to the students through the learning platform. Any feedback provided was both shown on screen and read aloud by the system to students. Different types of feedback were presented to students at different stages of their learning task. The feedback provided was based on interaction via keyboard and mouse, as well as speech. Screen display and voices were recorded.

In Y3, these recordings were analysed to identify English keywords for mathematical vocabulary and affect detection. Moreover, we annotated affective states (e.g., screen interaction and what the students said) before and after feedback was provided. We also used the HART mobile app

(Ocumpaugh, Baker, & Rodrigo, 2012) that facilitates the coding of students affective states in the classroom with the BROMP protocol (Ocumpaugh et al., 2015). This data then was used to train the Bayesian networks developed for TIS. More details about the training data collected and how these were used for TIS can be found in Mavrikis, Grawemeyer, Hansen, and Gutiérrez-Santos (2014), Grawemeyer, Mavrikis, Hansen, Mazziotti, and Gutiérrez-Santos (2014), Grawemeyer, Holmes et al. (2015), Grawemeyer, Mavrikis, Holmes, and Gutiérrez-Santos (2015), Grawemeyer, Mavrikis, Holmes, Hansen, Loibl, and Gutiérrez-Santos (2015b), and Grawemeyer, Mavrikis, Holmes, Hansen, Loibl, and Gutiérrez-Santos (2015a).

2.2 Intervention Model

The aim of this second formative evaluation trial was to test sequencing and switching between tasks.

Sequencing and switching (and when to provide adaptive support) are specified by the intervention model described in D1.3. The intervention model prescribes that tasks are sequenced based on each student's level of challenge – whether he is under-, over- or appropriately challenged by the current learning task. For the trial, RUB identified tasks that are typically more or less challenging for students based on the knowledge they require. The intervention model then could be formalized with IF-THEN rules: If the student is under-challenged with the first exploratory learning task then she will receive next a specific exploratory learning task that is typically more challenging. If it turns out that she is over-challenged with the first exploratory task, for example because she asks for a lot of TDS, she will receive a less challenging exploratory learning task next. Finally, if she is appropriately challenged with the first exploratory task she will receive a structured practice task that maps the to-be-practiced procedure to the previously explored concept.

In this formative trial, the selection of the next learning task was implemented in a Wizard-of-Oz setting. One of two experimenters took the role of the Wizard by identifying the student's level of challenge and selecting the learner-appropriate next task. The other experimenter took the role of the more distanced observer. Two students with different prior knowledge about fractions agreed to participate in the study. First, students were introduced to the basic functionalities of the system and were provided with a familiarization task. Students then learned for 30 minutes with the platform. Then, students were asked to complete a short knowledge test and a user experience questionnaire. Students were interviewed for suggestions how to improve the iTalk2Learn system and for identifying their prior experiences with learning with computers.

Results supported the predictions of the intervention model. One girl who had not yet been formally introduced into fractions highlighted that the learning tasks fitted to her needs very well and that she felt she could learn at her own pace. Another girl with very high prior knowledge repeatedly felt under-challenged by the tasks. Because she did not receive enough tasks that challenged her appropriately, she did not learn as much as could have been expected. Regarding user experience, students liked learning with the iTalk2Learn platform and were particularly excited about learning with different representations within the ELE.

In conclusion, this formative evaluation trial could show that also with an early version of the intervention model and in a Wizard-of-Oz setting it is possible to adapt to individual students' needs and to provide them with a unique sequence of learning tasks. In considering the children's comments, we included more challenging tasks and re-worded some of the tasks to aid clarity. This helped us to develop a more adaptive and intuitive learning platform.

2.3 Fractions Lab

The aim of this third formative trial was twofold: 1) to trial the final exploratory fraction addition and subtraction tasks IOE had designed, and 2) to provide final feedback to TL regarding the user interface of Fractions Lab (see D3.4.2).

During M29, IOE conducted this trial over two days with 50 Year-6 students at a suburban school. Small groups of students (approximately 10 at a time) worked through a series of paper-based tasks using Fractions Lab for a duration of approximately one hour in the school computer lab. The students completed questions related to the tasks on worksheets. This gave us an insight into how the students were interpreting the tasks and the level of challenge each task provided. At the conclusion of each session the students formed a focus group to discuss the tasks and the features of Fractions Lab. Because these students had previously used an earlier version of Fractions Lab when they were in Year 5, we explained to them how their earlier feedback had been taken into consideration and what the changes to the new Fractions Lab was as a result of their feedback. In this role they were co-designers and were asked to comment on the changes and newer features. As a result of the trials, some of the task descriptions were re-worded to aid clarity. Enhancements to the Fractions Lab user interface were also carried out, for example:

- the logo on the add/subtract boxes was re-designed to include + and -
- Fractions where the equivalence tool arrow is showing can now be dragged into the equivalence/add/subtract boxes, not only fractions that have a numerator / denominator arrow showing as was the case before
- 'Near-miss' drag and drop on representations was enabled

The students also raised some suggestions that were not implemented, for example:

- Include a 'clear all' option when you click on the bin because "it is tedious to drag them all in" [this was of low importance and low priority and was not implemented due to time and resourcing constraints]
- Make it possible to hold an arrow down so the fraction can change automatically without having to click every time a change needed to be made? [the touch and hold action on a tablet brings up a menu and therefore was not implemented to ensure compatibility between the different Fractions Lab versions]

2.4 Vygotsky Policy Sequencer

The aim of this fourth formative trial was to train the Vygotsky Policy Sequencer (VPS) for Fractions Tutor. UHi then conducted an evaluation in a laboratory study.

In preparation, our main concern was the so-called cold start problem. Cold-start problems are experienced in the performance prediction model underlying the VPS (matrix factorization, MF) when not enough data on tasks or on students are available. In Maths-Whizz, the cold start problem could be ignored because enough data on tasks was available and students with past history in the system could be selected for the formative evaluation in Y2. However, for Fractions Tutor (FT) the consequences of this prediction problem had to be evaluated because the FT data collected for training consisted of students that had interacted with a small number of tasks which were newly developed during the project. UHi in collaboration with RUB reduced the effect of the cold start problem with a practical and a theoretical approach. As described in D2.2.2, we designed and evaluated results of a novel data collection approach that allowed collecting enough data for the tasks although not enough time was available for the students to practice on the whole sequence, which is the classic approach. The novel approach used here consisted in collecting data in three different sequences, so that approximately the same amount of interactions was collected for all tasks. The formative evaluation of the VPS applied to FT consisted firstly in assessing the novel collected data, which meant evaluating the effect of the cold start problem both on tasks and on students.

Our analyses showed that at least six interactions of a student are needed to be able to use MF prediction, since with less interactions vanilla or rule-based classifiers perform better than our MF implementation. To further alleviate this problem, we implemented a novel approach to reduce the task cold start problem, which consisted in applying for the first time Transfer Learning to Educational Data Mining. We successfully combined iTalk2Learn data and data collected with a culturally adapted previous version of the same system to ameliorate the performance prediction. In contrast to classical Machine Learning methods, Transfer Learning exploits the knowledge accumulated from auxiliary data to facilitate predictive modelling with the use of different but similar patterns. This information is generally discarded in the common Machine Learning approach. Results were published in Voß, Schatten, Mazziotti, and Schmidt-Thieme (2015). For more information please see D2.2.2 and Voß et al. (2015).

2.5 Speech Recognition

The aim of this fifth formative trial was to improve speech recognition performance. Evaluation regarding the automatic speech recognition (ASR) component (and its associated models) in Y3 focused on its intended use within the iTalk2Learn platform as deployed for the summative evaluation.

In the platform, ASR is applied in the context of TIS, detecting relevant, domain-specific terminology and providing clues about a student's cognitive state. This task – the spotting of vocabulary and affect terms – can thus be viewed in terms of precision and recall, two widely used measures in the field of

Information Retrieval. These measures were consequently selected as the most appropriate measures for offline (formative) evaluation. They are expected to provide a realistic picture of performance which translates well into terms of online performance as observed in the deployed system (for summative evaluation).

During Y3, SAIL successively trained and evaluated four models, improving precision and recall in every iteration (see D3.3.2). For the final model, the f-measure (the harmonic mean of precision and recall) was, for German, 47% for affect words and 52% for math terminology, and, for English, 46% for affect words and 35% for math terminology. The model was implemented in the speech recognition module of the iTalk2Learn platform and tested with students in the summative evaluation pilots (see sections 2.7 and 2.8). Speech recognition proved useful in affect and vocabulary detection, so that we consider its performance a success for the present purposes.

2.6 Integrated Platform

The aim of this sixth formative trial was to identify and fix bugs of the integrated platform.

Once components had been integrated, a series of tests were conducted over the course of three months to ensure that integration was successful. This concerted consortium effort was led by BBK, RUB, and UHi. A schedule was set so that each test could focus on one component in one week and to provide sufficient time for bug fixing the following week. All consortium members participated in the biweekly testing sessions and recruited members of their organization who were not consortium members to participate as well. Bugs were reported directly to developers using GitHub and Google Forms. A smaller team convened in weekly integration meetings to discuss the progress on the platform and coordinate among consortium members. One stress test was also conducted locally during the General Meeting in Bochum in March which also allowed to test the platform within a local area network. Another work meeting took place in London in June to allow for face-to-face discussions and a concerted effort in finishing the platform for the summative evaluation. Through this rigorous testing process, we were able to identify and eliminate 155 bugs that were reported by testers. The testing was also able to assess risks identified in D5.2 and deploy appropriate counter measures. This is reported in more detail in section 3.1.

2.7 UK Summative Evaluation Pilot Study

The aims of this seventh formative trial were to stress-test the technology (i.e. to ensure the platform's readiness for deployment to schools), the research methods and research procedure, and to trial and confirm the summative evaluation's online and paper measures.

IOE undertook this ecologically valid pilot study in one of the schools to be used in the summative evaluation. To avoid contaminating the summative evaluation sample, the pilot study involved only children and their class teacher from Year 6 (whereas the summative evaluation involved children and their teachers from Years 4 and 5). A document outlining the pilot study was provided to the parents/carers of all the children in two Year 6 classes. Only those children whose parents or caregivers signed and returned a form giving their consent were included in the study (N = 27).

During the pilot study, which took place in the school's computer lab, all the children worked on individual computers to engage with the full iTalk2Learn platform (incorporating exploratory learning, structured practice and speech functionality). For a full description of the platform, how it was configured, what it involved, and how the children were asked to interact with it (the research study procedure), please see the details given in the description of the summative evaluation below.

The pilot study enabled us to check, refine and confirm a wide range of research procedure variables, for example:

(i) to confirm the individual computer configuration needed to run the iTalk2Learn platform in a typical school computer lab;

(ii) to confirm how the project's server could be connected into the school's network so that it might be accessed by the individual computers (i.e. to identify and resolve technical problems that might be encountered in other schools);

(iii) to confirm the efficacy of the project's equipment (i.e. headphones and microphones);

(iv) to trial and confirm the research procedures (e.g., the script used to introduce the system to the participating children, the protocol for the researchers, and the overall timings necessary to fulfil the project's requirement while working within school timetable constraints);

(v) to identify and mitigate any risks; and

(vi) to trial, refine and confirm the various paper-based and online measures.

The trial also provided additional data to further train the speech recognition functionality.

2.8 German Summative Evaluation Pilot Study

The aims of this eighth formative trial were to test different versions of the platform for the experimental conditions of the summative evaluation and to stress test the local area network installation. In addition, RUB conducted this study with the same goals and results of the UK pilot for the summative evaluation just described in section 2.7, so only notable differences are discussed here.

Firstly, RUB tested the German platform in a local area network with laptops provided by RUB because schools in the German summative evaluation could not offer adequate computer lab facilities (see also section 3). Moreover, not only the full platform, but also the individual platform versions of the conditions of the summative evaluation were tested. Finally, this pilot tested the platform with thirty students working simultaneously. The study was conducted with two fifth grade classes and two sixth grade classes of a school in a rural setting. Informed consent was obtained for N = 102 students who then participated in the study. The study revealed room for improving the speed in loading tasks which prompted BBK to redesign the way the platform used system resources. This is discussed in more detail in section 3.

3. Risk Mitigation and Consequences for Project Vision

This section discusses the risks in finalizing the platform and conducting the summative evaluation. It reflects how these risks impacted the project vision and describes how the project vision was realized in terms of the summative evaluation and beyond.

3.1 Platform Risks

D5.2 identified several risks for the summative evaluation that concerned the readiness of the iTalk2Learn platform. During the formative evaluation, we continuously evaluated these risks and were able to identify new risks that could be successfully mitigated prior to the summative evaluation. Table 3 presents risks by status of components at the end of Y2, indicates whether risks could be mitigated, and shows the status of the component in Y3 for the summative evaluation.

Table 3
Risk mitigation concerning the iTalk2Learn platform

Component	Status Y2	Risks mitigated?	Status Y3
Fractions Lab	Working. Based on a version of Unity Player that will not continue to be supported by major browsers	Yes	Working. For the summative evaluation, versions of Firefox were used that support this version of Unity Player.
Word recognition	First models trained Error rate of word detection unsatisfactory	Yes	Model optimized to recognize specific relevant words. Recognition provides useful input to TIS
Learning tasks	Limited amount available	Yes	Made enough content available to fully utilize learning time
Perceived Task Difficulty Classifier (PTDC)	First model trained Error rate of affect classification unsatisfactory	Yes	PTDC provides useful input to TIS
Vygotsky Policy Sequencer for Maths-Whizz/ Fractions Tutor	Working/ Training ongoing	Yes	Working, but requires interaction data from six tasks. Because these are not available in the summative evaluation, VPS is turned off for summative evaluation

Component	Status Y2	Risks mitigated?	Status Y3
Task-dependent support (TDS)	Manual delivery working Automatic delivery for one task implemented	Yes	TDS is provided automatically for all tasks implemented in the summative evaluation and one task in the follow-up study
Task-independent support (TIS)	Manual delivery	Yes	TIS includes speech-based indicators
Switching	Manual switching	Yes	Student Needs Analysis (SNA) switches automatically after every two tasks to ensure hypothesis 1 can be tested
Sequencing within ELE	Manual sequencing	Yes	The SNA provides automatic and adaptive sequencing
Integration	First tests successful	Yes	All components integrated. Platform was redesigned to better take advantage of system resources

As can be seen from the table, all risks that had been identified were prevented by realizing the necessary efforts in time for the summative evaluation. These are discussed now in more detail for Fractions Lab, sequencing, switching and integration. For these components, additional mitigation measures had to be taken, to account for constraints of the summative evaluation as is explained next.

Regarding Fractions Lab, major browser organizations will cease NPAPI support in the near future (e.g., Schuh, 2013). NPAPI is required for the version of Unity Player on which Fractions Lab is based. To ensure platform cross-compatibility, and given the limited time left in the project, we decided to stick with the NPAPI-based Unity Player for the summative evaluation. In schools, we used browser versions that do support NPAPI. TL has since created a WebGL-based version of Fractions Lab which does not require NPAPI support and which has been made publicly available (http://fractionslab.lkl.ac.uk/; see D6.3.3)

Regarding sequencing, while the VPS had been shown to produce substantive learning gains, the formative evaluation of the VPS described in section 2.4 also showed that the VPS requires interaction data from at least six tasks for a meaningful prediction. Within the constraints of the summative evaluation, specifically the lack of prior interaction data for students, the short learning time and the limited number of tasks available for sequencing would have incapacitated the VPS. Despite further adaptations to obviate this problem, reported in D2.2.2, we decided to use a pre-fabricated sequence of structured tasks during the summative evaluation. This allowed us to focus on adaptive sequencing of exploratory tasks by the student needs analysis (SNA).

Regarding switching, the SNA was intended to only switch from exploratory to structured tasks and vice versa when detecting a specific level of challenge for the student. Within the constraints of the summative evaluation, specifically the short learning time and the limited number of tasks available, the number of times students are switched theoretically could have varied from zero to ten. This variance posed a significant risk for investigating one of the central hypotheses of the summative evaluation: that the combination of exploratory and structured tasks promotes learning more than structured tasks alone. We therefore decided to switch from exploratory to structured tasks and vice versa every time a student had completed two tasks, regardless of the level of challenge detected by the SNA. The SNA did automatically tasks within the ELE depending on the detected level of challenge, as intended.

Regarding integration, testing of the platform revealed that its performance was compromised when more than fifteen students worked with it at the same time (see section 2.8). Within the constraints of the summative evaluation, specifically that class sizes in Germany are usually around 30 students, this would have doubled the sessions required to conduct the summative evaluation and likely reduced the number of participating schools due to the higher commitment of overall classroom time. In response, BBK was able to dedicate significant efforts to a redesign of the way the platform used the resources of the server that hosted the platform. We also purchased two powerful servers to ensure high availability of system resources. This had the additional advantage that evaluations could run in Germany and the UK at the same time. Both measures together enabled us to conduct sessions with up to thirty students per country at the same time.

3.2 Summative Evaluation Risks

In D5.2, we also identified a number of risks that could affect the summative evaluation independently of the readiness of the iTalk2Learn platform. Table 4 presents risks at the end of Y2, indicates whether risks could be mitigated, and shows the status of the component in Y3 for the summative evaluation.

Table 4
Risk mitigation concerning the summative evaluation

Risk	Status Y2	Risk mitigated?	Status Y3
Low number of schools and students volunteer	Recruitment efforts have started Experimental plan is designed with optional conditions	Yes	Experimental plan reduced to three conditions
Audio recording quality is low	Guidelines for audio quality are specified in writing	Yes	Audio quality sufficient for word and affect detection

Risk	Status Y2	Risk mitigated?	Status Y3
Schools lack computer labs	Laptops are reserved (Germany)	Yes	In UK, schools' computer labs could be used In Germany, a local area network with 30 laptops was set up at each school
Internet bandwidth is limited	Implementation of iTalk2Learn in a local area network has been successfully tested	Yes	Platform works without internet connection in local area network thanks to new offline speech production

As can be seen from Table 4, all risks have been successfully mitigated. Regarding participation in the summative evaluation, the overall number of schools was higher than anticipated (3 schools in UK and 6 schools in Germany). However, the number of classes from each school was low. For example, in Germany, in most schools only one class participated and three classes at the most. Had these participants been assigned to four conditions, statistical power would not have been high enough for reliable inferential statistics testing. We therefore dropped one experimental condition, reducing the experimental plan to three conditions.

Regarding computer labs and internet bandwidth, none of the schools in Germany was able to provide adequate facilities. In response, RUB hired additional student research assistants to set up local area networks with laptops at the schools. To prepare for the case that internet bandwidth may not be reliable enough, or in some cases even unavailable, BBK and RUB integrated and tested a speech production solution that works offline. This completely eliminated the need for internet access and increased reliability of the platform for the summative evaluation.

3.3 Reflection on Project Vision

How does the platform as it was implemented in the summative evaluation measure up against the project vision? We have successfully created a learning environment applicable in every day teaching, fully achieving the vision of the iTalk2Learn project. The platform adapts learning content to students' needs and provides intelligent support to help students master the learning content. This capability is based on input about the learner from a variety of sources. Importantly, not only traditional text-based sources and screen/mouse actions are used: the system also takes into account students' speech. The platform combines exploratory with structured tasks. This combination is based on a pedagogically-sound intervention model that applies research from math education and the learning sciences. Exploratory tasks are provided by Fractions Lab, a novel learning environment developed within the project that provides rich opportunities for learning through manipulating a diverse range of graphical representations.

The realization of the project vision regarding the testing of all technological components in the summative evaluation was limited to some extent because the VPS could not be tested in the summative evaluation. The external constraints of the summative evaluation did not allow for meaningful performance prediction. However, the extensive testing in the large-scale Maths-Whizz Y2 study has shown that the VPS works and provides significant advances in the state of the art of applying machine learning models to performance prediction. The VPS is integrated in the platform and can be used in future studies. Moreover, this does not mean that in the summative evaluation, the platform was not adaptive. Nor does it mean there was no adaptation of structured tasks. The SNA provided sequencing within the exploratory learning environment and adapted the structured tasks to match the previously completed exploratory task. In this sense, the project vision of building and evaluating an adaptive system was fully realized.

The platform over-fulfils the project vision regarding the use of speech for adaptivity. TIS exploits the first speech recognition model specifically tailored to children's speech. The affective state detector of TIS cross-validates the output from the speech recognition model with user's screen/mouse action, and the Perceived Task Difficulty Classifier (PTDC), a machine-learning model that classifies affect based on prosodic cues from speech. The affective state reasoner decides what type of feedback is most likely to improve students' affect and the affective state presentation model decides how the feedback should be provided. Each of these components of affective support implements Bayesian networks.

What are consequences for an "updated" version of the project vision, that is, how the platform could look like in the future? The iTalk2Learn platform already presents a significant innovation for its combination of exploratory and structured tasks, the use of speech for interaction and adaptation to learners' needs, and additional intelligent cognitive and affective support. Combining robust learning (consisting of mixing structured and exploratory tasks) with intensified adaptivity (e.g., using VPS performance prediction for sequencing tasks and switching) will increase the learning gains further, as the latter leverages possibilities of the former. Another idea is enabling students to learn collaboratively with the platform. Collaborative learning might further support students' exploratory behaviour and hence additionally support students' conceptual knowledge development. A collaborative extension to iTalk2Learn, similarly to the extension for Cognitive Tutor Algebra (Diziol, Walker, Rummel, & Koedinger, 2010; Rummel, Mullins, & Spada, 2012), could help investigate this question further. Additional directions for future development are discussed in section 6.1 based on the results of the summative evaluation. The summative evaluation is described next.

4. Summative Evaluation: UK and Germany Trials

The main focus of WP5 in Y3 was on the summative evaluation. As discussed in D5.2, the summative evaluation focused on the following main hypotheses:

H1) Combining structured practice and exploratory tasks promotes robust learning (referred to later as the *combination effect*).

H2) An adaptive system that interacts with learners through speech enhances learning more than an adaptive system that does not (referred to later as the *speech effect*).

The following sections describe the methodology and results of the summative evaluation.

4.1 Methods

4.1.1 Experimental design.

In order to investigate the two hypotheses, we compared three experimental conditions (C1 to C3), each one implementing a different version of the iTalk2Learn platform. The three conditions were specified as follows (see also Table 1 presented in section 1):

C1) The full iTalk2Learn platform (incorporating exploratory learning, structured practice and speech functionality).

C2) The iTalk2Learn platform without speech functionality (but incorporating exploratory learning and structured practice).

C3) The iTalk2Learn platform without exploratory learning (but incorporating structured practice and speech functionality).

To address hypothesis H1 (the combination effect), C1 (Full Platform) together with C2 (No Speech) were compared to C3 (No ELE). While to address hypothesis H2 (speech effect), C1 was compared to C2. These two comparisons provide statistically independent tests of the two hypotheses. Due to the observable differences between conditions (use of speech and different learning tasks), it was not feasible to run multiple conditions in the same classroom. Therefore, the study was run in a quasi-experimental design.

4.1.2 Participants.

In the UK, at the time of the study, fractions were typically taught towards the end of Year 5 of primary school. Therefore, participants in the UK were Year 4 and Year 5 primary school students aged between 8 and 10 years old from three schools. The three schools were from a rural, suburban, and inner-city area. Parental consent, for their involvement in the study, was obtained for all participating students. Seven students did not complete the study. Participating students were roughly stratified, according to previous teacher assessments of the children's mathematical ability, in three groups per

year group per school which were then randomly assigned to one of the conditions, resulting in the following distribution across conditions: N_{C1} = 61, N_{C2} = 56, and N_{C3} = 60.

In Germany, at the time of the study, fractions were typically taught at the beginning of sixth grade in secondary school. Therefore, participants in Germany were fifth and sixth grade secondary school students aged between 10 and 12 years old from five schools. One school was from a rural area while the four other schools were from a suburban area. Parental consent, for their involvement in the study, was obtained for all participating students. Four students did not complete the study. Participating students could not be stratified due to constraints of the participating schools, so students participated within their class, and classes within schools were randomly assigned to one of the conditions. Class sizes varied, and, due to a technical failure, data was lost for one class of 33 students (assigned to condition 3), resulting in the following distribution across conditions: N_{C1} = 100, N_{C2} = 59, and N_{C3} = 51.

4.1.3 Instruments.

Participants completed several instruments during the study. Fractions problems assessed procedural and conceptual knowledge. A questionnaire assessed attitudes to learning, mathematics and fractions. A second questionnaire assessed students' experience using the platform. A pop-up window within the iTalk2Learn platform asked students to report on the task they had just completed. For a subsample of participants, while they worked with the platform observers assessed their affect. These instruments are now described in more detail. In addition, all student interaction with the platform, including speech, was recorded.

Paper-based fractions problems. Two paper-based problems were presented to students. The first problem assessed representational variety of fractions knowledge. For this, the students were given a single sheet of A4 paper with the words 'one third' written in the centre. They were asked to write or draw on the sheet as many different versions or equivalences of one third as possible. The second problem assessed skill in using different fractions representations. For this, the students were given the sheet of A4 paper displayed in Figure 1. They were told that another student, called Amelie, had represented one third in five different ways (as a numerical symbol, as a number line, as a set, as a liquid measure and as a rectangle), and they were asked to represent two sixths in the same ways as Amelie.

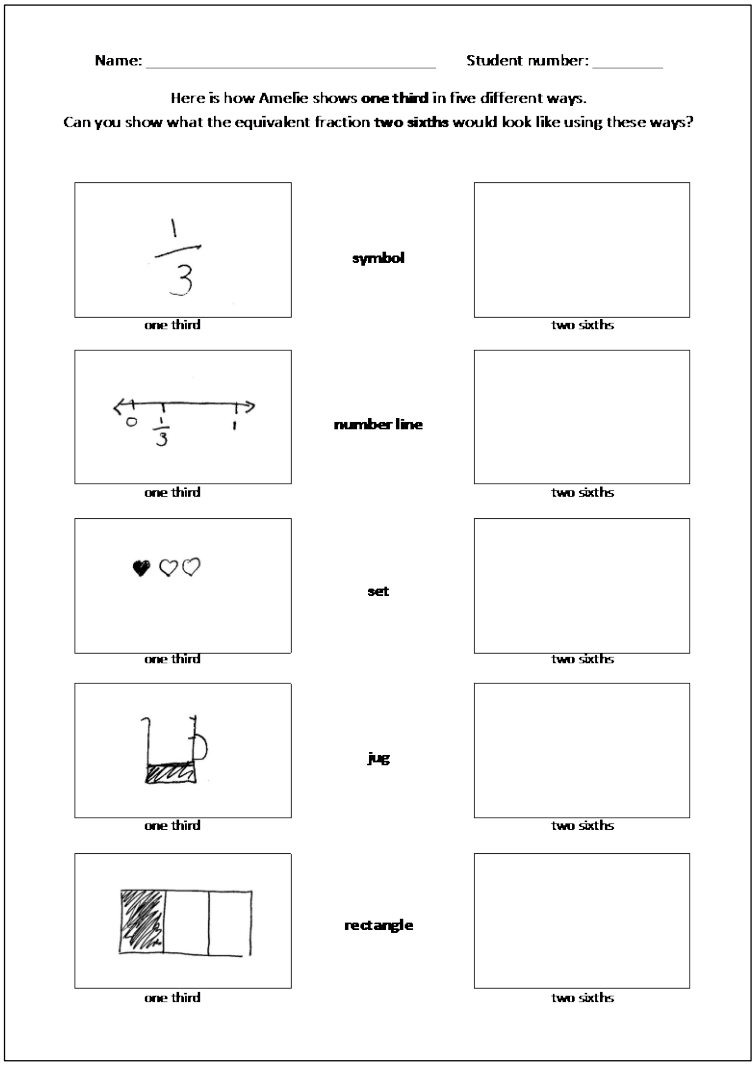


Figure 1. Equivalent fractions using different representations

Attitudes to learning, mathematics and fractions. Students responded to the following items on a scale from 1 (I do not agree at all) to 5 (I totally agree): Learning with a computer is fun, I have already used a learning program/platform, I often work with a computer, I like maths, I like fractions.

Online fractions problems. Two isomorphic versions of six fractions problems were designed. Students were randomly allocated one version at the first time of measurement and the other version at the second time of measurement. Three problems emphasised a procedural approach (see questions 22, 24, and 25 in Figure 2) and three emphasised a conceptual approach to understanding or calculating with fractions (see questions 20, 21, and 23 in Figure 2). The students received one point for each correctly answered problem and consequently obtained an aggregated score that we used as an overall measure of fractions knowledge. Internal consistency of this scale at pre-test was $\alpha_{UK} = .58$, $\alpha_{Germany} = .38$, and at post-test $\alpha_{UK} = .53$, $\alpha_{Germany} = 54$.

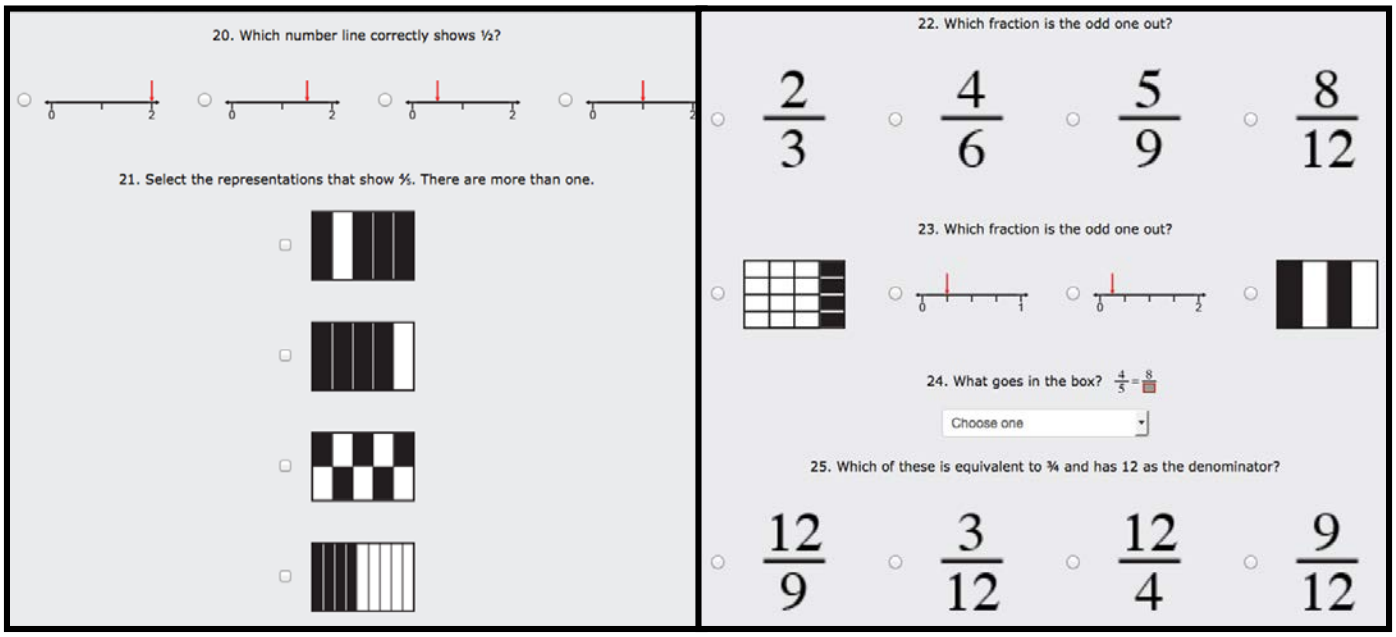


Figure 2. Online fractions problems.

User experience questionnaire. Students responded to the items presented in Table 5 on a scale from 1 (I do not agree at all) to 5 (I totally agree).

Table 5
User experience questionnaire

Item
I would like to work with the platform again
The platform was easy to use
I like speaking with the platform
The platform understood what I was saying
I always paid attention to the hints
The male robot's hints were helpful
The female robot's hints were helpful
The tasks were too hard
The tasks were too easy
After working at this platform for a while, I felt pretty competent using the platform

Item
I am satisfied with my performance in the platform
Learning with the platform was fun to do
While I was learning with the platform, I was thinking how much I enjoyed it
I thought learning with the platform was a boring
I think learning with the platform was very interesting
I thought learning with the platform was enjoyable
How much did using the platform make your head hurt?

iTalk2Learn task self-report. During the intervention, after completing each iTalk2Learn task and before they could progress to the next task, the students had to complete a brief pop-up questionnaire about the task (see Figure 3).

Figure 3. Self-report pop-up

Observational affect rating. While the students engaged with the system, the affective states of a subset of the UK participants (C1: *N* = 26; C2: *N* = 22) were monitored and noted using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP; Ocumpaugh et al., 2012). and the Human Affect Recording Tool (HART) Android mobile app (Ocumpaugh et al., 2015). BROMP gives strict guidelines on how the affective states of students are detected, for example by body posture, facial expression and engagement with the learning environment. The HART mobile app was then used to annotate the affective states with this protocol.

4.1.4 Procedure.

Individual sessions were run with up to 15 students in the UK and 30 students in Germany. Each session lasted approximately 90 minutes including breaks and was made up as follows.

During the first ten minutes, the students were introduced to the study and to the iTalk2learn platform (with the components being introduced depending on the experimental condition). To ensure that the introduction was as standardised as possible, it was scripted and was delivered by the same researchers in each session. The students were then asked to complete the paper-based fractions problems and the online measures. Students were given two minutes for each of the paper-based measures and ten minutes total for the online measures which were presented together in one browser window (questionnaire on attitudes to learning, mathematics and fractions, followed by online fractions problems).

Students then worked with the iTalk2Learn platform for approximately 40 minutes. After each iTalk2Learn task, students rated the iTalk2Learn task as described in section 4.1.3 before proceeding to the next task. During this main experimental period, the researchers adopted an intervention protocol that specified the allowable interactions and prompts (see Figure 4). For example, if a student requested help (noted as HU, for 'hand up' in Figure 4), a researcher would point to the screen to indicate where to access help from the system; or if the student was observed for at least 30 seconds to be stuck (not making progress or ignoring the task) (noted as OB, for 'observation' in Figure 4), a researcher would adopt the role of the class teacher by stepping in and guiding the student to the next appropriate stage; or if the student's system was observed to crash or the student reported a system crash (noted as OB/HU in Figure 4), a researcher would step in to reset the system and to get the student started again. The time of each intervention by a researcher was noted on the intervention protocol sheet.

Date: _____ Time: _____ School: _____
Condition 1 / 2 / 3 Researcher: _____

KEYWORDS	TRIGGER	PROMPT	TIME & STUDENT NUMBER
Nothing	OB	Please read and complete the task at the top of the screen	
Not answering + light bulb	OB	If you would like some help, click the light bulb	
Not answering - light bulb	OB	Please read and complete the task at the top of the screen	
Not answering, messing	OB	Please read and complete the task at the top of the screen	
Crash	OB/HU	I'll get Manolis to help you	
Finished, now what?	OB/HU	To continue, press the 'next' arrow.	
Don't understand Whizz	HU	Please read the task again	
Don't understand Whizz next	OB/HU	Click this arrow to move on.	
Don't understand the maths + light bulb	HU	If you would like some help, click the light bulb	
Don't understand the maths - light bulb	HU	Please read the task again	
Interface	HU	[point student to necessary step]	
Ask teacher	HU	[get student to click 'next' button again]	
Already done it	HU	It's not exactly the same task. Some of the numbers might have changed. See if you can answer it more quickly this time.	

Figure 4. Intervention protocol

In the last 30 minutes of the session, the students were asked to complete the final measures. Students were given two minutes for each of the paper-based fractions problems and twenty minutes total for the online measures which were presented together in one browser window (user experience questionnaire followed by online fractions problems. For the paper-based fractions problems, the students were given back the same sheets on which they had previously written their answers and they were asked to amend or add to their previous responses (using a pen, to distinguish their new responses from their original responses in pencil). For the online fractions problems, the students received the other, isomorphic version of the questionnaire.

4.1.5 iTalk2Learn platform.

The pedagogy of the iTalk2Learn platform is based on an intervention model for fostering robust knowledge which is described in D1.3 (also see Mazziotti et al., 2015). For the summative evaluation, the model was instantiated for the topic of equivalent fractions. It combined the ELE developed by the consortium, Fractions Lab, with structured practice environments in the form of ITSs. In the UK, this ITS was Maths-Whizz; in Germany, it was Fractions Tutor. The next section describes these learning environments and the tasks provided by them. Then, the adaptive support available to students is described. Finally, a section on the SNA explains how tasks were sequenced within and switched between learning environments.

Fractions Lab. As described in detail in D1.2 and D4.3.2, Fractions Lab is an ELE that provides tasks that aim to help the student develop conceptual knowledge of fractions. In the Fractions Lab interface (see Figure 5), a learning task is displayed at the top of the screen. Students can choose fraction representations (from the right-hand side menu) which they manipulate in order to solve (construct their own understanding of) the given task (for example, they can change the fraction's numerator or denominator, and find an equivalent fraction).

For example (as shown in Figure 5), an early task asked the student to create a fraction representation (with no limit on either the fraction or the representation), then to right click it, select 'find equivalent' from the resulting menu, and partition the fraction into 2, 3, 4 and 5. This particular task served both to introduce the student to available Fractions Lab functionality, and to introduce them to the idea and appearance of fraction equivalence with representations. While the students interacted with the system, they received adaptive support based on their screen interactions and their speech.

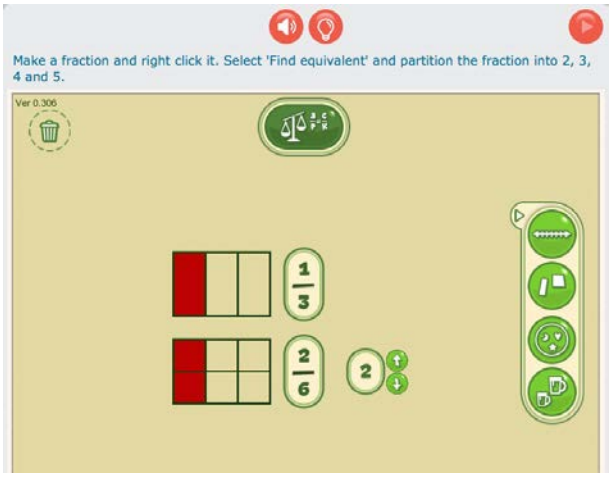


Figure 5. Screenshot of the Fractions Lab (ELE) interface

Maths-Whizz. As described in detail in D1.2, Maths-Whizz provides structured practice content that aims to help the student develop procedural knowledge of fractions. This content is delivered in three stages: a teaching page (which explains, procedurally, how to complete the following exercises

successfully), interactive exercises (questions with guided instruction and immediate feedback; see Figure 6) and a short test. When an incorrect answer is entered, Maths-Whizz provides feedback. Correct answers are rewarded with a celebratory response. The exercises use a range of graphical representations such as circles, rectangles, number lines, liquid measures and symbols within contexts that the students may be familiar with. Following an exercise, students are required to demonstrate their understanding in short tests, where no helps are available.

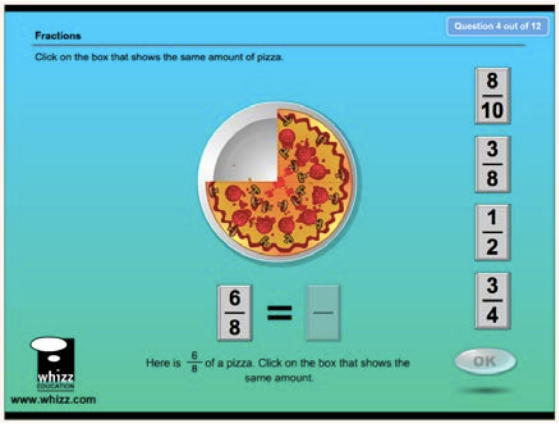


Figure 6. Screenshot of a typical Maths-Whizz equivalent fractions exercise.

Fractions Tutor. This web-based Cognitive Tutor for learning fractions (Olsen, Belenky, Aleven, & Rummel, 2013; Rau, Aleven, & Rummel, 2009, 2013; Rau, Aleven, Rummel, & Rohrbach, 2012) enables students to solve fractions problems step-by-step, and receive immediate feedback or ask for on-demand hints. Content is presented on the same page and revealed step by step while students solve the problem. See Figure 7 for an example.

Figure 7. Screenshot of a typical Fractions Tutor exercise.

Adaptive support. While the students interacted with the ELE (Fractions Lab) and with the structured practice environment (Maths-Whizz or Fractions Tutor), they were given automatic adaptive support: task-dependent support (TDS) and TIS. TDS was delivered by a green, male robot. TIS was delivered by a red, female robot.

TDS was built into Maths-Whizz and Fractions Tutor and not under the control of the iTalk2Learn platform. TDS for Fractions Lab was developed within the project and is detailed in D1.3. TDS provides problem-solving instruction (e.g. *"Remember that the denominator is the bottom part of the fraction."*), affirmation (e.g. *"The way that you worked that out was excellent."*) and reflection (e.g. *"Please explain why you made the denominator 12."*). For each of these types of feedback, four levels of complementary feedback were delivered in order: guidance (e.g. *"Did you know that you can click..."*), questions (e.g. *"How are you going to...?"*), didactic conceptual (e.g. *"You have changed the numerator. You need to change the denominator"*) and didactic procedural (e.g. *"You have changed the numerator to 12. You need to change the denominator to 12"*). How the TDS was delivered, whether interruptive (in a pop-up window that had to be interacted with before the student could continue) or non-interruptive (via an illuminated light-bulb that the student could choose whether or not to click), was determined by the student's affective state (as inferred by the TIS mechanism).

TIS was delivered while the student engaged with both the ELE (Fractions Lab) and the structured practice environment (Maths-Whizz and Fractions Tutor). TIS used the children's speech (while they solved structured or exploratory tasks) and interaction with the learning platform (while they solved exploratory tasks only). As described in D5.2 and D2.2.2, TIS aims to change a negative affective state (frustration or boredom) into a positive affective state (e.g. enjoyment) by adapting the feedback according to the student's affective state (e.g. asking students to talk aloud when frustrated helps them express their problems, which might move them out of their negative affective state). The speech indicators were used in different ways, with word recognition provided by a state-of-the-art speech recognition model trained with children's voices being used to determine whether students were talking aloud, whether they were using mathematical vocabulary, and whether they were saying something that indicated their affective state (enjoyment, surprise, confusion, frustration and boredom). Additional input for affect detection came from prosodic cues in the students' speech, with the PTDC using advanced machine learning models to extract from raw speech data whether the students were under-, appropriately, or over-challenged. TIS includes affect boosts (e.g. *"Well done. You're working really hard!"*), affirmation prompts (e.g. *"The way that you worked that out was excellent."*), instructive feedback (e.g. *"Use the comparison box to compare your fractions."*), reflective prompts (e.g. *"What do you notice about the two fractions?"*), and talk-aloud prompts (e.g. *"Please explain what are you doing."*).

SNA. As noted in D2.2.2, a key component of the full iTalk2Learn system is the SNA. Students began their iTalk2Learn session in the ELE (Fractions Lab). While the student was engaged with the ELE, the SNA drew on various inputs (e.g. student interaction, PTDC, word and affect recognition) to determine whether the student was under-, over-, or appropriately challenged by the task and thus to identify the next task appropriate for them. To ensure that within the limited available timeframe (40 minutes) every student experienced a range of exploratory and structure practice tasks, after each second task completed by the student, the SNA switched to the alternative type of task (i.e. when they had completed two exploratory tasks, they were switched to the structured practice environment, and vice versa). If the student was switched to the ELE, the level of challenge that they had experienced on the previous task was taken into account when calculating the next task. The first task in the structured practice environment was mapped to the fine-grain goal of the completed task

in the ELE (e.g. partition a fraction to find its equivalent). The next task in the structured practice environment stayed within the same fine-grain goal but increased the level of challenge based on a sequence determined by math education experts. Students continued in this fashion, alternating between exploratory and structured practice environments every second task until the 40 minutes was concluded.

4.2 Results

The following sections present results of preliminary analyses of the main measures used in the summative evaluation. Analyses of the other measures are time-consuming and underway at the time of this writing.

4.2.1 Online fractions problems.

For participants from the UK, Figure 8 presents the sum of scores on the online fractions problems at pre- and post-test for each of the three conditions. A repeated measures analysis of variance with time of measurement as the within-subjects factor and condition as the between-subjects factor showed a significant effect of time of measurement, $F(1,174) = 41.894$, $p < .001$, $\eta_p^2 = .194$, and a significant interaction effect of time and condition, $F(2,174) = 6.600$, $p = .002$, $\eta_p^2 = .071$. Planned contrasts revealed that students in conditions 1 and 2, who had received both exploratory and structured tasks, achieved significantly higher learning gains than students in condition 3, who had received structured tasks only, $F(1,174) = 5.048$, $p = .026$, $\eta_p^2 = .028$. Students in condition 1, who had received speech-based adaptivity, showed higher learning gains than students in condition 2, who had not received speech-based adaptivity, but this difference was not statistically significant, $F(1,174) < 1$.

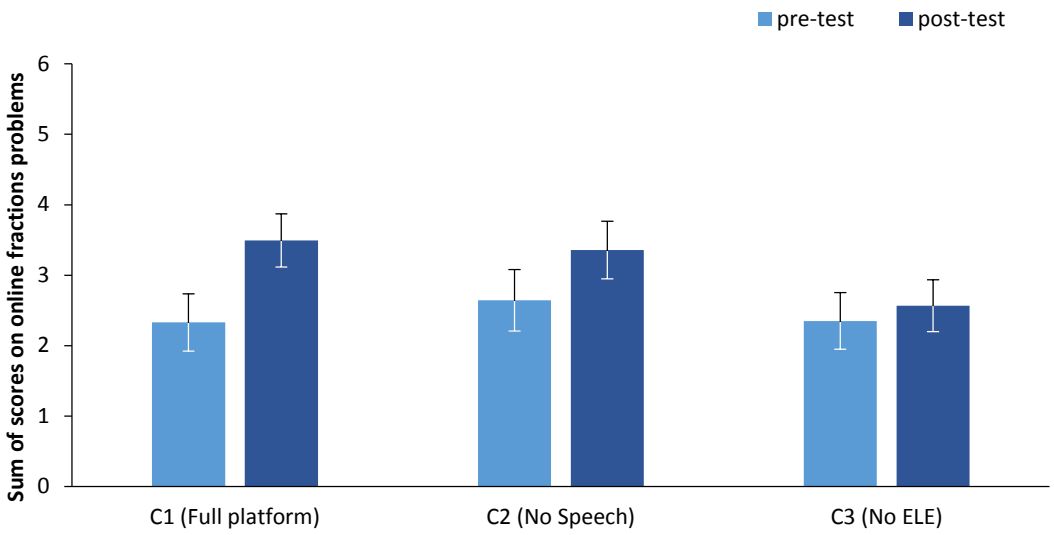


Figure 8. (UK) Sum of scores on online fractions problems as a function of condition and time of measurement.

For participants from Germany, Figure 9 presents the sum of scores on the online fractions problems at pre- and post-test for each of the three conditions. A repeated measures analysis of variance with time of measurement as the within-subjects factor and condition as the between-subjects factor showed a significant effect of time of measurement, $F(1,207) = 37.785$, $p < .001$, $\eta_p^2 = .164$, and a significant interaction effect of time and condition, $F(2,197) = 8.447$, $p < .001$, $\eta_p^2 = .075$. Planned contrasts revealed that students in conditions 1 and 2, who had received both exploratory and structured tasks, achieved significantly higher learning gains than students in condition 3, who had received structured tasks only, $F(1,207) = 64.535$, $p < .001$, $\eta_p^2 = .238$. Contrary to our first hypothesis, students in condition 2, who had not received speech-based adaptivity, achieved higher learning gains than students in condition 1, who had received speech-based adaptivity. To explore possible explanations, post-hoc analyses were conducted and revealed significant differences on the pre-test between conditions, $F(2,207) = 24.600$, $p < .001$, $\eta_p^2 = .193$. Least Significant Difference tests showed that students from condition 2 scored significantly higher on the pre-test than participants from condition 1 and participants from condition 3, both $p < .001$.

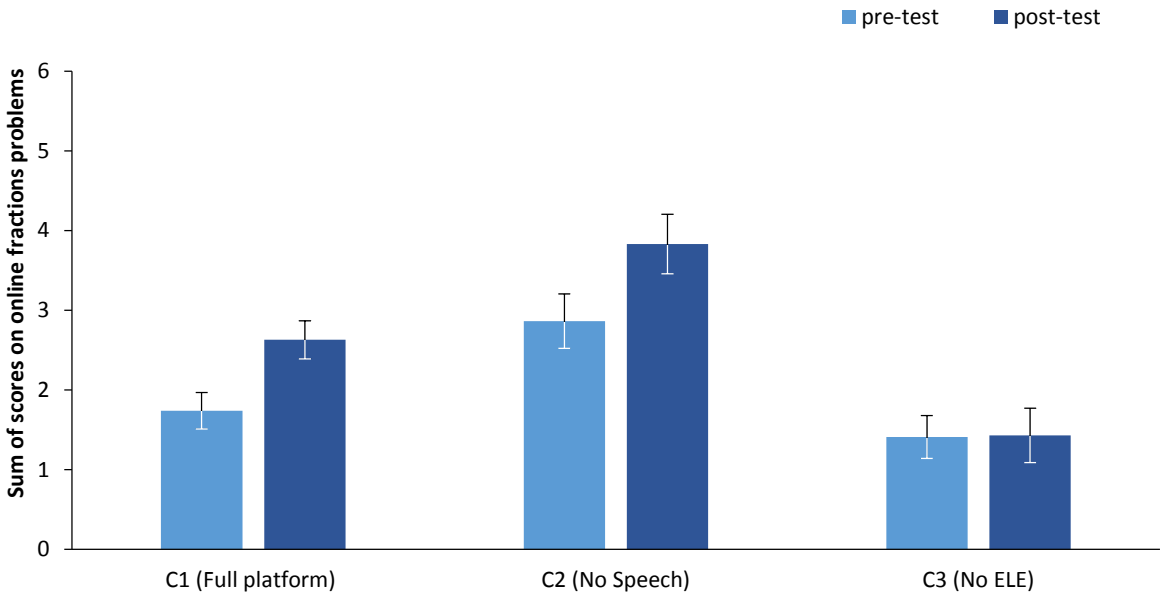


Figure 9. (Germany) Sum of scores on online fractions problems as a function of condition and time of measurement for Germany

4.2.2 Paper-based fractions problems.

Data from the paper-based fractions problems continue to be analysed. Study report 1 in the appendix presents analyses of misconceptions revealed by the first fractions problem ("show one third in as many ways you can") in the UK sample. This section presents analyses of knowledge gains revealed by the same fractions problem for both the UK and Germany.

In order to find additional evidence for the combination effect, we analysed the representational variety shown on the first paper-based fractions problem. Table 6 shows data for both the UK and Germany. These data are not available for all students who participated in the summative evaluation in Germany because some had to leave right after completing the online fractions problems. FL exposes students to a high number of different representations, so our hypothesis was as follows: Students learning with a platform version including FL show a larger increase in representational variety than students who learn with a platform version that does not contain FL.

For the UK, a repeated measures analysis of variance with time of measurement as the within-subjects factor and condition as the between-subjects factor showed a significant effect of time of measurement, $F(1,183) = 402.454$, $p < .001$, $\eta_p^2 = .687$, and a significant interaction effect of time and condition, $F(2,183) = 9.048$, $p < .001$, $\eta_p^2 = .090$ on the number of representations used to show one third. Consistent with our hypothesis, a planned contrast revealed that students who learned with a platform version including FL showed a larger increase in representational variety than their counterparts (i.e., condition 3), $F(1,183) = 7.506$, $p = .007$, $\eta_p^2 = .039$.

Table 6
Variety of representations as a function of country, condition, and time of measurement

	Number of representations used to show one-third			
	UK		Germany	
	Pre-test	Post-test	Pre-test	Post-test
Condition	*M (SD)*	*M (SD)*	*M (SD)*	*M (SD)*
C1 (Full Platform)	3.16 (1.43)	5.55 (1.83)	2.59 (1.56)	3.10 (1.57)
C2 (No Speech)	3.07 (1.44)	5.76 (2.12)	2.79 (1.17)	3.66 (1.35)
C3 (No ELE)	2.91 (1.55)	4.49 (2.21)	1.82 (1.04)	3.26 (1.85)

For Germany, a repeated measures analysis of variance with time of measurement as the within-subjects factor and condition as the between-subjects factor showed a significant effect of time of measurement, $F(1,203) = 38.406$, $p < .001$, $\eta_p^2 = .159$, and a significant interaction effect of time and condition, $F(2,203) = 3.334$, $p = .038$, $\eta_p^2 = .032$. Contrary to our hypothesis, students who learned with a platform version including FL showed a lower increase in representational variety than their counterparts. One explanation for this finding is that students in condition 3 already started with a lower number of representations used prior to interacting with FT and thus had more potential to "grow". To shed more light on these findings, we will conduct further analyses.

4.2.3 Evaluation of task-independent support.

In order to find additional evidence for the speech effect, we further investigated the feedback students received and its effect on student affect. The primary way speech is used by the system is by TIS to provide affective support. A comparison between the speech-enabled and the speech-disabled platform therefore allows to evaluate the effect of TIS on students' affect.

For these analyses, observations of student affect based on the BROMP protocol with the HART app were merged with platform log files. Results illustrate how TIS worked and its effectiveness. Students who were in condition 1 were less bored, while students in condition 2 engaged in more off-task behaviour. Flow correlated positively with the post-test score ($r = .307$) while off-task behaviour correlated negatively with the post-test score ($r = -.349$). Study report 2 in the appendix presents more details. A paper based on these analyses has been submitted to UMUAI.

4.2.4 User experience.

Table 7 presents descriptive statistics of the user experience ratings in the UK. In general, the platform was very well received. Participants' ratings of speech functionality, TDS (in the form of the male robot), and TIS (female robot) varied more strongly than other ratings. On average, TDS and TIS

were rated better than the speech functionality itself. Because the adaptive support tools built on the speech functionality, this may indicate that also in the participants' view, speech functionality was useful. Interestingly, the female robot's hints were perceived as more helpful when students worked with the full platform than when students worked with Maths-Whizz only. TIS was indeed more limited in condition 3 because it could not affect the delivery of TDS and provided affect boosts only based on speech recognition. Tasks were perceived as somewhat easy, participants felt competent working with the platform and did not think the platform made their head hurt too much. This indicates that the platform was successful in keeping participants in the flow. Participants found working with the platform interesting, enjoyable and not boring. Overall, these results speak to an engaging and accessible user experience.

Table 7
User experience ratings in the UK

	Condition					
	C1 (Full Platform)		C2 (No Speech)		C3 (No ELE)	
Items	*M*	*SD*	*M*	*SD*	*M*	*SD*
I would like to work with the platform again	4.25	1.07	4.29	.73	4.37	.97
The platform was easy to use	4.12	.88	4.11	.91	4.23	.81
I like speaking with the platform	3.31	1.55	[a]	[a]	3.66	1.45
The platform understood what I was saying	3.21	1.32	[a]	[a]	3.27	1.23
I always paid attention to the hints	3.95	1.06	3.77	1.13	3.97	1.16
The male robot's hints were helpful	3.43	1.33	3.07	1.34	[a]	[a]
The female robot's hints were helpful	3.90	1.12	[a]	[a]	1.84	1.36
The tasks were too hard	2.03	.86	1.91	.97	1.93	1.13
The tasks were too easy	3.65	.99	3.66	1.14	3.68	1.06
After working at this platform for a while, I felt pretty competent using the platform	4.45	.79	4.30	.84	4.25	.88
I am satisfied with my performance in the platform	4.47	.70	4.43	.87	4.53	.79

	Condition					
	C1 (Full Platform)		C2 (No Speech)		C3 (No ELE)	
Items	*M*	*SD*	*M*	*SD*	*M*	*SD*
Learning with the platform was fun to do	4.40	.72	4.32	.96	4.43	1.09
While I was learning with the platform, I was thinking how much I enjoyed it	4.00	1.04	4.05	1.00	4.14	1.28
I thought learning with the platform was a boring	1.77	1.13	1.93	1.14	1.85	1.30
I think learning with the platform was very interesting	4.35	.78	3.95	1.13	4.34	.98
I thought learning with the platform was enjoyable	4.43	.70	4.34	1.00	4.42	.94
How much did using the platform make your head hurt?	2.25	1.36	2.09	1.18	2.32	1.44

[a]. The component targeted by this item was not included in this condition, so the item was not applicable.

Table 8 presents user experience ratings in Germany. In general, the platform was also very well received by the German participants. Results largely mirror ratings by UK participants, but students from the UK generally reported a slightly more positive user experience than their German counterparts. German participants also found the platform understood them less well than UK participants. This may offer one explanation why participants in condition 1 learned less than participants in condition 2 in Germany.

Table 8
User experience in Germany

	Condition					
	C1 (Full Platform)		C2 (No Speech)		C3 (No ELE)	
Items	*M*	*SD*	*M*	*SD*	*M*	*SD*
I would like to work with the platform again	3.65	1.12	3.58	1.28	4.06	.93
The platform was easy to use	3.45	1.12	3.28	1.18	3.94	.84

	Condition					
	C1 (Full Platform)		C2 (No Speech)		C3 (No ELE)	
Items	*M*	*SD*	*M*	*SD*	*M*	*SD*
I like speaking with the platform	3.14	1.25	[a]	[a]	3.13	1.23
The platform understood what I was saying	2.42	1.08	[a]	[a]	2.47	1.18
I always paid attention to the hints	3.62	.99	3.63	1.00	3.94	.79
The male robot's hints were helpful	2.75	1.26	2.74	.89	[a]	[a]
The female robot's hints were helpful	3.38	1.22	[a]	[a]	2.74	1.09
The tasks were too hard	2.29	.95	1.83	.75	2.52	.95
The tasks were too easy	2.88	.98	3.08	.93	2.96	.92
After working at this platform for a while, I felt pretty competent using the platform	3.85	.98	3.78	1.29	4.16	.90
I am satisfied with my performance in the platform	3.75	1.05	3.86	1.07	4.25	.79
Learning with the platform was fun to do	3.76	1.09	3.67	1.16	4.39	.67
While I was learning with the platform, I was thinking how much I enjoyed it	3.09	1.20	3.00	1.22	3.75	1.00
I thought learning with the platform was a boring	2.23	1.18	2.31	1.21	1.69	.96
I think learning with the platform was very interesting	3.81	1.04	3.71	1.11	4.13	.87
I thought learning with the platform was enjoyable	3.64	1.07	3.60	1.09	3.93	.98

	Condition					
	C1 (Full Platform)		C2 (No Speech)		C3 (No ELE)	
Items	*M*	*SD*	*M*	*SD*	*M*	*SD*
How much did using the platform make your head hurt?	2.41	1.09	2.32	.94	2.22	1.04

[a]. The component targeted by this item was not included in this condition, so the item was not applicable.

5. Extended Evaluation: Exploring Thinking-in-Change (UK)

This section describes a study that was conducted in the UK to extend the summative evaluation to the coarse-grain goal of addition and subtraction, and to explore student's thinking-in-change. It collected additional data to investigate specifically:

1. the impact of the iTalk2Learn system on students' thinking-in-change about fractions, specifically:
   a) the role of representations in shaping students' thinking-in-change
   b) how students' thinking changes in relation to equivalence as a result of using the platform
   c) how students' thinking changes in relation to addition and subtraction as a result of using the platform
2. the impact of the SNA on student progression
3. the effect of feedback on student affect

The data are still being analysed at the time of this writing. Study report 3 in the appendix presents early findings on students' confidence levels in learning fractions. The following sections present the methodology of this study.

5.1 Methods

5.1.1 Participants.

After completing the summative evaluation, the students in the UK suburban school (N = 57) continued their involvement with this extended evaluation. A subsample (N =12) were interviewed pre- and post- intervention. These students were selected by their teachers and were chosen as students who were articulate and could discuss their mathematical thinking.

5.1.2 Instruments.

Like in the summative evaluation, the students completed online fractions problems. The problems for this extended evaluation addressed knowledge related to fraction size, comparison, addition and subtraction and emphasised a procedural approach. Students were randomly allocated one of two

isomorphic versions of four fractions problems at the first time of measurement and completed the second version at the second time of measurement.

In addition, students were interviewed. The interviews were designed to explore the issues described in Table 9.

Table 9
Topics covered in interviews conducted in extended evaluation

Topic	Pre	Post
Student's confidence in mathematics, and fractions learning in particular	✓	✓
Ability to write fractions	✓	
Ability to correctly order fractions	✓	✓
Explore range of representations used to show 1/4	✓	✓
Preferred FL representations	✓	✓
Understanding of equivalent fractions	✓	✓
Ability to add two fractions	✓	✓
Ability to subtract two fractions	✓	✓
Ability to identify addition misconceptions	✓	✓
Ability to identify subtraction misconceptions	✓	✓
Opinions on TDS and TIS feedback received		✓
Opinions on structured / exploratory tasks		✓
Opinions on using computers vs using pencil and paper		✓
Opinions on interruptive vs non-interruptive feedback		✓
Opinions on types of feedback (i.e. question, information, instruction)		✓

5.1.3 Procedure.

The study involved undertaking an additional session of 30 minutes per day for three further consecutive days. The online fractions problems were solved during the first day of the extended evaluation and the post-questionnaire was administered at the end of the final day. The subsample of 12 students was interviewed for 30 minutes prior to the intervention by one interviewer and for 45 minutes after the intervention by two interviewers.

5.1.4 Intervention.

Regardless of the condition the students had experienced during the summative evaluation, they were all involved during the extended evaluation with the full platform. However, there were three distinct differences between the full platform intervention for the summative evaluation and the extended evaluation. These are outlined in Table 10 and discussed further below. In order to distinguish between the two full platform conditions we refer to the extended evaluation's full platform as 'Condition 4'.

Table 10
Platform differences between the summative evaluation and the extended evaluation

	Summative evaluation (Condition 1)	Extended evaluation (Condition 4)
Task focus:	Fraction equivalence	Fraction addition and subtraction
Student support:	Full TDS	Partial TDS
	Full TIS	Full TIS
Intervention model:	Adaptive sequence based on SNA	Fixed sequence

Task focus. In order to extend the students' fractions knowledge, the tasks in the extended evaluation focused on addition and subtraction of fractions, moving from using fractions with like denominators to fractions with unlike denominators (see D1.2 Section 2.1 for further details of coarse-grain goals).

Student support. TIS continued to be available to students during the extended evaluation. One exploratory task utilised TDS. Due to time constraints we focused on TDS for the equivalence tasks because a) we assumed students would require more support earlier on; and b) more students were involved in the summative evaluation so testing TDS there was more important. If students were stuck, the researchers provided support that was reflective of TDS support.

Intervention model. A fixed sequence was set for the extended evaluation. The sequence was designed by a mathematics education expert who identified what might be 'typically' planned for set of a classroom-based lessons. The plan interweaves exploratory and structured tasks to introduce and consolidate learning, and moves through the coarse-grain goals in order (as per sequencer principles). Consideration was also given to the types of fraction representations that students would be exposed to, ensuring a variety over the course of the learning trajectory.

6. General Discussion

The discussion first focuses on the results of the summative evaluation. It then discusses possible future use cases of the platform and how the collected data can be analysed further. In the conclusion, we place the iTalk2Learn platform in the broader context of educational technology and its effectiveness to address societal challenges.

The iTalk2learn project developed an innovative speech-enabled learning platform that combines exploratory and structured tasks and tailors support in an adaptive fashion. The summative evaluation in Y3 investigated the robustness and efficacy of the platform for fostering fractions knowledge. Specifically, we investigated, firstly, whether a combination of exploratory and structured tasks promotes robust learning more than structured tasks alone (the combination effect; hypothesis 1). Across both countries we could confirm our hypothesis: Children learning with a combination of structured and exploratory learning tasks gained significantly more knowledge than children learning only with structured tasks. Secondly, the evaluation asked whether an adaptive system that interacts with learners through speech enhances learning more than an adaptive system that does not (the speech effect; hypothesis 2). The data presented here provided conflicting evidence for this hypothesis. In the UK, the speech-based system descriptively promoted learning more than the system without speech indicators. In Germany, the system without speech indicators promoted learning more than the speech-based system. This indicates that speech plays a more complex role in promoting learning than previously thought. These results are now discussed for both hypotheses separately.

Hypothesis 1. The summative evaluation provided clear evidence for the combination effect on students' robust knowledge. Beyond the fact that we could show the superiority of combining exploratory with structured tasks in the UK *and* in Germany, we could further show that the combination effect also holds true when using two different structured environments (Maths-Whizz in UK and Fractions Tutor in Germany). Knowing that combining both types of learning tasks helps students to gain robust knowledge is particularly important considering how difficult and challenging this mathematical topic is for young students (Charalambous & Pitta-Pantazi, 2007), and considering that students' fractions ability is a predictor for future maths performance (Siegler et al., 2012). Additionally, finding evidence for the combination effect underlines the need to foster both type of knowledge iteratively, as Rittle-Johnson and colleagues (2001) highlighted with their iterative model of knowledge development. Walking further down this path, we would need to investigate how children in condition 1 (and condition 2) scored on the procedural and conceptual knowledge test items separately. This way we could find out whether more practice time (as it was the case in condition 3) did lead to more procedural knowledge or whether fostering both types of knowledge adaptively did lead to more procedural knowledge. Preliminary analyses (not reported in this deliverable) suggest that in conditions 1 and 2, increases in procedural knowledge were similar to increases seen in condition 3, even though learning time was split between exploratory learning and structured practice in conditions 1 and 2. This may indicate that the combination of both learning tasks particularly fosters conceptual knowledge without harming procedural knowledge acquisition.

Interestingly, learning gains were very limited in condition 3 (No ELE). Different from the usual contexts in which Maths-Whizz and Fractions Tutor are deployed, in the summative evaluation students had very limited time to study very specific learning content. Moreover, participants had not worked with these learning environments before. Against this background, the clear learning gains observed in conditions 1 and 2 are even more impressive. The preliminary analyses of scores on fractions problems that emphasized conceptual versus procedural knowledge (not reported in this deliverable) suggest that the observed learning gains reflect primarily an increase in conceptual knowledge for students participating in conditions 1 and 2. These analyses also suggest that learning gains observed in condition 3 reflect primarily an increase in procedural knowledge which is comparable to the increase in procedural knowledge in conditions 1 and 2.

This combination effect prompts a series of follow-up questions. One of these questions asks for the component that makes the combination effect effective. For example, is the order of exploratory followed by structured tasks essential for realizing the combination effect? Or could the order be reversed? As we described in D1.3, we based the order of exploratory and structured tasks implemented in our studies on prior research that showed conceptual learning should be fostered first. This principle was not only realized within the first two tasks, but formed one of the rules of our intervention model employed throughout learning with iTalk2Learn. The rule states that if a student is over-challenged with a given task and thus has probably not yet fully understood the concept or accomplished the procedure, then it is best to provide the student with an exploratory, conceptually-oriented task to learn the concept first. The iTalk2Learn system now provides an additional, proven research context in which the generalizability of the prior research findings can be tested.

Another question concerns adaptivity and learning time. In the present study, the sequence of exploratory tasks within the ELE was adapted to the students' level of challenge. The sequence of structured tasks within the ITSs was not adapted to individual students. The sequencing approach of the VPS requires a certain number of interactions before being able to produce a meaningful prediction. The learning time in the summative evaluation was too short to allow for this. This affects not only the VPS in its approach, but any sequencer or switcher. A follow-up study could investigate whether the combination effect extends to a system that also includes more adaptive task sequencing for structured tasks when learning time is sufficient to allow the benefits of this adaptivity to emerge.

Hypothesis 2. The summative evaluation provided conflicting evidence of the speech effect. While in the UK students learned more with the speech-enabled platform as compared to the students learning with the platform version without speech indicators, in Germany this result pattern was reversed. At the time of this writing, the data collection had only finished four weeks ago. The conflicting evidence requires complex analyses that will feed into additional, planned publications (see D6.3.3). There are a number of research hypotheses that can be explored regarding the implementation and its effects on affect and learning in the two countries. At this early stage, we limit our discussion to one possible explanations: the benefits of speech may depend on prior knowledge of students. Specifically, the benefits of speech may be more pronounced for students with low prior knowledge. Students in condition 2 did not receive speech-based support and did not need it in Germany, perhaps because their prior knowledge, which was significantly higher than prior

knowledge of students in condition 1, was sufficient to master the tasks provided to them by the platform. It is also possible that the higher prior knowledge in condition 2 by itself contributed to higher learning gains in comparison to condition 1 (Matthew effect; Morgan, Farkas, & Wu, 2011; Stanovich, 1986).

6.1 Future Developments

This deliverable could only present preliminary analyses of the data collected in the summative evaluation due to the tight timeframe of projects that develop new technology such as this one. There are a number of questions that remain yet to be addressed. For example, we have not been able to fully analyse data collected in the log files. As a first step, log file analyses will allow us to check for every student that the platform was implemented as intended. In line with D1.3, it will be particularly interesting to find further empirical support for the effectiveness of our intervention model that is the nodal point where all single components of the iTalk2Learn system intertwine. For instance, we will investigate how adaptive our intervention model and therefore the system actually was by counting the frequency of students being appropriately challenged.

Moreover, based on pre-test scores, students can be selected that have low, medium or high prior knowledge. Case studies can illustrate how the system responded to these individual students. It would be particularly interesting to see whether the system provided similar sequences to students with similar prior knowledge.

We also plan to have a mathematics education expert evaluate the decisions made by the system: the expert can watch individual student actions in the system and listen to their speech. The expert could then assess whether she finds the support provided by the system and the sequence of tasks appropriate. Such a study could validate the intervention model implemented in the system.

Finally, log file analyses can address a number of further research questions concerning individual components of the intervention model, for example the accuracy of affect detection, the role it plays in providing support to students and how students overcame misconceptions while learning with iTalk2Learn. Initial analyses have been reported in section 4.2.3. A number of follow-up publications have been planned already (see also D6.3.3).

The platform also offers many opportunities for conducting further research. For example, enabling students to learn collaboratively with the platform seems to be very promising as collaborative learning might further support students exploratory behaviour and hence additionally support students' conceptual knowledge development. A collaborative extension to iTalk2Learn, similarly to the extension for Cognitive Tutor Algebra, could help investigate this question further (Diziol et al., 2010; Rummel et al., 2012).

The SNA in the summative evaluation used a simple, rule-based algorithm to sequence tasks. A follow-up study could develop a Bayesian network and train it with data collected during iTalk2Learn's summative evaluation to provide more intelligent sequencing of tasks (Mazziotti et al., 2015). This project could also further explore how best to combine exploratory and structured tasks.

The platform has successfully been implemented in classrooms during the summative evaluation. For better classroom integration, we would recommend two things: First, we will summarize our experience and write guidelines for the use of the iTalk2Learn system for teachers, as has been done already for Fractions Lab (see http://fractionslab.lkl.ac.uk/). This website also allows teachers to design tasks for their students. Additionally, we recommend to create a teacher dashboard that allows teachers to monitor students' progress. The Wizard-of-Oz tools programmed by BBK, which allows teachers to send messages to students and choose a sequence of tasks for them, provide a starting point for this.

6.2 Conclusion

In the aftermath of the PISA studies, which identified weaknesses of students in many European countries, especially in mathematics, the education of children in the elementary school grades received a lot of attention. Yet, most learning systems that have been developed for mathematics education have two significant limitations: first, they are usually constrained to text-based interactions and are thus hard to use by young learners who are still developing their basic literacy skills. Second, they are usually constrained either to exploratory tasks or to structured tasks and thus can promote robust learning only to a limited extent. While both task types allow for both conceptual and procedural learning, exploratory tasks are suited best for conceptual learning while structured tasks are suited best for procedural learning.

The iTalk2Learn platform overcomes these limitations by combining exploratory tasks from Fractions Lab, a newly-developed exploratory learning environment, and structured tasks from Maths-Whizz and Fractions Tutor, two proven intelligent tutoring systems. Further, it uses speech for interaction and for adaptation to learners' needs, and provides not only intelligent cognitive but also affective support. These innovations, tied together into one, unified platform, present a significant advance in the fields of computer science, mathematics education, and educational psychology. The summative evaluation examined the benefits of these innovations for learning in authentic classroom settings. The results clearly demonstrate that the combination of structured and exploratory tasks promotes learning more than structured tasks alone. This result was replicated in two countries which underlines the effectiveness of iTalk2Learn to foster robust fractions knowledge in students. The results also showed that the role of speech is more complex than previously thought which opens up new avenues for research. The summative evaluation demonstrated the transferability of the platform from the research to regular school settings: the platform can effectively support mathematics instruction in classrooms. Finally, it can serve as a testbed for future technological and theoretical developments.

The project has thus shown how the promise of educational technologies can be realized in tackling challenges in education. Before technology can be developed, a deep understanding of the challenge it should address is required. The work on common misconceptions and ways to represent fractions in iTalk2Learn was one way of doing so. Technology should then be developed with a pedagogically sound intervention model as its backbone that specifies how best to learn the targeted content, here realized by the SNA. It is important to involve prospective users in the development, not only to

ensure a good user experience, but also to foster a sense of ownership in students: students and teachers have had a hand in developing iTalk2Learn and provided vital input. Finally, the opportunity that lies in adapting to individual students' needs should be leveraged to the fullest extent possible: intelligent support can provide cognitive, but also affective support, be informed not only by screen and mouse action but also by natural language interaction, and its subcomponents can mutually influence each other, for example by taking affective states into account when delivering cognitive support. This three-year project has shown what a lot of work it is to simulate a teacher with educational technology. But the scalability and effectiveness of the iTalk2Learn platform made it a worthwhile effort.

7. References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Charalambous, C. Y., & Pitta-Pantazi, D. (2007). Drawing on a theoretical model to study students' understandings of fractions. *Educational Studies in Mathematics*, *64*(3), 293–316. doi:10.1007/s10649-006-9036-2.

Diziol, D., Walker, E., Rummel, N., & Koedinger, K. (2010). Using intelligent tutor technology to implement adaptive support for student collaboration. *Educational Psychology Review*, *22*(1), 89–102. doi:10.1007/s10648-009-9116-9.

Fiorella, L., Vogel-Walcutt, J. J., & Schatz, S. (2012). Applying the modality principle to real-time feedback and the acquisition of higher-order cognitive skills. *Educational Technology Research and Development*, *60*(2), 223–238. doi:10.1007/s11423-011-9218-1.

Grawemeyer, B., Holmes, W., Gutiérrez-Santos, S., Hansen, A., Loibl, K., & Mavrikis, M. (2015). Light-bulb moment? Towards adaptive presentation of feedback based on students' affective state. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (pp. 400–404). Atlanta, Georgia, USA: ACM. doi:10.1145/2678025.2701377.

Grawemeyer, B., Mavrikis, M., Hansen, A., Mazziotti, C., & Gutiérrez-Santos, S. (2014). Employing speech to contribute to modelling and adapting to students' affective states. 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. 568–569). Springer International Publishing. doi:10.1007/978-3-319-11200-8_73.

Grawemeyer, B., Mavrikis, M., Holmes, W., & Gutiérrez-Santos, S. (2015). Adapting feedback types according to students' affective states. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Lecture Notes in Computer Science. Artificial Intelligence in Education* (pp. 586–590). Springer International Publishing. doi:10.1007/978-3-319-19773-9_68.

Grawemeyer, B., Mavrikis, M., Holmes, W., Hansen, A., Loibl, K., & Gutiérrez-Santos, S. (2015a). Affect matters: Exploring the impact of feedback during mathematical tasks in an exploratory environment. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Lecture Notes in Computer Science. Artificial intelligence in education* (pp. 595–599). Springer International Publishing. doi:10.1007/978-3-319-19773-9_70.

Grawemeyer, B., Mavrikis, M., Holmes, W., Hansen, A., Loibl, K., & Gutiérrez-Santos, S. (2015b). The impact of feedback on students' affective states. In G. Rebolledo-Mendez, M. Mavrikis, O. C. Santos, B. Du Boulay, B. Grawemeyer, & R. Rojano-Caceres (Eds.), *Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education AIED 2015. Volume 7: International Workshop on Affect, Meta-Affect, Data and Learning (AMADL 2015)* (pp. 4–13).

Hausmann, R. G. M., & Chi, M. T. (2002). Can a computer interface support self-explaining? *Cognitive Technology*, *7*(1).

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and

Inquiry-Based Teaching. *Educational Psychologist*, *41*(2), 75–86. doi:10.1207/s15326985ep4102_1.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive Tutors: Technology bringing learning sciences to the classroom. In K. R. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 61–77). New York, NY, US: Cambridge University Press.

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.

LeFevre, J.-A., Smith-Chant, B. L., Fast, L., Skwarchuk, S.-L., Sargla, E., Arnup, J. S.,. . . Kamawar, D. (2006). What counts as knowing? The development of conceptual and procedural knowledge of counting from kindergarten through Grade 2. *Journal of Experimental Child Psychology*, *93*(4), 285–303. doi:10.1016/j.jecp.2005.11.002.

Lesh, R. (1999). The development of representational abilities in middle school mathematics. In I. E. Sigel & K. Tyner (Eds.), *Development of mental representation: Theories and applications* (pp. 323–349). Hillsdale, NJ: Erlbaum.

Martin, T., Petrick Smith, C., Forsgren, N., Aghababyan, A., Janisiewicz, P., & Baker, S. (2015). Learning Fractions by Splitting: Using Learning Analytics to Illuminate the Development of Mathematical Understanding. *Journal of the Learning Sciences*, 150814113842002. doi:10.1080/10508406.2015.1078244.

Mavrikis, M., Grawemeyer, B., Hansen, A., & Gutiérrez-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.

Mavrikis, M., Gutiérrez-Santos, S., Geraniou, E., & Noss, R. (2013). Design requirements, student perception indicators and validation metrics for intelligent exploratory learning environments. *Personal and Ubiquitous Computing*, *17*(8), 1605–1620. doi:10.1007/s00779-012-0524-3.

Mazziotti, C., Holmes, W., Wiedmann, M., Loibl, K., Rummel, N., Mavrikis, M.,. . . Grawemeyer, B. (2015). Robust student knowledge: Adapting to individual student needs as they explore the concepts and practice the procedures of fractions. In M. Mavrikis, G. Biswas, S. Gutiérrez-Santos, T. Dragon, R. Luckin, D. Spikol, & J. Segedy (Eds.), *Proceedings of the Workshops at the 17th International Conference on Artificial Intelligence in Education AIED 2015. Volume 2: Intelligent Support in Exploratory and Open-ended Learning Environments; Learning Analytics for Project Based and Experiential Learning Scenarios* (pp. 32–40).

Mercer, N., & Sams, C. (2006). Teaching children how to use language to solve maths problems. *Language and Education*, *20*(6), 507–528. doi:10.2167/le678.0.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities*, *44*(5), 472–488. doi:10.1177/0022219411414010.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of project LISTEN: Smart Machines in Education. In K. D. Forbus & P. J. Feltovich (Eds.), *Smart machines in education: The coming revolution in educational technology* (pp. 169–234). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=570950.570957

Noss, R., Poulovassilis, A., Geraniou, E., Gutiérrez-Santos, S., Hoyles, C., Kahn, K.,. . . Mavrikis, M. (2012). The design of a system to support exploratory learning of algebraic generalisation: CAL 2011The CAL Conference 2011. *Computers & Education*, *59*(1), 63–81. doi:10.1016/j.compedu.2011.09.021.

Ocumpaugh, J., Baker, R. S., Rodrigo, M. M., Salvi, A., van Velsen, M., Aghababyan, A., & Martin, T. (2015). HART: the human affect recording tool. In *Proceedings of the 33rd Annual International Conference on the Design of Communication* (pp. 1–6). Limerick, Ireland: ACM. doi:10.1145/2775441.2775480.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. T. (2012). *Baker-Rodrigo Observation Method Protocol (BROMP 1.0. training manual version 1.0.)*. New York, N.Y., Manila, Philippines: EdLab; Ateneo Laboratory for the Learning Sciences.

Olsen, J. K., Belenky, D. M., Aleven, V., & Rummel, N. (2013). Intelligent tutoring systems for collaborative learning: Enhancements to authoring tools. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Lecture Notes in Computer Science. Artificial Intelligence in Education* (pp. 900–903). Springer Berlin Heidelberg. doi:10.1007/978-3-642-39112-5_141.

Paramythis, A., Weibelzahl, S., & Masthoff, J. (2010). Layered evaluation of interactive adaptive systems: Framework and formative methods. *User Modeling and User-Adapted Interaction*, *20*(5), 383–453. doi:10.1007/s11257-010-9082-4.

Rajala, A., Hilppö, J., & Lipponen, L. (2012). The emergence of inclusive exploratory talk in primary students' peer interaction. *International Journal of Educational Research*, *53*, 55–67. doi:10.1016/j.ijer.2011.12.011.

Rau, M. A., Aleven, V., & Rummel, N. (2009). Intelligent tutoring systems with multiple representations and self-explanation prompts support learning of fractions. In V. Dimitrova, R. Mizoguchi, & B. Du Boulay (Eds.), *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (pp. 441–448). IOS Press.

Rau, M. A., Aleven, V., & Rummel, N. (2013). Interleaved practice in multi-dimensional learning tasks: Which dimension should we interleave? *Learning and Instruction*, *23*, 98–114. doi:10.1016/j.learninstruc.2012.07.003.

Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: Fluency matters too! ITS support for robust learning with multiple representations. In S. A. Cerri, W. J. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Lecture Notes in Computer Science. Intelligent Tutoring Systems* (pp. 174–184). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-30950-2_23.

Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, *93*(2), 346–362. doi:10.1037/0022-0663.93.2.346.

Rummel, N., Mullins, D., & Spada, H. (2012). Scripted collaborative learning with the cognitive tutor algebra. *International Journal of Computer-Supported Collaborative Learning*, *7*(2), 307–339. doi:10.1007/s11412-012-9146-z.

Schuh, J. (2013, September 23). Saying goodbye to our old friend NPAPI [Web log post]. Retrieved from http://blog.chromium.org/2013/09/saying-goodbye-to-our-old-friend-npapi.html

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M.,. . . Chen, M. (2012). Early Predictors of High School Mathematics Achievement. *Psychological Science*, *23*(7), 691–697. doi:10.1177/0956797612440101.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, *21*(4), 360–407. doi:10.2307/747612.

Teasley, S. D. (1995). The role of talk in children's peer collaborations. *Developmental Psychology*, *31*(2), 207–220. doi:10.1037/0012-1649.31.2.207.

VanLehn, K. (2006). The behavior of tutoring Systems. *International Journal of Artificial Intelligence in Education*, *16*(3), 227–265.

Verschaffel, L., & Corte, E. de. (1996). Number and arithmetic. In A. Bishop, K. Clements, C. Keitel, J. Kilpatrick, & C. Laborde (Eds.), *Kluwer International Handbooks of Education. International handbook of mathematics education* (pp. 99–137). Springer Netherlands. doi:10.1007/978-94-009-1465-0_4.

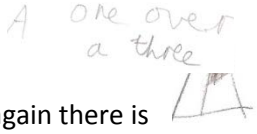
Voß, L., Schatten, C., Mazziotti, C., & Schmidt-Thieme, L. (2015). A transfer learning approach for applying matrix factorization to small ITS datasets. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros,. . . M. Desmarais (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 372–375).

Zakin, A. (2007). Metacognition and the use of inner speech in children's thinking: A tool teachers can use. *Journal of education and human development*, *1*(2), 1–14.

8. Appendix

Study report 1

Students' fractions misconceptions
Date(s): M33
Participants: 210 Year 4 and 5 students in the UK (8-10 years old)
Aim(s) of study: • To identify the misconceptions that are displayed through errors in fraction representations.
Method: As part of the summative evaluations, the students were provided with a blank sheet of paper to draw or write all the ways they think about one third. This task was given as a pre- and post-task as part of the summative evaluations. All errors were analysed and a framework for types of misconceptions according to conceptual understanding of fractions was formulated. It is possible to identify misconceptions according to representation type, or according to conceptual development. We have focused on the latter, to inform our knowledge for building students' robust mathematical knowledge.
Results/findings: ***Parts of the whole are not equal in size*** In all representation types students drew one third where the parts of the whole were not equal in size. This may show that they are yet to understand that when a whole is split up the parts must be evenly sized.    ***Not identifying the fractional part*** Some students drew the correct representation but omitted to identify the fractional part. In this instance they have understood that the whole divided into three represents one third, rather than the single part out of the thirds. ***Whole number bias*** Much has been written about whole number bias (REFS). There were textual / reading examples such as: "A one over a three"; "1 and 3"; "One on top of three"; "1 minus 3" that illustrate this significant issue in fractions learning. Another example showed one apple over three bananas.  Furthermore we saw whole number bias in one third being referred to as an ordinal number: "First, second, third"; "1-3rd". While the earlier examples are commonly found in the literature, the notion of fractional numbers being conceptualised as ordinal numbers has not been published to our knowledge. ***Misunderstanding numerator/denominator*** Within area representations students coloured the incorrect number of sections; colouring 1 and leaving three (to incorrectly show 1/4) and one student wrote "1 shaded and 3 not". This has been discussed in Hansen (2005, 2009, 2014).    There was occasional confusion with the order of how the numerator/denominator should be written when listing equivalent fractions: e.g. 6/2; 75/25. This may reflect students' lack of understanding of the numerator and denominator or could simply be an annotation error. 



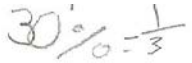
Some students correctly used time (showing 1/3 of an hour as 20 minutes) but others erroneously gave a time of 3:00 or drew a clock showing three o'clock. Once again there is a misconception related to fractions and what the numerator/denominator refer to.

Misunderstanding equivalence

Several errors showed that the students struggled with equivalence. Some of the errors could also be related to whole-number bias but are listed here for sake of closest-match.

Here, the students are finding 1/3 of the numerator and the denominator separately and stating "1/3 is one third of 3/9" or "half of 2/6 makes 1/3". This error straddles the whole-number bias literature and equivalence literature. To our knowledge, nothing has been published explicitly about this error.

There appears to be a direct translation error in symbol representations from fraction to decimal (e.g. from 1/3 to 1.3 or 0.3) and from fraction to fraction (e.g. 1/3 to 3/10)



Similarly, there were direct translation errors in symbol representations from fraction to percentage, with 1/3 being mapped to 30%.

Conclusions:

We note that drawing representations of one third is difficult for many students, and students can therefore be limited by their artistic ability. As a result, students may not show the full range of representations they know because they do not draw them, or the errors they make in drawings may be acceptable to the student and may not reflect a misconception.

It appears that we have explicitly identified some misconceptions that have not been written about before. "1/3 is one third of 3/9" is worthy of further investigation because we have additional data from other parts of the project that also point to this misconception. Publishing in this area would extend both the whole-number bias and equivalence literature.

We also noted that it was interesting that incorrect and correct representations sit alongside each other and often create a contradiction that a student appears to find acceptable. This reflects the messy nature of constructivism.

Next steps:

Our first next step is to broaden the dataset to include the German students' drawings of one third. We will then consider the impact of the platform on addressing students misconceptions. This will include analysing the students' errors in relation to the three conditions to see if there was an impact on the number of errors corrected or drawn post-test. Further analysis into how common the various errors were will also be carried out. Our findings on misconceptions will be submitted in a conference paper to the 40th Conference of the International Group for the Psychology of Mathematics Education next year and will then be further worked into a paper for the international mathematics education journal Mathematical Thinking and Learning.

In the meantime these misconceptions have been published on the project blog and are being published in a professional journal Mathematics Teacher, the journal of the Association of Teacher of Mathematics (ATM) in the UK.

Study report 2

Evaluation of task-independent support
Date(s): M33
Participants: 77 participants from the UK summative evaluation study
Aim(s) of study: • to evaluate the task independent support, students' in condition 1 that used the speech components with the affect aware support switched on were compared to students' from condition 2, where the speech components and the affect aware support were switched off. For the purposes of this study report, we refer to condition 1 as the affect condition and condition 2 as the non-affect condition.
Method: The participating students were roughly stratified, according to previous teacher assessments of the children's mathematical ability, and then randomly allocated to two sub-groups (approximately equal in size, with each group having approximately the same number of high, middle and low achieving students). The first group (N = 41) was assigned to the affect condition: the students were given access to the full iTalk2Learn system, which uses the student's affective state and their performance to determine the appropriate feedback and its presentation. The second group of students (N = 36) was assigned to the non-affect condition: they were given access to a version of the iTalk2Learn system in which feedback is based on the student's performance only. Two sessions, one for each condition, were undertaken in each school. At the beginning of each session, students completed an online questionnaire that assessed their knowledge of fractions (the pre-test). This was followed by 40 minutes during which the students engaged with fractions tasks in a version of the iTalk2Learn system that, according to the experimental condition, included either the affect-aware or the non-affect-aware support. There were two sets of support. One set, based on the student's interaction, was provided to both groups, whereas the second set, based on the student's affective state, differed according to the condition as follows: – Support based on interaction / performance: TALK ALOUD prompts were based on interaction only and were provided when students did not say anything for 30 seconds. This feedback was only provided in the affect group. The TASK SEQUENCE prompts were based on interaction only and were provided when students try to go to the next task when they have not completed the current task. The AFFIRMATION prompts were based on performance and were provided in both groups when students successfully completed the task. All of these feedback types were provided in both groups in a high-interruptive way (pop-up window). – Support based on affect / performance: AFFECT BOOSTS were based on student's affective state. This feedback was only provided in the affect group. INSTRUCTIVE feedback, OTHER PROBLEM SOLVING support and REFLECTIVE prompts were based on a combination of affect and performance within the affect group, but within the non-affect group these were only based on performance. Within the affect condition the presentation of the feedback (high- or low- interruptive) was based on their affect. In contrast, in the non-affect condition these feedback types were all presented in a low-interruptive way through the light bulb. While the students engaged with the system, the affective states of a subset of the students' (affect condition: N = 26; non-affect condition: N = 22) were monitored and noted using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP) and the Human Affect Recording Tool (HART) Android mobile app [1]. BROMP gives strict guidelines on how the affective states of students are detected by e.g. body posture, facial expression and engagement with the learning environment. The HART mobile app was then used to annotate the affective states with this protocol.

After the 40 minutes, the students completed a second online questionnaire that again assessed their knowledge of fractions (a post-test similar to the pre- test).

Results

Affect detection

In the affect condition, the affective states of the students were detected automatically by the system as it analyses their speech as well as their interaction, as described earlier. Additionally, the affective states were annotated by a researcher using the HART mobile app with the BROMP method, as described above.

The affective states that were detected automatically include flow, confusion, frustration, boredom, and surprise. With the HART mobile app the affective states included the same as the automatic detected ones with two added affective state: delight and eureka that humans were able to detect. Both of those data sources include time stamps, identifying when the particular affective state occurred. The affective state from the automatic detection and the HART annotations were matched according to their time stamp (with a 30 seconds window).

There was a moderate agreement between the automatic detection and the HART annotations, Kappa=.53, p<.001 (74.07% agreement).

The difference is partly due to the two affective states that were detected with the HART tool but that were not included in the automatic detection i.e. (delight and eureka). Additionally, we knew from our formative phase that surprise and boredom are difficult to detect automatically. Excluding those affective states a good (high) agreement between the automatic detection and the HART annotations is achieved, Kappa=.62, p<.001 (80.00% agreement). However, this result practically ignores the human annotation and implies that the annotated states along side two or several states are less transient that they probably are. Regardless, the result is quite satisfactory, particularly when considering that the effect of a misclassification is an intervention with a relative low cost to a student.

Adapting the feedback message types

In the affect condition, 1971 feedback messages were provided to students. On average, students received 48.07 (SD=14.58) messages (min=25; max=92). In the non-affect condition students received 2007 messages. Here, on average, 55.75 (SD=11.77) messages were provided to students (min=34, max=88).

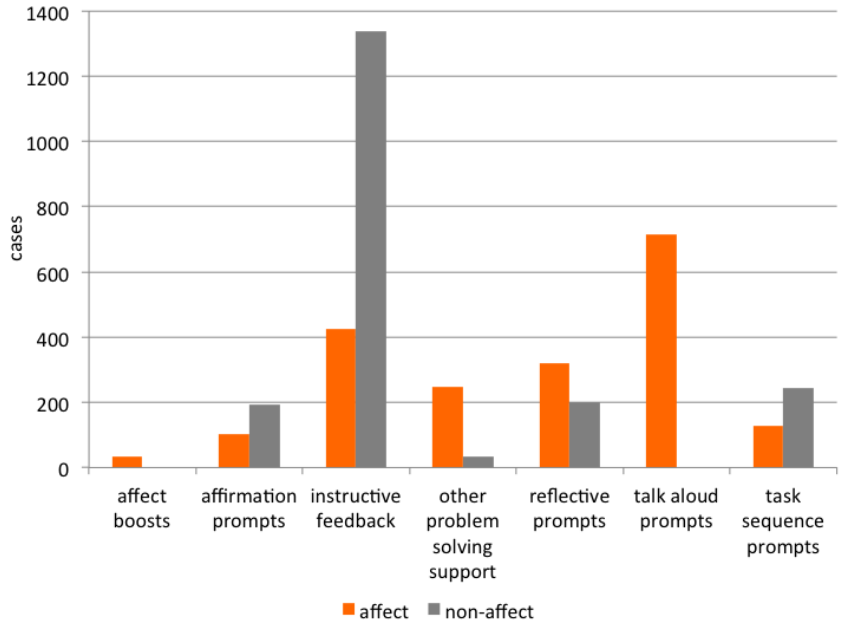


Fig. 1 Feedback types provided in the affect and non-affect condition.

Independent t-tests for each feedback types revealed significant differences between the groups. There was a significant difference in how often AFFIRMATION prompts were provided between the affect (M=2.51, SD=2.09) and the non-affect (M=5.33, SD=2.41) group (t(75)=-5.5, p<.05). There was also a significant difference in how many INSTRUCTIVE feedback was provided between the affect (M=10.32, SD=7.04) and the non-affect (M=37.14, SD=11.75) group (t(56)=-11.94, p<.05)). As well as for OTHER PROBLEM SOLVING support (affect: M=6.05 , SD=2.55 ; non-affect: M=0.97 , SD=2.21; t(75)=9.36, p<.05), REFLECTIVE prompts (affect: M= 7.80, SD=3.49 ; non- affect: M=5.53 , SD=2.21; t(68)=3.46, p<.05), and TASK SEQUENCE prompts (affect: M=3.12 , SD=2.60 ; non-affect: M=6.78 , SD=4.22; t(57)=-4.50, p<.05).

As described earlier, AFFECT BOOSTS and TALK ALOUD prompts were only provided in the affect condition (affect boosts: M=0.80, SD=1.40; talk aloud prompts: M=17.46, SD=5.92) and could not be compared with the non- affect condition.

Adapting the presentation of feedback

As described earlier, the feedback message was either displayed in a low- interruptive (light bulb) or in a high-interruptive way (pop-up window). The way in which the feedback was displayed depended on the type of message (interactive / performance) and on the student's affective state if they were in the affect condition.

In the affect condition, students viewed 1016 feedback messages (M=24.78, SD=9.67, min=11, max=54). In the non-affect condition, students viewed 963 messages (M=26.75, SD=10.61, min=12, max=56). An independent t-test revealed no significant difference between the groups in the number of feedback messages viewed (t(75)=-8.52, p>.05).

When feedback was low-interruptive (a glowing light bulb), students could either click on the light bulb and receive the feedback or they could ignore the light bulb and not receive the feedback. In the affect condition, 955 feedback messages were ignored (M=23, SD=7.54, min=8 , max=40). In the non-affect condition, students ignored 1044 feedback messages (M=29.00, SD=11.05, min=7, max=52). An independent t-test showed a significant differences be- tween the groups on whether or not they ignored the feedback (t(61)=-2.61, p<.05).

Affect and task behaviour

As described earlier, for a subset of students' (affect condition: N=26; non- affect condition: N=22) the affective states and task behaviour were monitored by using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP) and the Human Affect Recording Tool (HART) Android mobile app [1]. For each student, a set of affective states and task behaviour was annotated. Based on this, the percentage that a student was in a particular affective state and certain task behaviour was calculated. This was used for further analysis as described below.

Affect

Figure 2 shows the different types of affective states that were detected in the affect and non-affect condition.

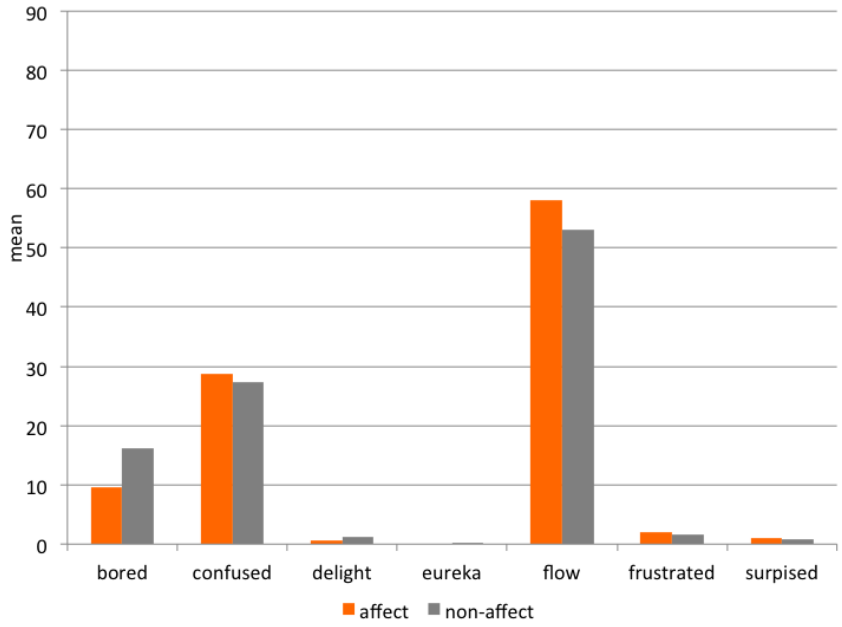


Fig. 2 affective states during the main evaluation session in the affect and non-affect condition.

In both conditions student's were mainly in flow (affect: M=58.12, SD=22.23; non-affect: 52.98, SD=17.41). This was followed by confusion (affect: M=28.77, SD=23.28; non-affect: 27.36, SD=18.21) and boredom (affect: M=9.54, SD=13.33; non-affect: 16.08, SD=7.45). Only a few were frustrated (affect: M=2.01, SD=3.15; non-affect: 1.54, SD=2.36), surprised (affect: M=1.03, SD=1.83; non-affect: 0.74, SD=2.07), or delighted (affect: M=0.53, SD=1.33; non-affect: 1.19, SD=2.50).

T-tests were conducted for each affective state. A significant difference between the groups was detected on boredom. Students in the affect condition were significant less bored than students' in the non-affect condition, t(40)=- 2.14, p<.05, d=.59. On all the other affective states no significant difference between the groups were detected.

Task behaviour

Figure 3 shows the different type of behaviour that occurred during the evaluation.

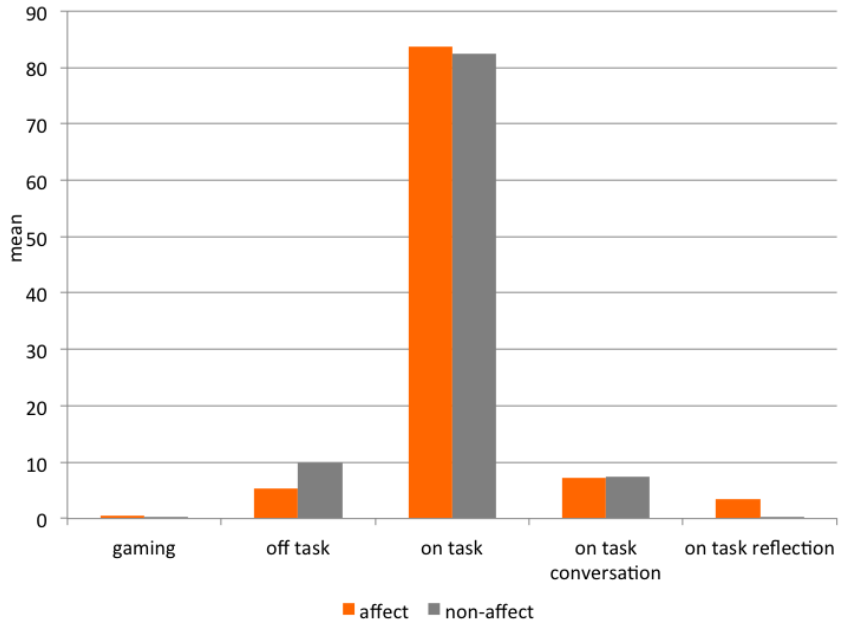


Fig. 3 Students' task behaviour during the main evaluation session in the affect and non- affect condition.

In both conditions, students were mainly on task (affect: M=83.58, SD=13.33; non-affect: 82.42, SD=8.29). Fewer students' did have an on task conversation (affect: M=7.24, SD=7.86; non-affect: 7.36, SD=6.02), were off task (affect: M=5.39, SD=6.48; non-affect: 9.87, SD=6.03), or reflecting on the task (affect: M=3.38, SD=9.86; non-affect: 0.23, SD=0.75). Very few were gaming the system (affect: M=0.41, SD=1.45; non-affect: 0.12, SD=0.55).

T-tests were conducted for each task behaviour. There was a significant difference on students' off task behaviour. Students in the affect condition were significant less off task than students' in the non-affect condition, $t(46)=-2.46$, $p<.05$, $d=.71$. On all other task behaviours no significant difference between the groups were found.

Affect, task behaviour and performance

Based on the pre- and post-test questionnaire, students' were given scores according to how well they answered questions about fractions. In order to investigate if there is a relationship between affect, task behaviour and performance, we correlated the variables from the HART data with the post-test scores.

There was a significant positive correlation between the affective state of flow and the post-test score ($r=.307$, $p<.05$). Additionally, there was a significant negative correlation between off task behaviour and post-test score ($r=-.349$, $p<.05$).

Performance and perception

Figure 4 shows the students' performance when answering fractions tasks before and after they have used the learning environment in the different conditions.

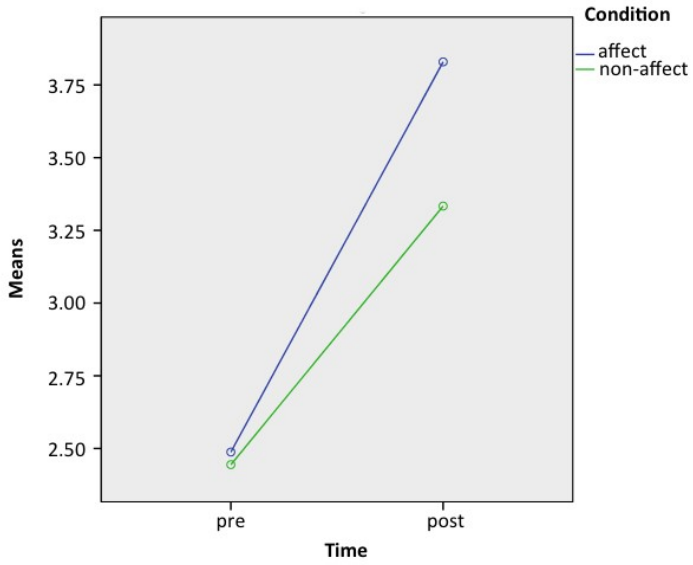


Fig. 4 Student's learning gains in the affect and non-affect condition.

In the affect condition students increased their knowledge of fractions from M=2.49 (SD=1.65) to M=3.83 (SD=1.46). In the non-affect condition students increased their knowledge from M=2.44 (SD=1.58) to M=3.33 (SD=1.71). An ANOVA repeated measures showed a significant increase of knowledge in both groups ($F(1,75)=43.94$, $p<.001$, $\eta 2=.369$).

Although, the difference in learning gains between the groups was not significant ($F(1,75)=1.81$, $p>.05$, $\eta 2=.024$), the overall tendency of the affect condition showing higher learning gains warrants further investigation.

In addition to students' performance of answering fractions tasks, we also asked students in the post-test if they found the learning platform interesting (rating scale 1: don't agree to 5: totally agree). Students in the affect condition found the platform on average more interesting than

students in the non- affect condition (affect: M=4.27, SD=0.74; non-affect M=3.80, SD=1.14). This difference was significant ($t(62)=2.22$, $p<.05$).

Conclusion:

During our evaluation student's affective states were annotated while they were using the system in either conditions. The results show that in the affect- aware condition students were less bored than students in the non-affect condition. Additionally, students' in the affect condition showed significant less off task behaviour then students' in the non-affect condition. These are important findings as boredom and off-task behaviour can have a negative impact in learning. Although our results show only a small difference between the affect and non-affect condition on learning gains, the improvement in higher learning gains in the affect condition are promising. Combined with anecdotal evidence that suggest students' higher level of engagement and reduced 'maths anxiety' indicate the potential of affect-aware student models at the heart of adaptive environments.

Reference

1. J. Ocumpaugh, R.S.J.d. Baker, and M.M.T. Rodrigo. Baker-rodrigo observation method protocol (bromp) 1.0. training manual version 1.0. Technical report, New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences., 2012.

Study report 3

Students' confidence levels in learning fractions

Date(s): M33

Participants:
12 Year 4 and 5 students (8-10 years old)

Aim(s) of study:

- To identify the extent to which iLearnFractions impacts on students' confidence in learning fractions, including:
  - equivalent fractions
  - adding fractions
- To identify what factors impacted on students' confidence in learning fractions

Method:
12 students were interviewed separately before the Part A summative evaluation and after Part B in Moorside Primary School, UK. During the interviews they were asked to rate their confidence (out of ten, with ten being the highest) in learning fractions, equivalent fractions and adding fractions. The students were not reminded of their pre-interview score during the post-interviews.

Results/findings:
"How has the platform helped you learn about fractions?"

Indicative comments about equivalence	**Indicative comments about addition**
- If you had 2/4 and a 1/2 next to it you could see the shapes shaded and compare them. - I got more confident with it and how to find equivalent fractions. It would go up, like 8 + 8 + 8. - It helped me because it was kind of like a chart when you split it into small bits. You double it and the denominator becomes bigger but the bits become smaller. - You could add 1 and 1 and 1 or 4 and 4 and 4. What you could also do was your times tables. - You could check equivalent fractions and it would tell you which was bigger and which was smaller.	- When you do the adding you click them and they merge together; that's quite good if you're learning to add or subtract fractions.
	Indicative comments about fraction size/comparison
	- The compare thing was good, so you wouldn't get told it, you'd do it yourself to see if your method was right. - I've learned about something if the numerator is larger than the denominator it is over a whole. I've learned that from the platform.

Comments about the platform overall

- It is more fun learning fractions on a computer than sums, seeing a different side of fractions really.
- It goes through with you how to do it without telling you so that has made me feel more confidence and it has made me feel good about fractions.
- I couldn't really add it before and now we've been working with IT fractions I'm a lot better at it.
- Using the platform, I couldn't really add fractions together and I didn't know how to and that helped me to do things I didn't know to do and I've become more confident after practising.
- I have improved. You know when in italk2learn it really makes things fun and it really explains things well when the robots come up.
- If you got it wrong it showed you the methods to do it.
- You can put fractions into FL and it will check them for you and you can learn what you've done wrong.
- Because with the system you can check and work out and check and see what you've done by changing it.
- The tasks were quite fun.
- Because using it has made it easier to understand fractions and it has made it easier to add and subtract them and things.
- When we have mental maths it is sometimes about fractions and I'd think 'oh no, do we have to' because I'm just not that confident about it but now that I've done iLearnFractions I know, because I've had some practice, that I might get it right this time. So I'm saying I'll put my hand up and give the answer instead because I'm feeling more confident with it.

- If you got it wrong then it wouldn't just give you the answer, it would explain how you could get the right answer. so it gave you some time and also gave you some advice.

Confidence data from interviews:

	Confidence in fractions			Confidence in equivalent fractions			Confidence in adding fractions		
No.	**Pre-**	**Post-**	**Diff**	**Pre-**	**Post-**	**Diff**	**Pre-**	**Post-**	**Diff**
100	6	8	+2	9	9	0	9	9	0
110	7	8	+1	8	8	0	7	7	0
113	9	10	+1	0	6	+6	0	10	+10
114	6.5	8	+1.5	3	5	+2	0	5	+5
115	7	9	+2	8	8.5	+.0.5	6	9	+3
125	7	9	+2	4	4	0	5	8.5	+3.5
216	7	9	+2	8	9	+1	9	10	+1
300	7	9	+2	5	5.5	+0.5	9	10	+1
301	7	8	+1	8	9	+1	7	7	0
312	3	7	+4	1	5	+4	6	[not asked]	N/A
313	6	8	+2	7	6	-1 *	5	7	+2
314	10	10	0	10	10	0	10	10	0
Total	**82.5**	**103**	**+20.5**	**71**	**85**	**15**	**73**	**92.5**	**29**
								/11	**/11**
Mean	**6.875**	**8.58**	**+1.71**	**5.92**	**7.08**	**+1.25**	**6.08**	**8.41**	**+2.95**

* The one student whose confidence reduced explained, "When you do equivalent fractions it is quite hard to find another equal fraction to do it. In the IT suite when we did iTalk2Learn it had a few challenging questions and I didn't really like it."

Conclusions:

Although we are aware that we cannot wholly rely on self-reported data, the findings from the interviews suggest that the platform generally had a positive impact on the students' confidence levels. On all occasions (bar one), the students self-reported increased confidence having used the platform. The mean averages show an overall increase in students' confidence in learning fractions (+1.71), fraction equivalence (+1.25) and addition (+2.95). Overall themes that emerged included:

Manipulating visual representations

Students reported being able to change fractions if they had made an error (e.g. 3/5 instead of 3/4) or to make an equivalent fraction using the built-in tools of Fractions Lab.

Trying things out and check their answers, then try again if necessary

Trial and error was mentioned by several students as a reason for their increased confidence in learning fractions.

Seeing equivalent fractions as related through multiplicative reasoning

By using the 'find equivalent' tool, some students began to articulate patterns they saw when making equivalent fractions.

Carrying out 'fun' tasks

From this anecdotal evidence it appears that the students who saw their platform experience as 'fun' may also gain further confidence. The one student whose confidence was reduced by one score reported how he found the tasks challenging.

These findings support others' work in this area (Ben-Naim, Marcus, & Bain, 2008; Hansen, 2008; Mavrikis, Gutierrez-Santos, Geraniou, & Noss, 2012). Our next task is to triangulate these findings with the data from Part A and the data from Part B in order to see if they reflect the larger picture.