In the Mood for Learning: methodology

Ivon Arroyo, Benedict du Boulay, Ulises Xolocotzin Eligio,
Rosemary Luckin and Kaśka Porayska-Pomsta

CSRP 603

November 2011

ISSN 1350-3162

University of Sussex

Cognitive Science
Research Papers

Introduction

This technical report arose out of a workshop on Evaluating Affect organised by Rosemary Luckin and held at Bath in 2006. The workshop was concerned with the many interactions between affect and learning, and particularly with the implications of these interactions for the design of technology enhanced learning. One of the outputs of the workshop was a cooperative writing project leading to a book with the working title of “In the Mood for Learning”. Only some of the material for this book has been produced, so it was decided, as an interim measure, to make the existing chapters available as a technical report.

There are three chapters. In the first Kaśka Porayska-Pomsta and colleagues consider the methods that are available for studying learner affect and for formalising the results of such studies in computer systems. In the second chapter, Ivon Arroyo and Benedict du Boulay explore the design issues that arise when building technology enhanced learning environments that take affect into account. Associated with this second chapter are a number of case-studies that look at specific examples of educational interactions where affect plays a key role. These include contributions by Ivon Arroyo and colleagues, Ryan Baker, Winslow Burleson and colleagues, Amanda Carr and colleagues, Sydney D’Mello, Cristina Conati and Manolis Mavrikis. In the final chapter Ulises Eligio and colleagues describe collaboration quality and emotional sharing around learning technologies from studies of children using concept map tool and a computer game.

1. CONTEMPORARY METHODS FOR AFFECT-RELATED KNOWLEDGE DISCOVERY AND EMOTION MEASUREMENT IN LEARNERS

Kaśka Porayska-Pomsta[1], Manolis Mavrikis[1], Sidney D'Mello[2], Cristina Conati[3] and Ryan Baker[4]
[1]London Knowledge Laboratory, Institute of Education, London, UK
[2]Institute for Intelligent Systems, University of Memphis, Memphis, USA
[3]Department of Computer Science, University of British Columbia, Vancouver, Canada
[4]Department of Social Science and Policy Studies, Worcester Polytechnic Institute, Worcester, USA

Introduction

This chapter of the Report considers the methods that are available for studying learner affect and for formalising the results of such studies in computer systems. Although most of the research methods that have been used before the advent of technology and affective computing are still in use (Coan & Allen, 2007), technology has brought with it new ways of studying the phenomena in question and, indeed, new questions. It has also opened the possibility of using the established methods in new ways, often in combination with emerging methodologies from data mining and machine learning. One of the major attractions of using technology to study affect in relation to learning is the fact that it allows us to build real-time dynamic models of affect in educational interactions (e.g. Conati & Zhou, 2002) to log such interactions and to test the models repeatedly and systematically.

This chapter intends to provide an introduction to the continuously changing methodological landscape of the current state-of-the-art in the field and ultimately to serve as a starting point for a broad spectrum of readers' methodological decision making in their own endeavours. The sections in this part of the report are also intended to illustrate the existing tensions between the different research perspectives and to demonstrate how these differences may be used constructively as a vehicle for a comprehensive exploration of the field.

Description of key methods

Researchers often have a wide choice of different methods through which to study learner affect. The choice that they make needs to be guided by the questions that they ask and the kind of technology that they want to design. Current research in the area can be broadly described in terms of two overarching goals that relate to (1) detection of learner affect and (2) acting on learner affect. The goals are related in that in order to be able to act on learner affect, it is necessary to know what the learner is experiencing at any given point, while choosing how to act involves the consideration of how the specific action will impact the learner's affective and cognitive states. When translated into a typical design of a learning environment, the goals are motivated respectively by the need to inform the learner modelling component responsible for tracking the learner's states, and by the need to know how to respond to the states that are being detected. Both goals are concerned with providing guidelines about the best design of technology-enhanced learning experiences, whereby “best” is measured typically by the effectiveness of the learning outcomes.

The two overarching goals are approached by different researchers in a different ways and typically involve a combination of traditional research tools such as questionnaires, self-reports and control measures, as well as purpose-built computational tools for accurate real-time data capture, such as interaction and decisions logs, and physiological sensors. Again the choice of a specific tool or a combination thereof depends on the exact research focus. For example, detection of learner affect often requires the researcher to identify the affective states that they want to model, to define them and to categorise them. In order to do this, it may be sufficient to rely initially on qualitative approaches such as self-reports from the learners or on teachers' annotations of the video and audio recordings of learners engaging in specific educational interactions. However, in order to be able to detect learner affective states in real time, additional tools are often required, such as physiological sensors and sophisticated inference mechanisms that are based on accurate information such as linguistic cues in learners' verbal responses, mouse and keyboard actions, time spent on task, information about help-seeking behaviour, etc. A further consideration relates to how to act on the detected states to enhance the learning experience for individual learners and its effectiveness. Although often separated from the question of how to detect learner affective states, acting on affect is crucial because it inevitably influences the learner's states during an interaction. Relying on human tutors/teachers as the source of information in relation to both detecting and acting on the diagnosis of their learners' affect, is the obvious option. However, accuracy of human tutors' inferences, effectiveness of their actions, their experience, and the ecological validity of the environment in which the data is collected are all, but not exclusively, issues to be considered, because they impact on how a study of affect will be conducted and how informative it will be.

In the following sections we describe the most common methods currently used to study affect for educational human-computer interaction. We show specific example studies that were conducted to illustrate the different methods or a combination thereof in-depth.

Types of instruments used to study affect

Most methods described in the following sections rely on a variety of instruments to measure affective states. Different types of questionnaires, including structured and semi-structured questionnaires, are often used to measure learners' overall affective and motivational **traits**, i.e. long-term affective characteristics (e.g. Gardner, 1983; Izard, 1991; Midgley et al., 2000; O'Bryen, 1996; Whitelock & Scanlon, 1996). Questionnaires are relatively easy to both design and administer and can be delivered online without the necessity of physical presence of the researcher. Typically, questionnaires are not appropriate for eliciting knowledge about affective **states**, i.e. *short-term*, transient characteristics of learners, which, ideally, need to be determined in *the-heat-of-the-moment*. Questionnaires do not lend themselves to such real-time knowledge elicitation, because, typically, they are administered in non face-to-face contexts, away from the immediate context of the situation studied. Furthermore, obtaining fine-grained reports on the specific emotional experiences of the learners often requires very precise, individual learner-dependent question formulation in order to enable the researcher to understand the specific nature of the emotions experienced by any given learner in a particular situation. Again, obtaining such detailed information is not feasible through questionnaires, which, normally, are aimed at gathering information from a multitude of learners simultaneously and contain open-ended questions, the answers to which may be difficult to reconcile and to classify systematically across the participants. Gaining access to learners' transient affective states, which is necessary to inform the specific pedagogical and TEL designs and which constitutes the focus of contemporary research in this area, requires the use of tools that allow the learner or the observer either to report on their states while they are engaging in a task or that are able to monitor the changes in the learners' behaviours in real-time. In relation to the first, there are a number of tools that are currently used to elicit knowledge about learners' affective states and many have been motivated by psychology research, where they are often categorised as (a) *forced-choice response* and (b) *free response* tools (Scherer, 2005). In relation to the second, the advent of physiological sensor technology has allowed to monitor the learners' physiological behaviours and to infer at least some of the information about their possible emotional states along with the changes in those states through relatively and increasingly unobtrusive means. Methods relying on physiological sensors are reviewed in a separate section later in this chapter.

Reporting tools: eliciting information about learners' emotions

Tools that are used in contemporary research in eliciting information about learners' emotional states are now discussed. The two main categories under which such tools are classified are: (a) forced-choice response and free-response tools respectively.

Forced-choice response tools are further categorised in terms of *dimensional* and *discrete response tools*. *A* ***discrete response*** tool provides the reporters with a pre-defined list of words describing the affective states of interest to the researcher (e.g. Cowie, 2005). The reporters are asked to rate their emotions along nominal, ordinal, or interval type of scales. For example, to ease the reporter's task, qualitative variables such as "little", "a lot", or Likert-like scales, can be used to report the degree to which an emotion were experienced by the learner. Although the purpose of relying on forced-choice response is to ensure homogenous data and to ease the researcher's task of analysing the data, both the definitions and the tools may influence the resulting reports. It is difficult to gauge the extent to which the learners' understanding and labelling of the affective states and their emotional experiences, when engaging in a task, actually correspond to the predefined labels of possible affective states. Often, an introductory session is needed to align the learners' labels with those of the researchers. However, during the actual task, learners may want to report on emotions that are missing from the list, but forced-choice response tools do not facilitate this. Furthermore, the use of forced-choice response does not ensure that the predetermined labels do not influence learners to report an affective state that they would not report had they not been primed by the choices provided. Finally, the results from different studies may be difficult to compare across, when different scales and sets of labels are used by different researchers (Scherer, 2005).

The **dimensional response** approach relies on structured description of emotion within a dimensional space. There are several of these instruments and the most notable examples Feeltrace (Cowie & Cornelius, 2003), NTX Feeltrace (Reidsma, Hofs, & Jovanovic, 2005) the Geneva Emotion Wheel (Scherer, 2005), and the Affect-Grid (Russell, Weiss, & Mendelsohn, 1989). When using such tools, the reporters are expected to locate their emotional states within the space represented by one or more dimensions. For example axes, such as valence and arousal (Cowie & Cornelius, 2003) or pleasantness and unpleasantness (Larsen, McGraw, Mellers, & Cacioppo, 2004) have been used as separate dimensions to obtain affect measurement. The Geneva Emotion Wheel arranges emotions in two-dimensional space and by denoting the distance from the origin represents the intensity of the associated feeling. Whilst, dimensional response approach allows the reporters to talk about affect in a systematic way, it can be still difficult for them to relate their affective states to the abstract dimensions. Moreover, this approach ignores the possibility that some states cannot be differentiated using general dimensions and that depending on the task at hand additional dimensions may be needed.

In contrast to forced-choice responses, **free response tools** – also referred to as self-reports – allow the participants to talk freely about their affective states as they are being experienced by them. In general, many researchers found that a composite approach consisting of affect labels, dimensions of emotions and free responses tends to provide the most consistent and defensible way of obtaining information about learners' affect. In the following sections we discuss the particular advantages and disadvantages of using self-reports, however for a comprehensive discussion of the issues involved in using the free-response methods for measuring emotions and for information about the procedures required to standardise the reports, the reader is referred to Scherer (2005).

Learners' self-report and retrospective annotation of affect

Learners' self-reports represent one of the commonly used methods for accessing learners' affective states. This method often relies on free-response measurements and questionnaires. Specifically, self-reports involve asking the learners about their experiences either as the learners experience them during a learning activity (*concurrent* self-reports) or after a learning activity is completed (*retrospective* self–reports). Student self-reports aim to elicit information about the learners' emotional and motivational states during a learning activity. The information gleaned through them can provide an intimate insight into the emotions experience by the learners and can be used to qualify data obtained through other means, e.g. physiological sensors. Although intuitively appealing and often very useful, this method needs to be applied with caution because it can interfere with the very emotions that it tries to gauge. Data obtained through self-reports is subjective in nature and depending on the learners' cognitive load at the time of reporting, their ability to express and remember their emotions, the length of time elapsed between the learning episode and reporting, as well as their multitasking abilities in case of concurrent reporting (they may have to report at the same time as engaging in learning) it may need to be validated by other sources of information. To ensure as objective results as possible, it is often best to combine this method with other sources of information.

Concurrent self-reports traditionally involve *think-aloud* or *talk-aloud* free-response measurements that allow the learner-reporter to verbalise their cognitive processes as successive states that are experienced during a learning activity (Ericsson & Simon, 1993). Specifically in relation to students reporting on their affect while engaging in an educational task D'Mello et al. (2006) refer to this method as ***emote-aloud***. Simply put, students express their emotions verbally as they perform a learning task.

Video recording of the learners engaging in the learning tasks can also be used for offline analyses. The video data is particularly valuable for emote-aloud as a way of providing additional information about the learners' emotional states that may be inferred from their verbal and non-verbal behaviours such as linguistic responses provided to the teacher, tone and amplitude of spoken utterances, facial expressions, posture as well as eye-gaze.

Retrospective reporting involves the learner-reporter accessing their memory trace of the information heeded successively while the learner was on task, *after* the task is completed. This is usually achieved by involving participants in an audio- or video-stimulated recall interview. Retrospective reporting can be elicited in a similar way to the emote-aloud, but it offers both the reporter and the researcher the opportunity to focus and to elaborate on specific aspects of the observed behaviours that may be of particular interest given their research questions. The instruments used for recording retrospective reports typically include structured or semi-structured interviews and thus can involve either free responses or can rely on appropriately designed questionnaires, dimensional emotion responses, or discrete emotion response (as outlined in section on the 'Types of instruments used to study affect'). These are often accompanied by video and audio recordings of interviews and it is usually good practice to rely on such recordings to ensure consistency and completeness of data, as well as to facilitate data triangulation.

To date, concurrent and retrospective self-reporting methods have been used predominantly to support the theoretical development of models of learner affect (Reinhard Pekrun, 2006; R. Pekrun, Goetz, Perry, & Titz, 2002) or to validate such models (Op 't Eynde & Turner, 2006). In more recent approaches data collected from self-reports have been used as the basis for developing computational learner models using machine-learning (S. K. D'Mello, et al., 2006; Mavrikis, D'Mello, Porayska-Pomsta, Cocea, & Graesser, 2010).

Self-reports can be facilitated directly within intelligent learning environments (ILEs). This approach can inform the method by which the specific ILE can adapt its feedback. Eliciting the reports in ILEs can be achieved through an appropriately designed interface, for example involving sliders or drop-down lists (see Case Study 2 in Box 1). It is important, albeit non-trivial, that the reporting tools are incorporated into such an environment in a seamless way, that is, a way that further supports, rather than interferes with, the process of learning. One of the more sophisticated methods that aims to incorporate the self-reporting task into a learning process is open learner modelling (OLM). Similar to traditional learner models, OLMs are used to track and reason about the learner's progress, as well as their cognitive and affective states. Additionally they also make available the results of such tracking and reasoning to the learner who, in some cases, can inspect and dispute the correctness of the system's inferences about them. There are numerous examples of OLMs being used, which employ different forms of visualisations to represent the student models generated in real-time. The different representations vary between simple skillometers to complex tools that allow the student to negotiate and change the content of the model (Mitrovic & Martin, 2002). Engaging learners in self-assessment as they learn facilitates access to their underlying mental states and it promotes their meta-cognitive skills, which are crucial to effective learning in general (Aleven, Koedinger, & Cross, 1999; Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Swanson, 1990).

Self-reports have been used also in the classroom context to enable a teacher to react according to the affective characteristics reported by the learners on a moment-by-moment basis. For example, Alsmeyer et al. (2007) provided school learners with hand-held devices to allow them to communicate their emotional experience to the teacher privately. This has given the teacher access to each pupils' emotional states at any point reported and thus, with an informed basis for understanding and responding to the learners' emotional experiences.

Advantages and disadvantages of using self-reports

Superficially, self-reports may be seen as relatively easy to create and to facilitate in the context of learner affect. The advantage of the free-response approach is that it enables the students to provide descriptions of their feelings in their own words: the labels used are not selected from a list of predetermined choices (Scherer, 2005). The fact that the report is personally meaningful to the learner provides a more accurate account of their emotional experience. However, employing this method can be time consuming and labour intensive for the researcher, as it usually requires the researcher to be present during the learning and reporting sessions. Such commitment may therefore impact on the number of reports that will be realistically generated. Introverted or shy learners will have greater difficulty in providing self-reports and may require additional prompting or specially designed materials to help them in the process. Such prompting may result in the greater level of intrusiveness and may bias the resulting reports. A further difficulty will be the ease with which the different emotion categories generated through self-reports may be reconciled across participants. The same emotions may be labelled differently by the participants and similarly, different emotions may be referred to by the same names. Therefore, post-processing of the data collected would be required. Nevertheless, despite the unavoidable difficulties that some learners with low meta-affective and reflective abilities will have in providing reports, the opportunity that this method offers of providing access to more truthful, non-primed statements of their emotional states than would be possible otherwise, renders it a very useful way of enriching the data gathered and of verifying the interpretations thereof.

Caution is advised in conducting quantitative analysis of the data collected, because interpreting and classifying the verbal reports is prone to bias (Scherer, 2005). Self-reports can be limiting for both researchers who employ solely these methods as research instruments, and for system designers who employ them as a basis for adapting the behaviours of the systems that they develop. This is because, as de Vicente (2003) among others relates, many researchers are anxious that the sheer act of learners reporting on their affective and motivational states may impact the very motivation and emotions. To ask someone to engage in a meaningful learning experience at the same time as to self-analyse places a significant cognitive demand on them – known as *cognitive load* – which may result in neither of the tasks being performed optimally. Furthermore, by attaching labels to specific emotions, the reporters may in fact reinforce or evoke such emotions in situations where they were experienced mildly or not experienced at all. Unfortunately, the extent to which this may happen or if indeed it happens is not known nor can it be easily measured. This limits the potential validity and reliability of the information gathered in this way. While this is a well-known limitation of all self-reporting methods, it is especially evident in relation to concurrent self-reports (Ericsson & Simon, 1993). The cognitive load is reduced, to an extent, in retrospective self-reports because the learner can engage with a learning task first and report on their emotions during that task later. However, a noteworthy limitation of retrospective self-reporting of affect is the distance between the time at which the learner is engaged in a task and the time of their report. Not only the affective states reported may be different from the ones experienced during the task (Masthoff & Gatt, 2006), but also, the learner may not remember exactly what they felt at specific points and their memories may be biased by post-hoc rationalisations of the experiences reported and whether or not the learning task was completed successfully. One approach to resolve problems associated with both approaches is to further clarify the responses collected during the task through *post-hoc* discussions with the learner. This method is often referred to as *retrospective* or *post-hoc walkthroughs* and requires access to a record of the learning episodes. Such record may consist of video- and/or audio-recordings of the learner engaging in a learning task, or if ILE is used, the recording of students' screen which may be synchronised with any video- and audio recordings of the learner, along with any verbal protocols collected from the learner while they engaged in a learning task.

Type and quality of data resulting from self-reports

Both concurrent and retrospective reports result in data that is subjective in nature and which may reflect the reporters' theories on what their affective states are or should be, rather than uncovering their actual states. Crucially, the quality of the data obtained relies heavily on the reporters' meta-affective skills, as well as their confidence and ability to verbalise their affective experiences. The quality is also dependent on the type and subtlety of the emotions to be reported, as well on the age of the reporters. In particular, studies with children suggest that the younger the learners are the more difficult it is to elicit coherent talk-alouds or self-reports from them and the more unreliable the reports may be (e.g. Conati & Maclaren, 2009). This is also conditioned by the fact that children's understanding of emotions is very crude until approximately the age of 8 and there are big differences between the specific age groups' abilities to recognise, categorise and label their own and others' emotions into fine-grained affective categories (Sayfan & Lagattuta, 2008). This means that any methods used for eliciting self-reports from young children must be adapted appropriately to their cognitive and affective capabilities. This typically means adjusting the questions asked and providing additional tools, such as pictorial representations of emotions (J. Read, MacFarlane, & Casey, 2002; J. C. Read & MacFarlane, 2006), for communicating their feelings. For younger learners information obtained through physiological sensors, such as eye-tracking devices or wrist-bands that measure heart rate and its variability (see the section on 'Physiological and behavioural sensing) may provide the most reliable insight into children's affective experiences during learning.

Table 1: Summary of the self-reporting methods, the instruments used and their advantages and disadvantages

Self Report	**Concurrent**	**Retrospective**
Instruments	video- audio- Free response (*think-, talk-, emote-aloud*)	video- audio- Free response (interviews) Dimensional response Discrete emotion response
Involves	Reporting emotions *during* a learning task	Reporting emotions *after* a learning task
Advantages	Provide *heat-of-the-moment*, stream of consciousness reports	• Allows to elaborate and focus on details; • Reduces cognitive load (in comparison with concurrent self-reports); • Easy to prepare, administer and elicit
Disadvantages	• Imposes high cognitive load; • Requires: • good meta-cognitive skills; • ability to coordinate engagement in task and self analysis; • May influence the emotions experienced Generates subjective data	• Requires more time per subject (time to engage in the task + time needed to elicit self-reports); • Imposes cognitive load (but less than concurrent self-reports); • Creates distance between the time of engagement in the task and self-reports; • Requires the reporters to have significant meta-cognitive skills; • Generates subjective data
For adults	Yes	Yes
For children	No	Possibly for older children, with appropriately designed tools

Case Study 1: WaLLiS

Mavrikis et al. (2007) used self-report during video-stimulated recall interviews. Learners were presented with replays of their own interactions with an Interactive Learning Environment called WaLLiS. Learners were asked to try to explain their actions and to report on their affect at specific points during their interactions. Explanations were elicited using semi-structured questionnaires in relation to affective and motivational factors such as learner's confidence, interest and effort. These were reported using a combination of discrete and dimensional emotional response. Learners were free to report what they were feeling during the task, but they were also asked to report specifically on whether their confidence, effort and interest were increasing, decreasing or remained unchanged after specific interaction points. To overcome the common difficulty of the distance between the interaction and the reporting time, the participants were asked to keep notes (a reflective journal) during or immediately after their interaction with the online tutor. To this end, students were provided with questionnaires that could be used as memory cues during the retrospective reports.

Case Study 2: MOODS

Self-reports have also been used to gather information about learners' affective and motivational states in real-time as the students interact with Intelligent Learning Environments (ILEs). For example, de Vicente (2003) describes the GUI of a tool which allows the students to report on their motivational states during a task of learning Japanese numbers, in terms of pre-specified categories (satisfaction, sensory interest, relevance, cognitive interest, confidence, and effort) by moving and placing a slider, provided for each motivational category, anywhere between the fuzzy-linguistic terms of low and high. Students can use such reporting tools either throughout their interaction with the software or at appropriate points, for example between activities, or before logging out from the system. de Vicente focused on the usefulness and usability of such tools within ILEs. He found that although the learners judged the tools as usable, embedding such tools in the learning environment leads the learners' to expect that their reports will affect the behaviour of the system. That, in the case of the MOODS system, the learners' self-reports had no impact on the behaviour of the learning environment, was reported as disappointing and could lead to frustration and demotivation in the longer term.

Box 1: Two case studies illustrating the use of the self-reporting method to elicit knowledge from the learner. Case study 1 illustrates the application of semi-structured interviews and reflective journals. Case study 2 demonstrates the use of online tools, such as sliders and multiple-choice questions facilitated by an appropriately designed graphical user interface.

Concurrent and retrospective annotation of affect by annotators

An alternative method for measuring affect is for an independent annotator to provide a report on learners' affect as they were using a learning environment or participating in a learning task. Such reports are often and interchangeably referred to as *annotations* or *codes* and the reporting task is referred to as the *reporting* or *coding* task. An annotator can either be a researcher, a peer student, a tutor or an individual with appropriate experience or training. The annotations can be conducted either concurrently with the learning task or retrospectively.

Learners' interactions are coded for affect using similar instruments as the ones described in the section on 'Types of instruments used to study affect', depending on the needs of the research being conducted. For example, semi-structured questionnaires can be used to prompt the annotator to elaborate on their reports. Alternatively, affective categories can be selected based on a theory and empirically gained information about those emotions of the learner that are important in the context being studied. In general, the more structured the type of information requested, the more data will be collected and analysed and the more amenable the data will be to a computational or statistical, rather than solely qualitative, approach.

In the **concurrent** version of the method with pre-determined categories, one or more annotators observe students engaging in a specific learning task and code for the target affective categories. The learning task may involve traditional, i.e. human-human, or technology-enhanced learning (TEL) interaction. In contrast, in **retrospective** coding learners are video-recorded while they perform the task. The video recordings can include close-ups of learners' faces along with any audio data generated during the learning sessions (e.g. sighs, gripes, etc.). When the educational intervention is administered through a computerised environment it is usually advisable also to record a video of the learner's computer screen so that the context of the educational session can be considered in the retrospective judgments of affect. As with retrospective self-reports, the retrospective affect judgment protocol commences *after* the educational session is completed. Particular attention needs to be paid when recording multiple sources of video in order to enable their retrospective synchronisation and to facilitate the annotator's task.

Both concurrent and retrospective methods can facilitate detection of any occurrence of a noteworthy affective state at pre-chosen important moments, or at regular or random intervals. In the first variant, annotators are instructed to volunteer judgments during emotionally charged episodes. For instance, in a study using the retrospective method designed to analyse the emotional reactions of students playing an educational game to teach number factorisation (Conati, Chabbal, & Maclaren, 2003), annotators were asked to report whenever they could detect any of the four specific emotions of interest, or general states of positive/negative affect. A similar retrospective protocol was used in a study investigating student affect during tutoring sessions with expert tutors (Lehman, Matthews, D'Mello, & Person, 2008) while de Vicente and Pain (2002) employed semi-structured questionnaires and asked postgraduate tutors to elaborate on their reports especially in relation to the evidence they relied on when making their diagnoses.

Alternatively, observers can be asked to make affective judgments at strategic points in the session (e.g. after a specific type of system's intervention) (S. D'Mello, Taylor, Davidson, & Graesser, 2008; A. Graesser et al., 2006). Within this variant, observations are often made at randomly selected points as well. The affect judgments at random points can serve as a control to the judgments at the theoretically selected points. In a third variant, observers make judgments at previously selected intervals and in a pre-determined order (S. K. D'Mello, et al., 2006; Rodrigo et al., 2007), giving evidence on the absolute frequency of different affective states, and their temporal dynamics.

When observations are fixed (as opposed to spontaneous), they are generally set to occur during a pre-determined observation period. Twenty seconds is the most common length found in the literature. While justification for the time window size is not often given in details, it seems reasonable to suppose that 20 seconds is long enough to be able to make valid judgments, without frequently seeing multiple affective states in one observation. This time interval is also convenient to work with. In some cases, the first affective state observed is the only affective state coded, for simplicity in later analysis; in other cases, all affective states observed during the observation period are coded, in order to get the richest possible picture of events. Note that calculating inter-rater annotation agreements crucially depends on the same intervals being judged by multiple annotators, and therefore researchers managing such annotations should be aware of the importance that their interval choices will have in ensuring the validity and reliability of their conclusions.

Concurrent variants of the method are especially suitable for obtaining affect information about a genuine classroom setting where either a learning environment is used or the learners engage in a formal educational task (e.g. Rodrigo et al., 2008; Rodrigo, et al., 2007). They build on prior methods for concurrent behavioural observation by researchers in classrooms (e.g. R. S. Baker, Corbett, Koedinger, & Wagner, 2004; Karweit & Slavin, 1982; Lahaderne, 1968; Lee, Kelly, & Nyre, 1999; Lloyd & Loper, 1986), largely duplicating these methods but changing the construct coded from behaviours (such as *off-task* behaviour and *gaming* the system) to affective states. Live observations of affect are less common in laboratory settings (though examples do exist in the literature – e.g. S. Craig, Graesser, Sullins, & Gholson, 2004), because in these settings it is usually substantially easier to capture all the relevant information on video and thus rely on the higher flexibility afforded by retrospective annotations.

When collecting live affect annotations in a classroom setting, observations are generally conducted using peripheral vision in order to make it less clear to the learners observed exactly when an observation is occurring. The purpose of this is to avoid, as much as possible, impacting or inhibiting the learners' behaviours. For related reasons, "warm-up" sessions are often conducted, where no actual data is collected, but observers are present in the classroom, taking notes, in order to accustom learners to the observers' presence. These methods are used to attempt to reduce the degree to which learners inhibit or suppress their affective expressions, though there is always some risk of this in live observation. This issue is considerably less critical in laboratory settings where it is possible to hide the observers behind a one-way mirror.

One of the most important methodological choices is whether to conduct concurrent or retrospective annotations. A challenge that relates to both approaches is that of selecting suitable annotators. This challenge is discussed in the following section.

Selecting annotators

It is often assumed among researchers that the closer the annotator is to the study population, the more likely they are to be able to identify the study population's affect correctly. This means that, at least in theory, judges who are themselves learners should be the best qualified to interpret the affective experiences of their peers. However, there is some evidence that learners are not, in fact, very good judges of emotions experienced by their peers (A. Graesser, et al., 2006) with some experts claiming that the ability to detect emotions accurately requires considerable teaching or tutoring experience (Goleman, 1995; Lepper & Woolverton, 2002) or prior training in assessing emotions (Sayette, Cohn, Wertz, Perrott, & Parrott, 2001).

There is also some evidence that cultural differences between observers and the learners might have an impact on the accuracy of the observers' judgments of learners' affective states. While, following Ekman et al. (1987), it is commonly assumed that expressions of basic emotions (e.g., anger, fear, sadness) are universal and are reasonably recognisable cross-culturally, there is also evidence that recognising affect in specific contexts is difficult across cultures. particularly among cultures with exposure to Western mass media (Russell, 1994). In particular, researchers, who used live observation method to study affect in classrooms, have found that observers from different national backgrounds than the study population sometimes achieve poor inter-rater reliability, when compared to observers from the same national background as the study population (personal communication, Mercedes Rodrigo). In addition, our experience has been that observers who are no longer learners themselves find it more difficult to assess learners' affect.

Expertise in the subject domain taught and, crucially, tutoring experience may have significant impact on annotators' ability to provide coherent, consistent, and accurate reports. For example, Porayska-Pomsta (2004) found that tutors' domain expertise impacts whether they focus more on the content (e.g. correctness of student answers, difficulty of the task) or pragmatic information (learners' hesitation in answering, learners' interaction styles, the way in which learners seek help, etc.). For those tutors who are not used to considering the pragmatics of the interaction explicitly, the reports often focus solely on the content, giving little information about the tutors' judgements and interpretations of the learners' behaviours and their underlying emotional causes. The more experience the tutor has in delivering teaching, the more likely he or she might be able to pay attention to the signs of changes in the affective states of the learner, especially if such states have negative valence Porayska-Pomsta et al. (2008).

Another important issue to consider when deciding on the appropriateness of annotators is their familiarity with the learning environment with which the learners are interacting and how easy it is for the annotators to interpret such interactions. For example, Mavrikis (2008; 2007) discuss that tutors may find it difficult to report on learner affect when watching replays of learners' interacting with a complex web-based environment for mathematics. In these studies, the variety of materials (multiple-choice questions, open learning activities), the length of the interactions (up to an hour), and the complexity of the learning environment – learners were able to solicit help from the system – raised questions as to the appropriateness of relying on tutors' expertise to report on a situation to which they are neither accustomed nor specifically trained to interpret. Although in previous research (e.g. de Vicente, 2003) this method seemed to be appropriate for use in a mock-up environment, where the correctness of student assessment was relatively simple and could be judged on the basis of 'right' and 'wrong' answers, the method may not be scalable to more complex learning environments.

Finally, in retrospective reports, the tools and the medium through which the materials are being presented to the annotators and the bandwidth of information contained therein have to be considered carefully. There are additional factors that can potentially impact on the ease with which annotators will be able to judge and report on student affect. For example, Porayska-Pomsta et al. (2008) found that when tutors are asked to judge learners' affect retrospectively, they often opt out from committing a judgement because they cannot reconcile the information that they can observe on the screen, e.g. learners' mouse movements, learners typing and deleting half-constructed responses, etc., with learners' verbal (typed) behaviours.

Advantages and disadvantages of live and retrospective annotations

One of the main advantages of the data collected from annotators, as opposed to the data collected from self-reports, is that so long as inter-rater reliability is validated, there can be fairly high confidence that all reports of an affective state involve that same construct (a construct is a conceptual variable that cannot be directly measured such as intelligence or emotion). By contrast, if multiple students report on their affect, it is difficult to be certain that they are all referring to the same construct. For instance, some students may have higher tolerance for frustration than others, and thus report the same actual emotional experience as frustrated or not frustrated. Additionally, applying this method is unlikely to have intentional bias, particularly if the observations are conducted by hypothesis-blind observers, or as part of an exploratory study with no explicit hypotheses.

External annotations may lack the "internal perspective" that learners' self-reports can give, or the long-term information about a specific learner's responses that a tutor or a peer familiar with the learner can provide. Other factors that impact the reliability of observer-based methods are the degree to which the learner displays his/her emotional expressions, and the intensity of the emotions to be recognised. If the learners choose to control their emotional displays, or if the intensity of the emotions is not strong enough to generate visible reactions, an external observer may systematically overlook some instances of learner affect. The more fine-grained the target set of emotions, the more difficult it is to tell them apart. For instance, in (Conati, et al., 2003) two observers were unable to consistently recognise instances of the four player's emotions of interest (pride/shame towards oneself, admiration/reproach towards a game agent) because often the two negative emotions and the two positive emotions were expressed similarly. However, better results were achieved in coding for positive vs. negative affect.

In summary, methods based on external observers' annotations are reliant on the attentiveness, and expertise of the observer and can be limited by circumstances that reduce the visibility of the learner's emotional reactions. However, these methods are particularly suited for achieving high ecological validity as their particular strength is that they can be conducted under realistic situations, whereas self-reports are typically used under the laboratory conditions or are generated *post factum*.

Live observations have the disadvantage of being very resource-intensive. It is recommended to have multiple observers, in order to both reduce the time between observations of a given student, and to enable calculations of inter-rater reliability. One approach used to deal with cases where appropriately trained personnel are limited in number, is to conduct an inter-rater reliability session with multiple observers early in the project, and in later observation periods to involve only a single observer (e.g. Lehman, et al., 2008). Another disadvantage of purely live observational measures is that it is difficult to test objectively the observed affective states at a later time. If no self-reports or other objective measures are collected during the learning activity annotations, the observed judgments of affect cannot be validated.

Retrospective reports from annotators also suffer from limitations. In particular, such reports often represent theories about the events presented *post factum* rather than being representative of diagnoses *in-the-heat-of-the-moment*. Moreover, it is quite difficult, in ecologically valid situations, to have access to the student's facial expressions and body language in addition to the recorded action during a given task. This means that in practice tutors have to rely on a limited bandwidth as discussed in the section on 'Selecting annotators'.

One advantage of the retrospective protocol is that it is easier to have multiple judges to participate in the affect judgment process. For example, peer learners can first judge the students' affective states, followed by trained judges, and expert teachers. Reliability among these different types of judges can then be computed in order to obtain a measure of construct validity (see next section). The ability to incorporate multiple perspectives on the learners' affective states represents an advantage of this methodology. Although, it is possible to include multiple judges in concurrent observations (Rodrigo, et al., 2008; Rodrigo, et al., 2007), this imposes additional logistic constraints, such as considerations about where the observers will be located, and may increase the degree to which students change their behaviour in response to the presence of observers.

Another important advantage of this method is that affective judgments can be solicited at theoretically relevant points that might be unknown during the learning session. For example, one set of judgments can be made at a given set of points to answer some theoretical question. At a later time, the videos can be recoded for affect at another set of points to answer new theoretical questions. Therefore, in contrast to live annotations, where the observation points must be decided *a priori* or in real-time, the retrospective protocols afford the possibility of reusing the data to test theoretical questions as they emerge. Table 2 summarises the two variants of the external-annotators method, the instruments required and their advantages and disadvantages.

Table 2: Summary of methods based on external annotation of affect, the instruments used and their advantages and disadvantages

<table>
<tr><th>External annotation</th><th>Concurrent</th><th>Retrospective</th></tr>
<tr><td><b>Involves</b></td><td>Observer annotating students' interaction with emotional judgements <i>during</i> a learning task</td><td>Annotator reporting on students' affective states <i>after</i> a learning task</td></tr>
<tr><td><b>Instruments</b></td><td>Dimensional response
Discrete emotion response
Think-alouds</td><td>video- audio-recordings
Free response (interviews)
Dimensional response
Discrete emotion response</td></tr>
<tr><td rowspan="2"><b>Advantages</b></td><td colspan="2">After validating inter-rater reliability, one can be fairly confident that, compared to student self-reports, data are referring to the same psychological construct.</td></tr>
<tr><td>Provide <i>heat-of-the-moment</i>, stream of consciousness reports</td><td>• Allows to elaborate and focus on details;
• Reduces cognitive load (in comparison with concurrent self-reports);Easy to prepare, administer and elicit</td></tr>
<tr><td rowspan="2"><b>Disadvantages</b></td><td colspan="2">- lack the "internal perspective" that a student's self-reports can give
- rely on the degree to which the learner displays his/her emotional expressions
- require observers to be able/trained to judge emotional experiences</td></tr>
<tr><td>• resource-intensive; multiple observers are required
• difficult to test objectively the collected results</td><td>• reports often represent theories about the events presented <i>post-factum</i> rather than being representative of diagnoses <i>in-the-heat-of-the-moment</i>.
• Potentially reduces the bandwidth of available information to the reporter</td></tr>
</table>

Case Study 3: Graesser et al (2006).

One example of retrospective affect reporting is the Graesser et al. (2006) study. In this study, 28 students were tutored on topics in Computer Literacy with an Intelligent Tutoring System. Videos of the students' faces and computer screens were recorded. These videos were subsequently used in three follow-up analyses that were inspired by the retrospective protocols. In one analysis, the students themselves engaged in a retrospective affect annotation task (retrospective self report) by watching their videos and providing emotion annotations. In a second analysis, untrained peers, trained annotators, and experienced teachers viewed the students' videos and provided affect judgments (retrospective annotation by external observers). Finally, in a third analysis, two experts in pedagogy analysed the videos at points when the learner was having an intense emotional experience (e.g., frustration and confusion) and recommended a set of strategies to alleviate these negative emotions. These suggestions were subsequently incorporated in the design of a computer tutor that automatically detects and responds to student emotions.

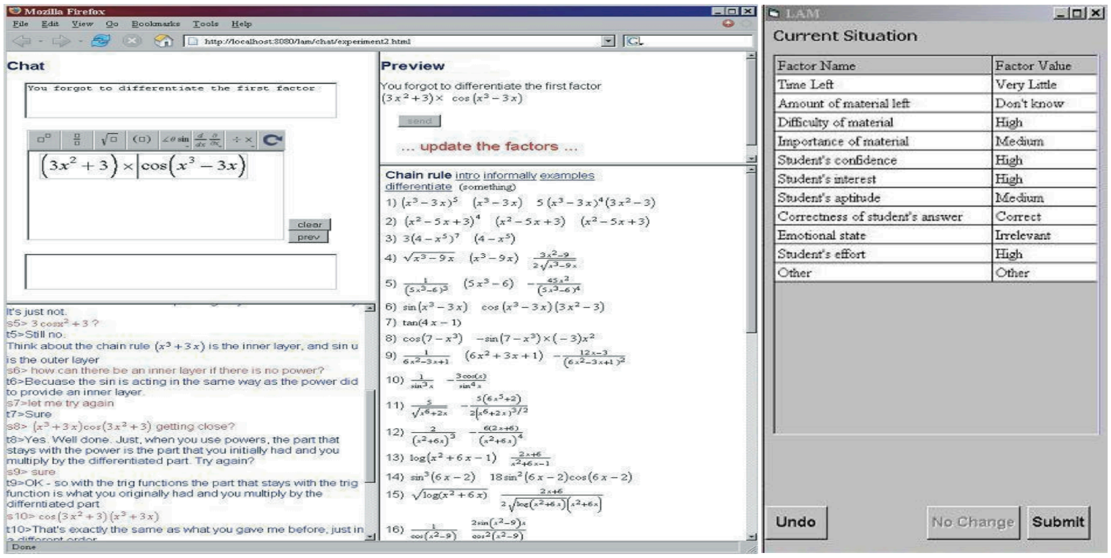
Case Study 4: A_MOODS

de Vicente (2003) augmented his original self-reporting tool (see Case Study 2) in order to allow replay of the learners' previous interactions for the purpose of eliciting affective diagnosis of the learners' by independent judges. These tools serve to interpret learners' behaviours in terms of their affective states *post hoc*. The judges were given access to the learners' actions within the system, minus the self-reporting sliders. The learners' actions included the answers that they provided to the questions during the online tutorial, as well as their mouse movements. The judges were being presented with the learners' longer-term traits (e.g. control, fantasy) before being asked to annotate the learners' short-term states (e.g. confidence, interest). Although before commencing the annotation task the annotators were sceptical about being able to provide the required judgements, after the task they all conceded that it was relatively easy. From their verbal comments along with the online annotations, de Vicente inferred rules corresponding to a combination of specific traits and the online behaviours of the learners, including their mouse movements, as the rules' preconditions and the specific affective states as their conclusions.

Box 2: Two case studies illustrating the use of the retrospective reporting method. Case study 1 illustrates the application of the method in order to infer and triangulate affective diagnosis from the three parties: (1) learners, (2) their peers and tutors and (3) pedagogy experts. Case study 2 demonstrates the use of online post-hoc reporting tools by independent judges, with only a minimal amount of tutoring experience, using a video screen-capture of the learners' interactions and specially designed reporting tools.

Tutors' reporting of affect

The previous sections described self-reporting methods and instruments that can be used to elicit information about learner affect directly from the learners, as well as methods involving external annotators. However, for most part, understanding what affective states specific behaviours represent depends on the context of the situation in which these behaviours occur. Replicating the relevant behaviours out of context is virtually impossible at a later stage. This is why researchers often design realistic situations where the tutors who are asked to annotate the data are also actively involved in generating that data, i.e. they are involved in the tutorial sessions – the point being that the tutors' input is particularly valuable when they have actually tutored the learners about whose affect they are asked to make judgements. As with self-reports and external annotations, tutor reports of learners' affect can be done concurrently or retrospectively and the methods are often combined for more informative data. If the concurrent method is used, the tutor is asked to simultaneously deliver the tutorial and to code or to comment on the student emotions observed in the *heat-of-the-moment*. This has the added advantage of resulting in a data that consolidates both the tutor judgements of the student affective states and their tutoring actions.


Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8080/lam/chat/experiment2.html
clear
prev
send
Done
LAM
Undo
No Change
Submit

Chat

You forgot to differentiate the first factor

$(3x^2+3)\times\cos(x^3-3x)$

It's just not.
s5> $3\cos x^2+3$?
t5>Still no.
Think about the chain rule (x^3+3x) is the inner layer, and sin u is the outer layer
s6> how can there be an inner layer if there is no power?
t6>Becuase the sin is acting in the same way as the power did to provide an inner layer.
s7>let me try again
t7>Sure
s8> $(x^3+3x)\cos(3x^2+3)$ getting close?
t8>Yes. Well done. Just, when you use powers, the part that stays with the power is the part that you initially had and you multiply by the differentiated part. Try again?
s9> sure
t9>OK - so with the trig functions the part that stays with the trig function is what you originally had and you multiply by the differntiated part
s10> $\cos(3x^2+3)(x^3+3x)$
t10>That's exactly the same as what you gave me before, just in a different order

Preview

You forgot to differentiate the first factor
$(3x^2+3)\times\ \cos(x^3-3x)$

... update the factors ...

Chain rule intro informally examples
differentiate (something)

1) $(x^3-3x)^5$ (x^3-3x) $5(x^3-3x)^4(3x^2-3)$
2) $(x^2-5x+3)^4$ (x^2-5x+3) (x^2-5x+3)
3) $3(4-x^5)^7$ $(4-x^5)$
4) $\sqrt{x^3-9x}$ (x^3-9x) $\frac{3x^2-9}{2\sqrt{x^3-9x}}$
5) $\frac{1}{(5x^3-6)^3}$ $(5x^3-6)$ $-\frac{45x^2}{(5x^3-6)^4}$
6) $\sin(x^3-3x)$ $\cos(x^3-3x)(3x^2-3)$
7) $\tan(4x-1)$
8) $\cos(7-x^3)$ $-\sin(7-x^3)\times(-3)x^2$
9) $\frac{1}{6x^2-3x+1}$ $(6x^2+3x+1)$ $-\frac{12x-3}{(6x^2-3x+1)^2}$
10) $\frac{1}{\sin^3 x}$ $-\frac{3\cos(x)}{\sin^4 x}$
11) $\frac{5}{\sqrt{x^6+2x}}$ $-\frac{5(6x^5+2)}{2(x^6+2x)^{3/2}}$
12) $\frac{2}{(x^2+6x)^3}$ $-\frac{6(2x+6)}{(x^2+6x)^4}$
13) $\log(x^2+6x-1)$ $\frac{2x+6}{x^2+6x-1}$
14) $\sin^3(6x-2)$ $18\sin^2(6x-2)\cos(6x-2)$
15) $\sqrt{\log(x^2+6x)}$ $\frac{2x+6}{2\sqrt{\log(x^2+6x)}(x^2+6x)}$
16) $\frac{1}{\cos(x^2-9)}$ $\frac{2\sin(x^2-9)x}{\cos^2(x^2-9)}$

Current Situation

Factor Name	Factor Value
Time Left	Very Little
Amount of material left	Don't know
Difficulty of material	High
Importance of material	Medium
Student's confidence	High
Student's interest	High
Student's aptitude	Medium
Correctness of student's answer	Correct
Emotional state	Irrelevant
Student's effort	High
Other	Other

Figure 1. LeActiveMath data collection interface, specially designed to elicit real-time knowledge from the tutors about the situational factors, including the affective and motivational states of the students. The left-hand-side of shows the chat interface, including the maths editor (top left), the dialogue window (bottom left), the preview window, where the tutor can check their typed response before sending it to the learner (top right), and prepared set of problems that the tutor can quickly insert into their responses by simply clicking on the desired equation. The situational factor interface has some pre-defined factors. The tutors can select a value for a factor in the list using a drop-down list in the right hand side column or they can add new values. The tutors can also add new factors by selecting the "other" option at the bottom of the list.

For example, in Forbes-Riley et al. (2008) tutors and students interacted through an adaptive *Wizard-of-Oz* tutoring system, in which learner's uncertainty was manually annotated by the tutor in real-time. In Porayska-Pomsta et al. (2008) the tutor's task was to tutor the learners remotely, through a purpose-built chat-interface. In this case, the tutors had the additional task of talking-aloud about any and every possible aspect of the interaction as they engaged in tutoring, as well as of selecting values for a set of situational factors from an interface specially selected for this purpose (see figure 1). This provided the researchers with access to the decision processes involved in the tutor's diagnoses and facilitated subsequent data analysis. Additionally, following each completed interaction, the tutors were invited to participate in post-task walkthroughs, which allowed the tutor and the researchers to revisit specific interactions and to discuss the tutors' reports in detail.

Advantages and disadvantages of tutors reporting on affect

In contrast to data collected from annotators, collecting reports from tutors has the crucial, advantage of generating data that helps not only in diagnosing learners' affective states but also in modelling pedagogy and designing appropriate responses to learner's affect. In particular, it allows the researchers to link the tutor diagnoses of learner affective states, with the information that tutors rely on when performing such diagnoses, and with the way in which they act on such diagnoses.

The disadvantage of this method is that it imposes a significant cognitive load on the tutors. For example, in the most extreme cases, tutors are asked not only to tutor in real-time, but also they need to type in order to communicate with the student and report their observations of the student. The high cognitive load may affect the quality of the tutoring and of the resulting data and, consequently, it may increase the effort needed to explain the data *post-hoc*. It is therefore important to invest in an interface design that is as seamless as possible, that is, one that reduces the effort required of the tutor. One way of lessening the cognitive load is to prepare a set of problems that the tutor might give to the student and allow the tutor simply to click on each to reduce the amount of typing needed. This is particularly useful in the domains such as mathematics where formulae, the typing of which is cumbersome and time consuming, are part of the tutoring interaction (e.g. Kaśka Porayska-Pomsta, et al., 2008). The interface can also be designed to prompt or even "force" the tutor to make an observation. Whilst this may initially add to the cognitive load, with appropriate prior training, the tutors tend to become more efficient at systematic reporting and find that it eases their task, because they do not have to remember to report their observations in addition to engaging in all the other tasks (Porayska-Pomsta et al., 2008).

Type and quality of data resulting from tutors' reports

The data collected through participant-tutor reports include information about what states the tutors think the learners experience during specific interactions, the learners' behaviours that lead to tutors' specific judgements, and concrete tutoring actions that tutors commit as a consequence of those judgements. Such rich data can be used as needed depending on the specific questions asked and may be combined with other data such as pre- and post- tests of learners' knowledge. Together such rich data can inform about what affective states lead to increased learning (e.g. A. Graesser, Chipman, King, McDaniel, & D'Mello, 2007) and what specific tutoring actions are most effective.

A crucial aspect of data obtained from tutors' reports is information about the sources of evidence that the tutors rely on when making their decisions. These may include anything from the amount of time that the learner takes to answer a question or to solve a problem, linguistic cues, such as unfinished sentences, question marks at the end of statements – to the nature of the solution provided by the learner. In some studies tutors also point to their specific actions as potentially impacting learners' affective states. For example, Porayska-Pomsta et al. (2008) report that tutors sometimes anticipate dips in learners' confidence when they set a question or problem difficulty level as high. This kind of evidence can be compared with and complemented by results derived from studies using learners' self-reports (Section 4). In this way, the two different perspectives: that of the tutor and the learner can be reconciled to derive a more accurate model.

Table 3. Summary of the instruments required during the tutor participant and the advantages and disadvantages of the method

Tutor participant annotation	**Concurrent**	**Retrospective**
Involves	Tutor reporting on student's affect during a computer-mediated tutor-student interaction	Tutor revisiting the concurrent annotations after the learning episode
Instruments	Tool for computer-mediated interaction Free response (talk-aloud) GUI for: - Dimensional response - Discrete emotion response	video-; audio-recordings Free response (interviews) Dimensional response Discrete emotion response
Advantages	- Helps in diagnosing students' affective states as well as in modelling pedagogy and designing appropriate responses to learner's affect - Tutor reports can be compared with simultaneous student reports Repeated use can result in tutors' improved reporting skills	
	Provides *heat-of-the-moment*, stream of consciousness reports	• Allows to elaborate and focus on details;
Disadvantages	- relies on the bandwidth of available information in the computer-mediated interaction	
	• requires an investment of time on the researcher part to prepare and implement the reporting tools that allow the tutoring and the reporting to take place concurrently • increased cognitive load which may affect both the quality of the teaching delivered and of the reports	Time consuming to prepare the materials for the post-hoc walkthroughs and to administer It may be difficult to bring the tutors back for the post-hoc sessions Tutors may want to change their assessment of the student or misremember the specific situations Increases the amount of data to be consolidated.

Reliability and validity of annotation: multiple human judges

After exploring a number of methodologies that allow to monitor learners' affect during tutoring sessions, we now turn to an analysis of the validity of the affect annotations obtained from these methods. Validity in affect measurement is critical, because akin to most psychological variables affect is a construct (i.e., an inferred conceptual entity). Therefore, it cannot be directly measured and one can only approximate its true value. This approximation raises critical validity concerns in the measurement of human emotions. These include conclusion validity, internal validity, construct validity, and external (or *ecological*) validity (Rosenthal & Rosnow, 1984).

Conclusion validity pertains to the ability to infer a relationship (not necessarily causal) between any two variables of interest (e.g. are increased levels of happiness related to increased learning gains?) (Shadish, Cook, & Campbell, 2002).

Internal validity is concerned with establishing whether a relationship between two variables is causal (i.e. does happiness cause positive learning gains?) (Campbell & Stanley, 1963).

Construct validity involves determining whether the operational definitions of a construct accurately reflect the construct (Campbell & Fiske, 1959). Simply put, are we measuring what we are claiming to be measuring?

Ecological validity is related to the extent that any observed relationship can generalise to other people, places, and times (Shadish, et al., 2002).

Although, each of these validity measures is important to the scientific study of affect, this short discussion will focus on construct and external validity only. This is because our immediate concern is whether we can accurately measure affect (construct validity) and whether our measurements generalise (external validity). We are less interested in this discussion with what relationships exist between affect and other variables (conclusion and interval validity).

Construct validity is an important concept in affect measurement because human emotions cannot be directly observed. Hence, we have to rely on operational measures such as self-reports, observer judgments, or monitoring physiological and bodily correlates of affect expression. For example, consider the problem of measuring student happiness during a learning task. Since happiness cannot be directly measured (unlike physical measurements such as height, weight, etc.), it would have to be inferred from operational definitions such as self-reports, observer judgements, tutors' annotations or by monitoring facial expressions. In this hypothetical situation, establishing construct validity pertains to the extent to which each of these operational definitions accurately reflects the desired construct (e.g. happiness). For example, simply monitoring the presence of smiles would be an unstable operational definition for happiness because people also smile when they are unhappy (e.g. a grimace) and there is a difference between true (or Duchenne smiles, i.e. unposed smiles) and forced smiles (Ekman, Davidson, & Friesen, 1990).

Although, it is generally difficult to establish construct validity, Campbell and Fiske's (1959) landmark paper provides useful guidelines. Perhaps the most basic requirements are the demonstration of reliability, convergent validity, and discriminant validity.

Reliability implies that the same or similar measurement device should produce measurements that are highly correlated. For example, affect judgments provided by two annotators observing a learner) should be strongly correlated. Or an automated system that detects affect by monitoring facial features should demonstrate similar performance under varying conditions such as backgrounds, lighting, etc.

Convergent validity means that measurements produced by different measures that are theoretically related to a construct should be highly correlated. Therefore, in order to establish convergent validity in measuring affect, multiple measurement schemes should be employed and these should be strongly correlated. For example, subjective experiences of affect (Measure 1) can be correlated with facial expressions (Measure 2) (Bonanno & Keltner, 2004; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005).

It is important to emphasise the difference between multiple measures of a construct (e.g. self-reports and nonverbal behaviours) and to refer to multiple exemplars of the same measure (i.e. two researchers observing a learners' emotions). For example, affect judgments made by two observers can correlate highly with each other (i.e. high reliability), yet this correlation is insufficient to establish any degree of convergent validity (i.e. multiple measures were not used). On the other hand, convergent validity could be established if self-reports of affect were correlated with judgments made by researchers.

As an example, consider a study by Graesser and colleagues (2006) that investigated the validity of affect judgments by incorporating several contrasting measures of learner affect. This study involved 28 learners interacting with AutoTutor, an ITS with conversational dialogue (Arthur C. Graesser, Person, Harter, & TRG, 2001). Learners' affective states were measured after the tutoring based on video recordings of the learners' face and a screen capture of their tutorial sessions. The judges were (i) the learner (i.e. self reports), (ii) an untrained peer, (iii) two trained judges, and (iv) an accomplished teacher (S. D'Mello, et al., 2008; A. Graesser, et al., 2006). The trained judges had been trained on how to detect facial action units (i.e. facial muscle movement) according to Paul Ekman's Facial Action Coding System (FACS) (Ekman & Friesen, 1978). They also had considerable experience interacting with the ITS. Hence, their emotion judgments were based on contextual information as well as facial features.

The results of the study indicated that trained judges exhibited reliability (they agree with each other) as well as convergent validity (their judgments match self-reports). Agreement between the self-reports, peers, and accomplished teachers was near zero. Therefore, from a methodological perspective, affect judgments by the participants themselves (i.e. self reports) combined with trained judges, seems to be a useful approach towards establishing a degree of construct validity.

Graesser et al. (2006) study just discussed, along with several others, supports a number of conclusions about the difficulty of emotion measurement by humans (Ang et al., 2002; Grimm, Mower, Kroschel, & Narayanan, 2006; Litman & Forbes-Riley, 2004; Shafran, Riley, & Mohri, 2003). For example, Litman and Forbes-Riley (2004) reported kappa scores of .40 in distinguishing between positive, negative and neutral affect. Ang et al. (2002) reported that human judges making a binary frustration-annoyance discrimination obtained a kappa score of .47. Shafran, Riley and Mohri (2003) achieved kappa scores ranging from .32 to .42 in distinguishing among 6 emotions.

Statisticians have claimed that kappa scores ranging from 0.4 – 0.6 are typically considered to be fair, 0.6 – 0.75 are good, and scores greater than 0.75 are excellent (Robson, 1993). Based on this categorization, the kappa scores obtained in these studies would be considered to be poor. However, such claims of statisticians address the reliability of multiple judges when the phenomenon is more salient and when the researcher can assert that the decisions are clear-cut and decidable. However, it is important to understand that affective judgments are fuzzy, ill-defined, and possibly indeterminate. A kappa score greater than 0.6 is expected when judges code some simple human behaviours, such as facial action units, basic gestures, and other visible behaviour, but it is unlikely that perfect agreement will ever be achieved in affect measurement and there is no objective gold standard.

An important concern for research involving affect is the ecological validity of the studies. Ecological validity (or external validity) pertains to our ability to make inferences about behaviour in the real world from controlled laboratory research. Barrett (2006) discusses some common threats to ecological validity in affect measurement. One issue is a sampling problem that plagues designs that adhere to *culled* sampling procedures. In these situations the judges of affect are exposed to a small, carefully selected, sample of behaviour available in the data collected. This selective sampling process results in overly optimistic recognition rates that do not reflect the realistic difficulty in decoding affect. For example, a meta-analysis by Elfenbein and Ambady (2003) report that cross cultural face perception accuracies were lower for studies that used the entire data set when compared to those utilising a culled sampling procedure.

Another important violation of ecological validity occurs when actors are used to portray affective expressions. Examples of these include caricatures of emotion as prominent in the facial expression literature (Ekman & Friesen, 1976), or actors posing various expressions. The use of actors is sometimes defended as producing prototypical emotional expressions that resemble real emotional productions (Banse & Scherer, 1996). However, it is unclear whether actors portraying artificial affective expressions in controlled laboratory environments will ever yield affect classifiers that will generalise to naturalistic expressions in real-world environments

Physiological and behavioural sensing

Scherer (2005) provides one of the most comprehensive and formal accounts of *Emotion*. According to him, emotion is a multimodal component process that incorporates (i) *cognitive component,* responsible for a person's ability to appraise objects or people as well as events and events' consequences, all of which facilitate the appropriate emotional (re)actions in a person, (ii) *neuro-physiological component* that relates to bodily manifestations or symptoms of a person experiencing an emotion, (iii) *motivational component*, related to a person's action tendencies, (iv) *motor expression component* (behavioural manifestations such as facial and vocal expressions) and (v) *subjective feeling component* which corresponds to the emotional experience itself. Whilst these components may function independently of one another (Russell, 2003) Scherer proposes that the entire episode of emotion is defined by a coordinated and synchronised process involving all of the five components (pp. 698-699).

To date there are no studies that tackle this process in its entirety. Instead, the separate components have been studied individually and, often, in relation to very limited number of emotional states (i.e. typically in relation to the basic emotions). Four of the Scherer's emotion components – cognitive, motivational, motor expression and subjective feeling – can and have been studied using the observational and self-reporting methods described earlier. However, the neuro-physiological component requires different tools than the traditional methods afford – namely **physiological sensors**. In this section we introduce physiological sensors as means for detecting emotion, and we place their usage in the broader methodological context introduced in this chapter. Additionally, we introduce **behavioural sensors** as devices that are related to physiological sensors and that can enhance both the study of the motor expression component and the interpretation of the data obtained through the physiological sensors

Physiological sensing

Physiological sensors are devices that allow researchers to detect, to monitor and to measure precisely physiological events, such as a person's heart rate, body temperature and neural patterns of the brain during a specific task. Numerous devices have been used over decades for medical and forensic purposes, most notably as part of lie detectors, and many more sophisticated (both more accurate and less intrusive) tools have been built over the past decade for use outside the specialised medical fields. Increasingly, such devices are becoming cheaper, more reliable and easier to employ. The most common tools include wearables such as gloves that sense the wearer's skin conductance, wrist bands that are able to detect heart rate and heart rate variability, EEG (electroencephalography) devices that allow to study electrical activity of the brain through a net of sensors placed on a participant's head, Blood Volume Pulse (BPVs) devices able to detect blood flow, respiration rate detectors, and, more recently, infrared thermal cameras able to detect people’s facial and/or bodily thermal features.

Measurements of skin conductance, also known as galvanic skin response (GSR – see case study 6 in box 3), rely on the idea that the electrical conductance of the skin varies depending on its level of moisture and, as sweat glands are controlled by the sympathetic nervous system, this can be indicative of psychological and physiological arousal. Fluctuations in the skin conductance levels, therefore, can be indicative of changes in, if not of the exact nature of, the affective states of the persons monitored. GSR is used to obtain precise data about possible changes in learners’ arousal levels. Such data is then used as the basis for understanding the emotional and cognitive paths for individual learners during learning episodes.

Electrocardiograms (ECG) can be used to detect changes in heart rate and inter-beat intervals of the learners to determine the heart rate variability (HRV). In a general population, a low HRV can indicate a state of relaxation, and an increased HRV may indicate a state of stress or frustration (Gunes & Pantic, 2010) Similarly, respiration rate (R) can be measured through an appropriate device and can be indicative of a relaxed state (low R), or aroused negative emotions such as fear or anger (high R) (Chanel, Ansari-Asl, & Pun, 2007).

Before the advent of wireless technologies, such devices, although very useful, were also quite intrusive as they restricted the mobility of the wearer (mostly through a multitude of wires that bound a device worn, and the wearer, to a computer) and the equipment was both specialised and often prohibitively expensive. This intrusion level was often criticised by researchers for potentially being one of the factors that to influence the learners’ emotions and learning and therefore, similar to the issues discussed in relation to self-reports, the ecological validity of the data obtained through those means was questioned. Recent developments in wireless technology made GSR in particular, viable in studying changes in emotional states of learners in natural contexts as opposed to in artificial, laboratory conditions.

One of the most prolific groups known for investigating the use of such devices is the Affective Computing Group at the MIT Affective Computing Lab. The group reports success in developing and using numerous physiological sensing tools including wireless wrist-bands and photoplethysmography (PPG) comprising a magnetic earring sensor and wireless earpiece that can be worn in the earlobe (Fletcher, Poh, & Eydgahi, 2010) Recently the group also reported the development of a low-cost method for measuring multiple physiological parameters including the blood volume pulse (Picard, 2010; Poh, McDuff, & Picard, 2011). Infra-red thermal cameras (ITC) are increasingly used in emotion detection research as they are non-intrusive, they do not rely on contact with the human body and therefore they offer great possibilities in education contexts, especially as relates to increasing the ecological validity of those contexts and the behaviours observed therein. The information obtained through ITC, in particular the increased blood perfusion in the orbital muscles, has been shown to correlate with anxiety and stress levels in humans (Pavlidis, Levine, & Baukol, 2001, as cited in Gunes and Pantic, 2010). It is important to note, however that ITC use for detecting emotion is in its infancy, not least because no model of the relationship between blood flow, skin temperature and bodily muscle activity is as yet available (Gunes & Pantic, 2010). Equally, the correspondences between the data obtained through physiological sensors, the emotions experienced and the effect that these have on learning for different individuals are yet to be established and validated across different learning contexts.

Behaviour sensing

Behaviour sensing relates to gathering data about behaviours, which, although observable, are difficult to record and sometimes register overtly with the "naked eye/ear". The behaviours in question involve eye pupil dilation, eye gaze and eye fixation posture or acoustic characteristics of someone's speech. All of these behaviours, individually and in combination, can be manifestations of a person's emotional states. Among the most frequently used behavioural measures are a variety of cameras able to detect a user's eye movement and eye-gaze direction at any given point during a person's interaction with a system, posture chairs able to detect a person's posture while s/he engages in a task and computer mice able to detect the pressure of a person's clicks. With the advent of commercially available and affordable touch and multitouch devices, the pressure and the nature of touch, e.g. one or more fingers, whole hand, etc., can also be used for behaviour sensing purposes. Acoustic sensors are able to detect and analyse voice frequency, amplitude and tone, while the cutting-edge motion sensing technologies (also now commercially available – e.g. Kinect) are employed to detect the direction and nature of a person's whole body. It is important to note that, as with physiological sensing, behaviour sensing research is relatively new and therefore the models developed for mapping between the raw data obtained through behaviour sensors and the emotions and learning effects are as yet to be fully developed and evaluated.

Nevertheless, there exists a body of work that correlates different behavioural reactions to people's specific affective and mental states, e.g. body posture (see case study 5 in box 3) - to extreme emotions such as fear, surprise, joy (Burleson, 2006); eye gaze fixation on the correct parts of another human's face – to their ability to empathise with others (Klin, Jones, Schultz, Volkmar, & Cohen, 2002), or prolonged eye gaze fixation on a specific aspect of a task or an object – to sustained level of attention or flow (Findlay & Gilchrist, 2003). Furthermore, recent developments in the computer-vision field also carry an increasing promise of automatic detection of emotion based on facial expression recognition (smiles, frowns, etc.) through a detailed analysis of muscle tones and muscle tensing, relative position of eyes, nostril and mouth contours, thus facilitating a more precise study of the physiology of the motor expression component (Zeng, Pantic, Roisman, & Huang, 2009).

Advantages and disadvantages of physiological and behavioural sensing

Physiological and behavioural sensors allow researchers to collect significant amounts of data in a systematic way. The definite advantage of such sensors is that they facilitate access to the behavioural manifestations of emotions other than those that can be easily observed with the naked eye. In many cases, and crucially for education, such access further facilitates in-depth and quantitative study of the relationship between emotion and cognition. For example, eye tracking devices are able to detect accurately what the learner attends to at any given point during a learning task, whether or not the learner fixates on any particular features of a task, and the length of any such fixation. This information is important to our understanding of the individual learners' states, such as the state of *flow* or being *stuck*, hesitation and anxiety, confusion and eureka. The eye-tracking data can also be indicative of the learning patterns of individual learners (e.g. through monitoring gaze shifts) that a human observer might find impossible to notice and record accurately. Use of such sensors also potentially reduces the subjectivity that is typically involved in humans making judgements about the occurrence of their own and of others' emotional states – but not the interpretation of what those states might actually be. An EEG device placed on a learners' head will simply record the areas of the brain which are activated when the learner engages in a particular task, a wireless wrist band will inform about a person's heart rate during a task, while an eye tracker will detect the exact coordinates within which the learner is fixating their gaze. This level of accuracy is invaluable to research and facilitates a collection of large amounts of data from which different patterns of emotional experiences can be inferred.

As with most techniques, physiological and behavioural sensors have their limitations. In particular, their use is often criticised for potentially interfering with the learning task or with users' emotional states, because they are either visible to the user or worse – they have to be worn by the user. Wearable devices vary in the degree of their intrusiveness, for example wireless wristbands tend to be less intrusive and easier to get used to than, say, the EEG headgear. Much effort has been dedicated in the last decade to creating sensors that are less intrusive and less visible, for example through improvements made to the quality of the cameras built into the desktop computers and the emergence of reliable wireless technology. With such improvements these devices allow the researchers to record users' interactions with a system in a way which, in principle, does not adversely impact the their affect any more than the use of the system itself does. For example, the recent ECHOES project (Kaska Porayska-Pomsta, Bernardini, & Rajendran, 2009; Kaska Porayska-Pomsta et al., 2011), which aims to support young children in acquiring social interaction skills through use of an enhanced-reality environment, employs multitouch screens and vision component which relies on two off-the-shelf wide fish lens cameras that are mounted on the top and the bottom of the screen. No devices are worn by the children, who are free to move naturally in front of the screen, and to interact with the environment without being physically constrained by the vision technology.

There are, however, issues associated with the use of physiological and behavioural sensors that are less easy to resolve. While such sensors provide very concrete information about learners' physiological states at various points during their learning, such information still needs to be correlated with the learners' own reports of the emotions and/or the judgements of the outside observers or teachers/tutors. Furthermore, in order to register the relevant information some devices, including eye-trackers, facial expression recognisers, and acoustic processing tools need to be calibrated and/or trained, using predominantly supervised machine learning methods, on existing data such as images of different facial expressions, or acoustic data that are annotated by human judges *a priori*. As was discussed in the previous sections, obtaining such annotations can lead to inconsistencies and disagreements that may be difficult to reconcile. Furthermore, machine learned classifiers are often limited in their ability to generalise to new, previously un-encountered data, especially if such data represents spontaneous behaviours (e.g. Zeng, et al., 2009). Also, while able to derive typical classes, data points that represent rare or individual occurrences of specific behaviours, which nevertheless might inform our understanding of the learner affect, might be either ignored or misclassified.

Calibrating eye-trackers typically involves a learner or user being engaged in an introductory activity that stimulates them to look at different pre-defined regions of the workspace, typically a computer screen. With most current devices, eye-gaze can be detected reliably only if the learner's eyes remain within that workspace and require them to sit relatively still and face the workspace directly. This can be very restrictive. Rapid movements, temporary eye-occlusions, profile position of the learner in relation to the workspace can all be problematic and will result in the loss of vital information. For example, in ECHOES, children engage with the environment through a 42" multitouch screen. While, in this case, the size of the screen is dictated by the importance of encouraging physical movement in children and their exploration of the different regions of the virtual space portrayed, this set up makes it difficult for the eye gaze tool to record all of the needed information – young children tend to move a lot and it is often difficult to obtain sufficient view of their faces and eyes to enable the estimation of where they are looking at specific moments and what facial expressions they might be adopting. While both calibration and control of the user movements may be easier with adult population, restricting learners behaviours will impact on the design and nature of the interaction facilitated. The ECHOES example serves to illustrate a frequent tension that exists between researchers' desire to obtain precise (low-level) data such as is facilitated through physiological and behavioural sensors, and high-level pedagogical recommendations, for example allowing young learners to have freedom of movement, that might be necessary to enhance learning. The tension lies in the fact that, often, the calibration required for complete and reliable data collection through such sensors and the desire to provide learners with freedom of movement can be mutually exclusive. In ECHOES the information about the eye-gaze of the child is crucial to determining whether the child is attending to the socially informative regions of the depicted situation, but equally the freedom of movement allowed in this environment is needed to allow the children to express themselves, to self-regulate and to embrace the environment on their own terms.

Much work is currently dedicated to finding solutions to the problems posed by eye-tracking technology (Chen & Lemon, 2009). Unfortunately, the solutions are not as yet robust enough to always provide a reliable method in the environments other than the laboratory ones. Similarly, automatic facial expression recognition is still largely limited to the laboratory settings. While the devices are becoming more robust, they are often trained on posed or acted expressions and have limited success with natural, spontaneous expressions occurring in everyday situations (Zeng, et al., 2009). Acoustic processors are also problematic, for although they can be a rich source of information about people's possible emotional reactions such as excitement, anger, fear, surprise, etc., increasingly researchers discover that annotating acoustic information with emotion categories needs to be supported by the lexical and syntactic information that accompanies the sounds (Devillers & Vidrascu, 2006).

It is important to consider that data obtained from a single type of sensor (e.g. only eye gaze or only heart rate variability) often has limited informative value. Emotion is triggered through multiple sensors with which our bodies are equipped. All of touch, smell, sight and sound contribute to what emotions we might experience at any given time and the extent and duration of those emotional experiences. In order to understand fully the data from one sensor, it is often necessary to obtain further physiological information from other types of sensors. For example, both fear and joy may be manifested by an increased heart rate. Further information such as facial expression or acoustic information or both may be needed to clarify which one of the two states is being experienced. Thus in order to correctly interpret the specific emotional states, it is necessary not only to have access to the individual sensor data but also to be able to *fuse* the relevant low level data to arrive at the high level interpretations. Although humans are adept at fusing information around them to interpret their environment, they do so based only on imprecise data. They do not have access to exact heart rate and heart rate variability of other people, often they are not able to precisely locate where another person is gazing and they certainly have no idea about exactly which regions of other people's brains display heightened activity at any given point. Current approaches to combining such precise information involve video recordings of people involved in a task being annotated for specific emotions. The annotated data together with the sensor data corresponding to the points at which the emotions were thought by the annotators to occur is then used to learn the appropriate sensor data –specific emotion classifications. The success of the current approaches is still limited and does not extend beyond a small and typically posed emotional displays. For a comprehensive review of both the state of the art in physiological sensor use and capabilities as well as issues related to fusion the reader is referred to (Zeng, et al., 2009 and Gunes and Pantic, 2010).

Case Study 5: D'Mello and Graesser (2009).

D'Mello and Graesser explored the reliability of detecting learners' affect by monitoring their gross body language (body position and arousal) during interactions with an ITS called AutoTutor. Training and validation data on affective states were collected in a learning session with AutoTutor, after which the learners' affective states were rated by the learner, a peer, and two trained judges. They used an automated body pressure measurement system to capture the pressure exerted by the learner on the seat and back of a chair during the tutoring session. They extracted two sets of features from the pressure maps. The first set focused on the average pressure exerted, along with the magnitude and direction of changes in the pressure during emotional experiences. The second set of features monitored the spatial and temporal properties of naturally occurring pockets of pressure. They constructed five data sets that temporally integrated the affective judgments with the two sets of pressure features. The first four datasets corresponded to judgments of the learner, a peer, and two trained judges, whereas the final data set integrated judgments of the two trained judges. Machine learning experiments yielded affect detection accuracies of 73%, 72%, 70%, 83%, and 74% respectively (chance=50%) in detecting boredom, confusion, delight, flow, and frustration, from neutral. Accuracies involving discriminations between two, three, four, and five affective states (excluding neutral) were 71%, 55%, 46%, and 40% with chance rates being 50%, 33%, 25%, and 20% respectively.

Case Study 6: Affective Diary

An interesting use of sensors that may aid self-reflection and self-reporting task is when such sensors are used to provide information to the users about their physiological reactions and events in which they participate. For example, Lindström et al. (2006) devised mobile affective diaries that harvest bio data through wearable wireless devices such as wristbands (i.e., they collect GSR data), which the users wear throughout the day. This data is downloaded by the user in the evening via Bluetooth onto a tablet PC and is represented to the user as animated characters with different body postures and colours, denoting movement and arousal. The users can click on the different representations in order to change the characters, if they feel that a different representation of their mood throughout the day is more appropriate. Among other things, such as enabling the users to relive parts or all of the events in the day, such devices can help people reflect on their experiences, debate and verbalise their affective states. The use of such methods for collecting data and providing people with an accurate memory prop, carry a promise of allowing access to a combination of the individual users' physiological states, their behavioural correlates and the corresponding subjective feelings, thus bringing us closer to a synchronised study of the Scherer's emotion components. In the context of education these methods may provide a basis, especially for young children engaging in self-reflection and emotional self-regulation, both of which are crucial to learning and social success at school.

Box 3. Case studies illustrating use of different sensors: case 5 uses behavioural sensors – posture chairs to detect learners' emotions in order to assess the reliability of these types of sensors; case 6 uses GSR to as part of a more complex set up to detect and record the user's emotions throughout the day in order to represent the data gathered to trigger reflection.

2. DESIGNING FOR AFFECT

Ivon Arroyo[1], Benedict du Boulay[2], Manolis Mavrikis[3]
[1]Computer Science Department, University of Massachusetts, Amherst, USA
[2]Human Centred Technology, University of Sussex, Brighton, UK
[3]London Knowledge Laboratory, Institute of Education, London, UK

Introduction

The preceding chapter has underlined the complex nature of the phenomena that cross over into cognition, motivation and metacognition. This chapter explores the issue of designing computer-based learning environments that take different aspects of affect into account.

An issue for this kind of design is that historically learning systems have tended to focus on what is to be learnt and understood and on the learner's idiosyncratic reconstruction of that material. So dealing with affect inevitably has the feel of an "add-on" to something which is essentially rational rather than emotional.

The possible scope of affective learning approaches - designed to elicit particular emotional states in students and/or to respond to students' emotions - is immensely broad. Effective teachers respond to student affect on a daily basis, e.g. by analyzing the moods and motivations of students, and reacting to these in a kaleidoscope of ways, or indeed by reframing whole topics in such a way as to enhance their appeal or reduce the chances that an adverse reaction will occur to a touchy subject. In fact, they often devote as much time to the achievement of students' motivational goals as to the achievement of their cognitive and informational goals (for example Lepper & Hodell, 1989). Yet, an in-depth analysis of the different mechanisms through which educators address affective traits or states, and their effectiveness toward enhancing learning or student motivation for learning, has yet to be undertaken. Different approaches to this research are starting to appear, some of which are shown as case-studies at the end of this chapter. Some of this work links into the notion of emotion as a social construct and thus into the importance of the social interactions underpinning learning. From the point of view of design this raises the issue of the degree to which the interaction between a learner and either an online tutor or an online peer is, or can be construed by the learner as *social* in terms that are at all similar to those that apply to interactions with other humans (see .e.g. Nass, Fogg, & Moon, 1996). It also raises the issue of the relative value of systems supporting social interaction either directly (e.g. through online collaborative activity) or as foci for social interaction (e.g. by being used by more than one student at a time who talk around the machine – either spatially or temporally) as part of their repertoire of activities.

Finally, the previous chapter on methodology reminds us that building systems that take affect into account requires additional tools for data-gathering not just to enable the system to react appropriately but also to assist researchers undertake evaluations.

Designing technology for affect

Although some computer-based learning environments are designed (implicitly or explicitly) to take student affect into account in a non-adaptive manner, only few exist that detect and respond dynamically in even limited ways to students' affective states or traits. This chapter explores the many ways in which a learning environment can be designed to take student affect into account and the ways and the points at which it can provide affective responses. So the chapter (i) categorizes different ways to take students' affect into account in the context of an educational scenario, (ii) describes work that has been done to promote desirable emotional states or traits, and (iii) identifies new areas of research.

This chapter examines several of the issues that need to be confronted when taking affect in the design of educational technology into account. It does not attempt to be a "how to do it guide", but rather to elucidate the issues involved. The case-studies also play such an illustrative role. The focus of the chapter is on how macro-adaptive and micro-adaptive technologies can be designed to customize instruction to individual or group student affect *dynamically*. This is in contrast to systems that are non-adaptive dynamically and where any affective dimension of the interaction is fixed from the start and is the same for all users.

When designing technology for affect, it is important to consider three dimensions: (1) Categories of Technology, *i.e.,* what kinds of technologies are being considered and their roles in relation to the student; (2) Degree of Adaptability, *i.e*., how adaptive or adaptable to affect the technology could or should be; and (3) How the System can Respond, *i.e*., the specific ways that the technology provokes or responds to student affect.

Categories of technology

When discussing affective technologies, it is important to keep in mind the kinds of learning technologies under consideration. There are two general categories: (1) "tools for learning" and (2) "learner-centered learning environments."

The first category – "tools for learning" – involves systems that do not have an explicit teaching agenda or student model that dynamically drives the system's behavior. Such tools might include those designed to facilitate collaboration between student and teacher such as chat forums, as well as programming languages like LOGO, hardware blocks such as Lego, or a simple browser to facilitate student research of a topic. In many cases the tools will have been designed to provoke specific affective reactions in their users. It should not go unnoticed that Lego blocks are bright colours, or that the look and "feel" of various browsers are the way they are (partly to be easy and effective to use, but also to be pleasing to the eye). Their designers might harbor expectations regarding how their constructs should be used in teaching, but the tools themselves cannot reason about that.

In the second category of technologies - "learner-centered learning environments" – the software embodies either a pedagogical decision-making component (tutoring software, inquiry-based software, web-based courseware), or an underlying pedagogical plan. Such systems are often known as intelligent learning environments or intelligent tutoring systems. Examples include systems that are very much learner-driven, as well as systems that act as tutors, mentors, coaches, peers, *etc*. Regardless, the behavior of the system is determined by an explicit representation of both a teaching agenda and a model of the student. When designing for affect one must consider both the affective component of the teaching agenda and the affective component of the model of the student.

Degrees of adaptability

The second major consideration when designing for affect is the degree of adaptability of the software. The degree of adaptability can be divided into three kinds: non-adaptive; macro-adaptive; and micro-adaptive. Non-adaptive – as the name implies – indicates that the system behaves the same for all users, ignoring individual or group variation. The Macro-adaptive approach involves adaptation to relatively stable dimensions such as self-efficacy, gender, ethnicity, age and the like. Typically such a system operates in a small (often binary) number of modes optimised for the characteristics of a particular target audience. The Micro-adaptive approach involves adaptation to (possibly more transient) individual affective differences, e.g. across a single lesson, caused by changes in mood or changes in degree of effort. Each of these degrees of adaptability is discussed in more detail below.

Non-adaptive affective technologies. Non-adaptive affective technologies are designed to bring about an overall emotional response from users that is productive in the learning process or that may be a goal in and of itself (*i.e.,* to 'motivate' the student), but that behave in the same way for each member of the target audience, regardless of the user's emotional state. Such technologies (as illustrated in the first column of Table 1) are designed to ensure that the software elicits certain emotions in order to be affectively effective for a chosen target demographic. Such emotional states may include interest in or liking of a domain: for example, software that uses representations of animals to teach children math in order to elicit liking of the subject through its association with animals, based on the theory that young students like animals. Alternatively, the technology might make the student feel respected and appreciated, such as an online course on diversity designed to make the reader feel respected rather than judged. Another example of technology designed to elicit emotional states is a learning tool that is aesthetically beautiful in order to stimulate pleasure or satisfaction in its use ("It's a joy to use"). One could design technologies to elicit other emotions as well, including tenderness (smart toys), or curiosity (software with videos about an interesting phenomenon in nature that connects to the taught subject). Outside the technological realm, some studies with real teachers have investigated curriculum design that is emotionally effective in the classroom. These studies take into account the cultural background of the student to introduce topics that are likely to generate positive feelings or those that will make the student identify more with the taught subject (Rosiek, 2003). Such technologies are placed in the first column of Table 1.

Macro-adaptive affective technologies. Macro-adaptive affective technologies (identified in the second column in Table 1) respond to relatively permanent affective traits or attitudes of the students (such as self-concept in students' ability to succeed in the subject, level of interest, baseline reported frustration while working in the domain, dislike of the domain, susceptibility to boredom and so on) by adjusting the appearance or response of the technology in an effort to address those attitudes. This characterization does not necessarily have to be related to emotional traits, but could instead be based on non-emotional traits, for instance clustering students by hyperactivity level, learning disability, or by gender. For instance, software could be modified to include digital learning companions that adopt the role of a peer-motivator instead of an older expert. Such changes in the role of the digital character could prioritize the coping needs of students with low self-efficacy beliefs, as opposed to a role that follows an expert model which might satisfy the student's cognitive needs without paying attention to his or her affective needs (Kim, 2007). In all cases, once the student has been classified into a particular group, macro-adaptive affective technologies will provide appropriate emotional scaffolding to students who fall into that targeted group, but they will not keep track of or address individual, and possibly transient, emotional states as the session progresses. Although macro-adaptive technologies exist (Ainsworth, 2007; Arroyo, Beck, Woolf, Beal, & Schultz, 2000), the authors know of very few technologies that are macro-adaptive to affective traits, or that provide emotional scaffolds by adjusting instruction to effective traits. One example is the work of du Plessis (1998), who built a system that adapted to various personality traits, such as introversion and extroversion. Macro-adaptive technologies for affect is an area with clear potential for further research.

Micro-adaptive Affective Technologies. The third set of technologies along the continuum of adaptability are micro-adaptive affective technologies (those that correspond to the third column in Table 1). These adaptive technologies are individualized, in the sense that they respond to each student's emotional state as the student progresses through the specific emotional arc of a learning episode. These micro-adaptive technologies involve two important steps: (1) the diagnosis of specific emotional states (such as frustration or anxiety) in real time while the student interacts with the software; and (2) reaction via a specific emotional pedagogy.

A system that closely resembles a pure micro-adaptive affective technology is the work of Kapoor, Picard and Burleson (Kapoor, Burleson, & Picard, 2007), in which the software diagnoses whether a particular student will click on an "I am frustrated" button. Diagnosing whether a person is frustrated can be recognized either by self-reports (the student clicks on a button that reports they are frustrated), or as in this case, utilizes devices that create a high bandwidth of communication with the student (a mouse that detects hand pressure and a glove that detects hand humidity). Meanwhile, other attempts have been made to employ students' direct actions with the system only to detect and adapt to transient emotional and motivational states such as confidence, effort, etc. (e.g. S. K. D'Mello, et al., 2006; del Soldato & du Boulay, 1995; Mavrikis, et al., 2007). Other research involves detecting specific behaviors – such as gaming the system - that are related to emotional states such as frustration and boredom (R. Baker et al., 2008; Johns & Woolf, 2006).

The second component of research involving micro-adaptive learning technologies explores how best to react to student emotional states after the emotion has been identified. This area has been much less studied, and there is ample room for further research. A summary of progress so far is given in Section 4.4.

The information about categories of system is summarized in Table 1, below. For each cell within the Table there are issues about the nature of (a) the recognition of factors that indicate the affective state of the student (Recognition); (b) the ways in which the system is able to respond to differing affective states (Capability); and (c) the ways the system chooses to respond (Performance).

Table 1. Different possible forms of design of affective technologies, setting forth a space involving the design of technologies taking student affect into account. The rows consist of the distinction between the two kinds of technology; and the three columns embody the distinction between the degree of adaptability of the technology. The degree of explicitness of the agenda and the model means that the horizontal line between the two rows is somewhat fuzzy.

	NON-ADAPTIVE	MACRO Adaptive or Adaptable	MICRO Adaptive
Tools for Learning	Tools designed to elicit a certain emotional response	Tools adaptable to user emotional *Traits*	Tools adaptive to user emotional *States*
Learner-Centered Learning environments	Learning Environments designed to elicit an emotional response	Learning environments adaptable to user emotional *Traits*	Learning environments adaptive to user emotional *States*

Responding to affect

Take a moment to think of the ways in which designing for affect has impacted your lives. What comes to mind? Maybe an exciting scene in a popular film? Or perhaps a website that instils confidence in the viewer. Or perhaps a darkened restaurant that sets a calming mood; contrast that image with a fast food diner. We now invite the reader to think of a software learning environment that responds to students' affect. What comes to mind? It might be the kind of nuanced perception of other learners as described in Chapter 2: perhaps a teacher-character engaging the student in a kind, encouraging way?

There are many more different ways that a learning system can address affect. At a more detailed level when designing macro and, particularly, micro-adaptive learning environments we need to consider the particular affective phenomena that the system should try to deal with. Are they clearly emotional such as "frustration", or perhaps more to do with motivation, such as the "value" placed on the outcome of the learning, or even related to meta-cognition such as "self-efficacy" (Avramides & du Boulay, 2009). These are not either/or categories, as many systems operate across the affective/motivational/meta-cognitive territory, often employing essentially relatively simple binary distinctions (e.g. frustrated or not frustrated, or confident or not confident) or even a single binary distinction between positive and negative overall student affect. Thus, design is not just about the tactics and strategy that the system might deploy, but also about the nature of the data about the students' affective states that is available, and the designer's beliefs about exactly how that data relates to students' states (does sitting still imply concentration or boredom, for example?). We envisage the system as having a number of channels of input that give clues about the affective state of the individual. These might include how hard they are working and how accurately, their physical and facial demeanour, what they say about how they are feeling, what they say about what their goals are, and the overall educational situation in which the interaction is taking place.

The learning system will need to consider some or all of these clues as to the affective states of the participants and then reason about (i) what those clues imply in terms of those affective states, and the (ii) what kind of reaction by the system would best move matters forward in a productive way. It might be that the student is starting to make a lot of mistakes and is looking a bit listless. If that student normally likes the kind of activity being undertaken, perhaps a bit of a chat about why things do not seem to be going as well as usual would be the best, or maybe an invitation to stand up and stretch for a moment might work too. By contrast, a different student showing the same symptoms may be finding the work too easy, or indeed too hard, and an adjustment to the task itself may be what is needed.

The space of potential clues to the student's affective state together with the space of possible reactions has been mapped by du Boulay et al. (2008). This space covers the cognitive, the metacognitive, the affective, the meta-affective (what the student understands about his or her affective state and the extent to which the student can regulate it), the physiological, the meta-physiological, the contextual and the meta-contextual.

For this chapter we have simplified the design considerations down to a focus on responses rather than including clues to affective state. We cluster responses into two broad categories: those that concentrate directly on affective matters ("emotion focused") and those that concentrate on cognitive reactions ("problem-focused"). We should note that concentrating on the affective directly (e.g. by being encouraging) may well have an effect on the cognitive state of the student and *vice versa*.

In broad terms, the kinds of question to be considered in designing affective systems include: for which affective states should the system gather information? what actions or changes should be utilized to address these states? How should the output of the affective theory be represented to the user (i.e., perhaps a smiley face from an embodied peer, or perhaps a reorganisation of a concept map to reflect the representations of the user's learning style).

Emotional response to affect

In order to promote positive affective states in students one must first determine what would be appropriate messages to transmit to students in order that they find the learning experience more congenial and beneficial, and/or are therefore more willing to persist working on the task, or perhaps on future tasks. Of course it is important to acknowledge that learning is sometimes hard work and frustrating. One possible response from the system is to help the student come to terms with this fact of life and help them find ways to deal with it in a mature manner. When conceptualizing the appropriate messages, there are myriad details to take into account: How should the system respond affectively to affective states or traits of negative valence? What should be done when a student is frustrated, or 'unmotivated'? Shall we praise students when they are doing well?

Carol Dweck's early research on human motivation sheds some light onto these questions. Her work focused on helpless and mastery-oriented response patterns in schoolchildren (Diener & Dweck, 1978, 1980). Some students persist in the face of failure while others quit as soon as the going gets tough. She discovered that students' implicit beliefs about the nature of intelligence have a significant impact on the way they approach challenging intellectual tasks: students who view their intelligence as an immutable internal characteristic tend to shy away from academic challenges. On the other hand, students who believe that intelligence can be increased through effort and persistence tend to seek out academic challenges (Carol S. Dweck, 1999b; Carol S. Dweck, Chiu, & Hong, 1995).

Students with an immutable view of intelligence place a high value on success because they worry that failure, or having to work very hard at something will be perceived as evidence of low intelligence. Consequently, they make choices that maximize the possibility that they will perform well. By contrast, students with an "incremental" theory of intelligence are not threatened by failure because they believe that their understanding can be increased through effort and persistence. As a result, these students tend to pursue academic challenges that they believe will help them grow intellectually (Carol S. Dweck, 1999b).

This research on the impact of praise suggests that teachers, parents and even technology may lead students to accept an entity view of intelligence. By praising students for their intelligence, rather than effort, many adults are sending the message that success and failure depend on something beyond the students' control. When these students perform well they have high self-esteem, but this crashes as soon as they hit an academic stumbling block. Students who are praised for their effort are much more likely to view intelligence as being malleable, and their self-esteem remains stable regardless of how hard they may have to work to succeed at a task. Therefore, these students are more likely to be willing to come through and reach their full academic potential (Carol S. Dweck, 1999a, 1999b).

Dweck's work suggests that messages that highlight the malleability of intelligence, instead of its inherent nature, could encourage a student to persist at a task. But how to apply this to the context of software learning environments? Burleson and Picard's work on educational software, where a character talks with the student while solving a Towers of Hanoi problem, emphasizes this

malleability of intelligence to the student. The idea was to add phrases such as “the mind is like a muscle and through exercise
and effort you can grow your intelligence” within a short dialog. This kind of intervention has shown promising results. This system is explored further under ‘case-studies’ in this section.

On another line of research, psychology literature suggests that *empathetic* responses might work well in situations were the student does not feel well about the learning experience. Responses that display other forms of emotional and social intelligence (such as synchronization with the user’s feelings) might also prove effective (Goleman, 2006a, 2006b). Teaching students how to cognitively recognize and cope with their own and others’ negative feelings falls into this general category, regardless of whether their behaviors are essentially physiological such as sweaty hands and heart rate, or feelings such as frustration and disappointment.

Indeed, feelings are *contagious*, the giving and receiving of feelings accompanies any human encounter. Studies have shown that when we register a feeling from someone else, there are signals in our brain circuits that imitate that feeling in our bodies (Hatfield, Cacioppo, & Rapson, 1994). Such research suggests that an optimistic pedagogical agent could possibly affect students in a positive way, making them more cheerful about the activity at hand.

Another promising approach in response to affect is to support students in the development of *meta-affective* abilities; that is, the degree to which the student can recognise and regulate their affective states. The system can encourage them to anticipate future affective states, such as nervousness before a presentation, or reflect on past affective states such as frustration when a task was too hard. Many students that suffer from math anxiety are not aware that they are suffering from anxiety, and are not practiced in regulating their emotional state.

Many of the approaches listed above are involved in some sort of therapeutic approach to help people achieve personal growth. One of those is traditional Rogerian Psychotherapy (Rogers, 1961), developed by the humanist psychologist Carl Rogers in the 1940s and 1950s. It is one of the most widely used models in mental health and psychotherapy. The basic elements of Rogerian therapy involve showing congruence (genuineness), empathy, and unconditional positive regard toward the patient, which in this case would be the student. Based on these elements the therapist creates a supportive, non-judgmental environment in which the client is encouraged to reach their full potential. Person-centered therapy is used to help a person achieve personal growth and/or come to terms with a specific event or problem they are having. The therapist encourages the patient to express their feelings and does not suggest how the person might wish to change, but by listening and then mirroring back what the patient reveals to them, helps them to explore and understand their feelings for themselves. The patient is then able to decide what kind of changes they would like to make and can achieve personal growth. The therapist’s role is mainly to act as a facilitator and to provide a comfortable environment but *not* to drive and direct therapy outcomes.

Such software could help students in this way by implementing an emotional therapeutic dialog with the student, in a similar way to the ELIZA software (Weizenbaum, 1966). ELIZA was a computer program designed in the sixties, which parodied a Rogerian therapist, largely by rephrasing many of the patient's statements as questions and posing them to the patient. ELIZA worked by simple parsing and substitution of key words into canned phrases. There is a clear potential for pedagogical agents to engage the students in such therapeutic dialogs.

Cognitive response to affect

Affective, cognitive and physiological states are intertwined (see e.g. Forgas, 2008). For instance, when a student tries to solve a problem that is proving challenging, the cognitive effort may also be reflected in feelings of fatigue, confusion or frustration, paralleled with signs of physiological stress. The extent of these feelings and ways of expressing them may depend on the individual.

Such responses are not necessarily negative. Indeed Forgas (2007) argues that certain kinds of cognitive activity are enhanced when someone is in a negative affective state. Human performance has been linked to levels of stress or arousal. When a student has a very low level of stress, or when a student has a very high level of stress, performance at any task will not be optimal. Rather, performance under a moderate level of stress optimizes results. This reveals a 2-dimensional inverted U-shaped relationship between stress and performance (Yerkes & Dodson, 1908): a student achieves optimal performance when neither bored at one extreme, nor anxious at the other end of the inverted-U-shaped curve.

In much the same way that a human tutor will naturally respond to affect by giving a simpler activity to an anxious student and a more challenging one to a bored student, Murray and Arroyo proposed that the challenge level of an activity within an Intelligent Tutoring Software can similarly be adjusted. When adding a second dimension of challenge level, a zone of optimal challenge is revealed, as shown in Figure 1. This parallelism between the feelings and challenge level of an activity can be seen in the work of Csikszentmihalyi (1990) who suggested optimal activities in the "flow" channel moving outward as skills are gained, and certainly before apathy sets in.

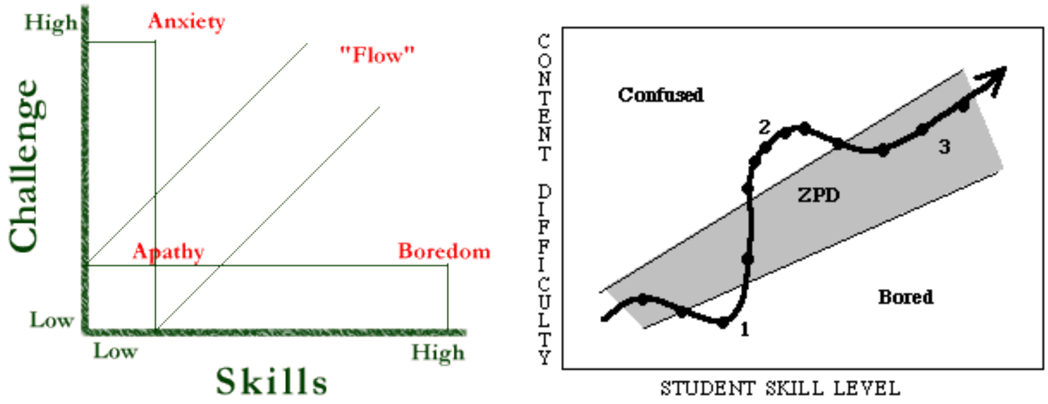


Figure 1. Adapted from Csikszentmihalyi (1990) and Murray & Arroyo (2002)

Similar to a human tutor, software might be designed to react to affect by adjusting the difficulty of the content, or by providing a worked-out example, a more thorough explanation, or a video of a person solving a similar (analogical) problem, thus making the task more manageable and easier to solve. Moving the student to a more "fun" part of the software is another possibility.

Meta-cognitive response to affect

Metacognition entails the knowledge and awareness of one's cognitive processes and the efficient use of this self-awareness to self-regulate these cognitive processes. Meta-cognitive scaffolding can encourage students to self-reflect and self-monitor by showing them their progress, and so self-regulate their problem solving process by helping them to carry out a plan to solve the problem. There is a clear potential for metacognitive feedback to help students maintain interest and engagement in the tasks at hand.

For example, Luckin and Hammerton's (2002) Ecolab II displayed back to students the degree of challenge that was inherent in the learning tasks that they had undertaken, as well as the extent of the help that they had sort. So the advice to a child who had undertaken only the easy tasks in the domain (ecology) and not asked for much help was to try the more tricky tasks. By contrast the the advice to the child who had tried very difficult tasks, but who had needed lots of help, was to ease back both on the difficulty of task and on the requests for help.

Another example of affecting students' affect with metacognitive feedback is Arroyo and colleagues' use of progress charts and metacognitive tips in a mathematics tutoring system (see the case-study in Section 4.3.6). Controlled experiments showed that the inclusion of these interventions in between the standard activities of the software helped students not only learn more but also exhibit higher appreciation of the software. The results of that study suggested that the reminder of students of their learning goal, and highlighting that they were making progress affected their attitudes towards the system and towards learning in general.

Other relevant research comes from the literature of teacher education. For instance, in the particular case of mathematics, Sheila Tobias (1995) was probably one of the first to address several myths about the fixed nature of mathematical intelligence. An interesting contribution of this book is the description of several interventions in math anxiety clinics across the United States. Many of those have to do with the de-mystification of math, but also several tools to help students develop "meta-affective" skills such as keeping affective diaries in parallel to their mathematics solutions to problems (e.g. how did I feel while solving that last problem? What can I do next time to not feel so frustrated?).

Contextual response to affect

Finally there is the issue of the physical and social context within which the learning takes place. It is important to recognise there is a strong cultural component to providing effective emotional support (Rosiek, 2003). This too can be adjusted. For example, peers can be introduced if the student is working alone, or removed if working in a group (this is part of what is called social scaffolding). The location and ambient characteristics can be adapted in various ways, *e.g.* turning music on and off, adjusting the lighting conditions, providing low-level distractions such as facing the student towards a window or an open door or not, or indeed moving from classroom to home to library to museum or wherever.

Case-studies

We conclude the chapter with 5 case-studies. These illustrate different ways that contemporary research is approaching the issue of design for affect. In order to provide some kind of non-technological baseline we start with a brief description of the work of Rosiek. There she describes ways in which teachers of a variety of disciplines have reframed tricky topics in ways that make them learnable. "Tricky" might mean issues with a high affective component e.g. race, or it might be issues that are abstract or difficult to grasp and need a very specific analogy tailored for that particular group of students.

Non-technological design

Rosiek (2003) reports on examples of ways that teachers have organised and framed lessons so as to provide emotional scaffolding to students. These lessons were taught by human teachers and computers were not involved. The reason for including this case-study is that it indicates future directions that the design process for learner-centered learning environments might go. Please note that the following is our interpretation of Rosiek's work.

Review of

Emotional Scaffolding: an exploration of the teacher knowledge at the intersection of student emotion and the subject matter

Jerry Rosiek
University of Alabama

Sometimes the "emotional scaffolding" in the title was provided to enable the students to identify better with the material to be learned, by framing it in more familiar terms, for example. Sometimes the scaffolding was provided to circumvent unhelpful negative reactions to the material, by focusing the students on equivalent but less emotionally-laden examples. At the heart of her paper is Table 1 below. This divides the scaffolding into two kinds: that designed to foster constructive emotions,
And that designed to reduce *unconstructive* emotions. The terms "constructive" and "unconstructive" are helpful in this context and rather better than the more usual "positive" and "negative", particularly as a normal arc of emotion in learning passes through what would normally be regarded as negative emotions. Moreover, Forgas (2007) argues that a mildly negative affective state enhances some kinds of cognitive processing.

In addition to promoting the constructive and reducing the unconstructive, the table is also divided into two rows to distinguish *implicit* approaches from *explicit* ones. This distinction is related to notions that have been introduced earlier. So for example, helping students focus on the "value" of what is to be learned is regarded as an implicit approach, as it treats the nature of what is to be learned. Helping the students recognise and deal with the emotions associated with learning directly is called an explicit approach, and has already been mentioned in Section 4.4.3 under the heading of meta-affective responses to affect.

Table 1. Implicit and Explicit Emotional Scaffolding (adapted from Rosiek, 2003)

Approach to Emotional Scaffolding	Attempts to Foster Constructive Emotions About the Subject Matter	Attempts to Reduce Unconstructive Emotions About the Subject Matter
Implicit	1. An effort is made to foster a constructive emotional response to the subject matter by associating it with something students find familiar or interesting	2. An effort is made to avoid triggering an unconstructive emotional response to the subject matter by approaching it in an unfamiliar context

Explicit	3. An effort is made to foster a constructive emotional response to the subject matter by drawing attention to it and offering students reasons why the effort to learn it is worthwhile	4. An effort is made to undermine an unconstructive emotional response to the subject matter by drawing attention to these emotions and making light of it or assuring students it is “not as bad as it seems”

In her paper, Rosiek provides telling accounts from work with teachers that illustrate each of the four cells of the table. These run from lessons in mathematics to the history of social and judicial inequality in the United States. From a design point of view, it is clear that in order to provide this kind of emotional scaffolding the (human) teachers needed to have a wealth of knowledge about their students: what they really enjoyed doing, how what they did enjoy doing could be related to the educational matter in hand, how they were likely to react emotionally to the topic to be learned and examples of how that topic could be slightly reframed to lessen and divert its emotional sting. While this would be hard to do automatically in the general case, given the current state of the art in Artificial Intelligence, special cases ought to be possible.

Ivon Arroyo and Benedict du Boulay

Designing around goal orientation

The next case-study looks at issues around how collaboration between pupils can be affected by their individual goal orientations and how the technology supporting the collaboration can take that into account.

Creating contexts for productive peer collaboration: some design considerations

Amanda Carr (nee Harris)[1], Rosemary LUCKIN[2] and Nicola YUILL[3]

[1]School of Human and Life Science, Roehampton University, UK
[2]London Knowledge Lab, Institute of Education, University of London, UK
3CHaTLab, Department of Psychology, University of Sussex, UK

(adapted from Harris, Yuill, & Luckin, 2007)

Research from psychology and education shows that working collaboratively in pairs or small groups can have positive effects on children's learning and development. However classroom observations suggest that the quality of children's collaborative discussion tends to be poor (Baines, Blatchford, & Kutnick, 2004). Much collaborative work in schools occurs around the computer and typically involves pairs or small groups working together (Light & Littleton, 1999). The frequency of this activity coupled with its challenges provide a natural context for developing systems to support co-located classroom interaction. However, in order for technology to fulfil this role we need to know more about when and why children's collaborative activity is more or less productive.

Achievement motivation and collaborative learning

Children differ a great deal in their motivation towards learning. Some children focus on understanding new material and developing competence (mastery-oriented). Other children focus on demonstrating their knowledge and gaining favourable judgements of ability (performance-oriented). These two orientations represent different underlying goals structures and are associated with important differences in the way children cope with difficulty and challenge (Ames, 1992). For example, mastery goals are associated with high levels of task engagement, the use of meta-cognitive and self-regulated learning strategies, effort and persistence (C. S. Dweck & Leggett, 1988). Performance goals, on the other hand, are associated with more surface-level learning strategies, the avoidance of challenging tasks and a concern with comparative evaluations of ability (Ames, 1992; C. S. Dweck & Leggett, 1988). Given these different approaches to learning there is a strong theoretical basis for predicting that achievement goals will influence collaborative behaviour in distinct ways and may therefore account for some of the variation in the quality of interactions observed in the classroom.

For example a child motivated by performance goals may find coordinating activity with a collaborative partner difficult as this inevitably minimises individual markers of achievement in favour of group progress. The social comparative nature of a performance orientation may, therefore, lead the child to perceive a collaborative context as either a forum for the public display of individual ability or a threatening environment in which a lack of ability may be exposed. This represents potential conflicts for the child between the demands of a collaborative task and individual perceptions of what might constitute success on that task.

On the other hand, the mastery-oriented child may find a collaborative context more appealing as they are more likely to view peers as sources of information rather than as a threat to their competence beliefs (Darnon, Muller, Schrager, & Pannuzzo, 2006). The high levels of task engagement associated with mastery goals may enable these children to engage with and participate in collaborative learning activities more productively.

Empirical Studies

We undertook two studies which, although used different methods, had a common aim; to identify the influence of achievement goals on children's behaviour when collaborating with a peer. Both studies used behaviour observation methodology which distinguishes our approach from other achievement goal work which relies on learners' self-reports after a particular learning experience (for example Darnon, et al., 2006). In both studies we observed same-gender pairs of primary-aged children (7 to 10 years old) interacting around a desktop computer using a single input device; a configuration children are most familiar with in the classroom. Each child's contribution to discussion was considered in relation to individual differences in goal orientation (Study 1) and contextual differences in an emphasis on mastery or performance goals (Study 2). In drawing together the results from both studies we will present a mastery- and a performance-oriented profile of collaborative behaviour.

Study 1

Twenty-two children from two urban primary schools in the South East England were divided into 11 same-gender, same-school pairs and observed using a software system called Joke City. The software prompts reflection on language ambiguity and humour in jokes and riddles. Two video recorded interactive sessions per pair, each between 15 and 30 minutes long, were analysed. Our coding scheme distinguished between those behaviours we identified as indicating productive interaction (*descriptions, justifications, disagreements)* and those indicating less productive interaction (*commands, submissions, dismissing and off-task behaviour*) as well as comments indicating meta-cognitive awareness and regulation (*description of self, other and joint understanding*). The proportional frequency of language categories for each child was summed and correlated with mastery and performance scores measured using a teacher-rated version of the Patterns of Adaptive Learning Scales (PALS)(Midgley, et al., 2000).

Results

The achievement goal scale yielded a mastery and a performance score for each child. The overall mean mastery score was 3.14 (SD .72) and the overall performance mean was 2.9 (SD .46). Our analysis revealed that mastery goals were positively correlated with disagreements ($r = .58$, $p < 0.01$) and negatively correlated with submissions ($r = -.38$, $p < 0.05$). Submissions involved children giving in to a partner without indicating understanding. We also found a relationship between performance goals and meta-cognition; performance goals were positively related to references to the child's awareness of their own knowledge or understanding ($r = .51$, $p <0.05$) while negatively related to references to their partner's knowledge or understanding ($r = -.49$, $p <0.05$).

These results are consistent with some of the key characteristics associated with mastery and performance goals identified in the literature and suggest which behaviours in particular, within a general mastery or performance profile, are relevant within a collaborative context. However, the range of mastery and performance scores measured using the PALS questionnaire suggested that neither goal orientation was particularly strong. We now turned to examining the extent to which it might be possible to manipulate goal orientation and foster a particular style of interaction.

Study 2

In this study we examined the influence of goal-oriented contexts on children's collaborative behaviour while playing a computer-mediated logical reasoning game called Zoombinis (Hancock & Osterweil, 1996). The game consists of a series of puzzles which players have to solve by using thinking skills such as identifying patterns and reasoning about evidence. This is done by combining and organising the Zoombini's features (e.g., hair colour, type of footwear) and different features of the environment.

In order to assess the influence of goal-oriented contexts we distinguished between children who had strong goal orientations from those who appeared neutral and therefore may be open to external motivational cues. From a sample of 61 children between 8 and 10 years, we identified 34 who were neutral in terms of goal orientation and were suitable for pairing. These children were matched in same-gender, same-class pairs (all children in this study came from a single primary school in a city in the South East of England). Eight pairs were randomly assigned to a mastery-focused condition and 9 to a performance-focused condition.

All pairs were given instructions to work together after which they received different instructions pairs about the object of the game depending on which condition they had been assigned. In the mastery-focused condition pairs were told that the object of the game was to work out good strategies for solving the puzzles and that it did not matter if they lost any Zoombinis in trying to achieve this aim. In the performance-focused condition pairs were told that the object of the game was to get as many points as possible and that each Zoombini was worth one point. Pairs were observed playing Zoombinis during a single session for approximately 25 minutes in length.

With reference to both the achievement goal and collaborative learning literatures we extended our coding scheme to focus specifically on behaviours which might characterise a collaborative interaction as being mastery or performance oriented. We were interested in three types of participation. Firstly, we distinguished between complex problem solving (*reasons, justifications and explanations*) and simple problem solving (*procedural, descriptive, suggestions without elaboration*). Secondly, we measured expressions of awareness of knowing and thinking and distinguished between metacognitive awareness of held knowledge or understanding (*I know*) and awareness of a lack of knowledge or understanding (*I don't know*). Thirdly, we were interested in exploring differences in children's help-seeking behaviours.

Results

Analysis was conducted by calculating proportional frequencies of each language category for each participant. In order to explore whether there were any differences in problem-solving language we conducted a mixed ANOVA with problem-solving type (simple and complex) within subjects and goal orientation between subjects. This revealed a main effect of problem solving type with children making more simple problem solving utterances overall (F(1. 30) = 432.41, $p < 0.001$). However there was a significant interaction between goal orientation and problem solving (F (1, 30) = 5.36, $p < 0.05$) where children in the mastery-focused condition made significantly more complex problem-solving utterances than children in the performance group. In relation to meta-cognitive awareness children in the performance group made more meta-cognitive positive comments (I know) (Mean = 9.6, SD = 4.5) than children in the mastery group (Mean = 7.4, SD 5.02). However, the large standard deviations suggest this was highly variable and the difference between groups did not reach significance. In relation to help seeking children sought help either by accessing the help function provided by the game (*task-focused*) or by requesting help from the researcher (*external help).* A count of each type of help for each child was measured. Chi square analysis revealed that children in the performance group were significantly more likely to use external help than children in the mastery group ($x^2(1) = 7.56$, $p <0.01$) but no difference between groups in use of task-focused help.

Discussion

In both studies, mastery goals were associated with behaviours more conducive to productive interaction and more likely to promote learning. In Study I, the stronger children's mastery goals were, the more they engaged in constructive disagreements and the less they tended simply to submit to their partner's suggestions. In Study II, mastery goals were associated with complex problem-solving which involved providing justifications and explanations for suggestions. In both studies therefore mastery goals appeared to engender a willingness to engage in a process of argumentation and discussion. The core feature of a mastery orientation is identified as the desire to gain understanding and master tasks without concern for making public mistakes or exposing a lack of ability (Ames, 1992; C. S. Dweck & Leggett, 1988). When children are collaborating, this motivation becomes evident in the level of the complexity of their discussion. Providing explanations and justifying one's perspective will necessarily carry with it the risk of exposing an incorrect solution or understanding of the problem. However, it is through making one's understanding explicit and discussing alternative perspectives that new insight is gained in the course of interaction (Wegerif, Mercer, & Lyn, 1999). Holding mastery goals seems to complement and support this process.

Similarities between studies suggest that a performance orientation may be associated with an individualistic style of peer interaction. In Study I, we found that performance goals were related to types of meta-cognitive talk which indicated a focus on the self. The stronger a child's performance goals the more likely they were to make self-referring meta-cognitive utterances and the less likely they were to refer to their partner's knowledge or understanding. In Study II, children in the performance goal group also used a higher number of self-referring meta-cognitive statements. These results suggest that holding performance goals may inhibit a child's ability to engage in the coordinated, joint activity associated with productive collaborative interaction (Roschelle & Teasley, 1995). A core element of a performance orientation is the desire to demonstrate ability (Ames, 1992; C. S. Dweck & Leggett, 1988). Collaborative partners may therefore provide the performance-oriented child with the opportunity to affirm their own competence.

Implications for design

The above discussion suggests two issues for designing technology for collaboration in the classroom. Firstly, children will behave differently during collaborative interactions depending on whether they are pursing mastery or performance goals. Secondly, most children are very responsive to goal-oriented cues embedded in the learning context. These findings suggest that a) it is crucial to consider goal orientation when constructing a model of the learner and b) using goal-oriented cues in the design of the collaborative context may encourage more adaptive approaches.

Modelling collaborative learners

A collaborative learning context poses a challenge to any teacher, either human or machine, as the characteristics of not one but two or more learners need to be considered simultaneously. The machine teacher is at a further disadvantage given the constraints of technology to 'listen' to the verbal exchange that is fundamental to such a learning context. With this limitation in mind we suggest three categories of goal-oriented behaviours which a system could detect and monitor:

- Help-seeking: Children pursuing performance goals will tend to look for help that offers solutions and also provides reassurance.
- Complex problem-solving: Children pursuing mastery goals tend to engage in more complex problem-solving in which they provide justifications and explanations for their suggestions. One approach which could be applied and which has been borrowed from asynchronous collaborative environments, is to provide sentence openers for participants to select in order to give the system an indication of the nature of the discussion (Robertson, Good, & Pain, 1998).
- Disagreements: Mastery goals are more likely to be related to constructive disagreements. Recent developments in how collaboration is represented through technology, for example through interface design and hardware configuration (Kerawalla, Pearce, Yuill, & Harris, 2008), affords the possibility of monitoring both individual activity (through separate input devices) as well as the process of coordinating that activity (interface representations of agreements and disagreements).

Designing collaborative contexts

We identified earlier the difficulties many children have with engaging constructively in collaborative activity. In designing technology we need to be mindful of these difficulties and design in a manner that supports the *development* of collaborative skills. Our classroom observations suggest that mastery goals may promote more productive interactions and our experimental manipulations suggest that it is possible to encourage children to adopt particular goals. Building on these findings we suggest that fostering a mastery context could be achieved through techniques such as:

- Motivational instructions for the task: Inherent in our instructions were messages about the purposes and meaning of the collaborative task. If we want to create systems which encourage children to engage collaboratively in deep learning, it would seem more appropriate to foster environments in which learning itself is the goal rather than receiving external rewards.
- Scaffolding goal-oriented contexts: The emphasis of scaffolding prompts may be designed to highlight particular goals for the interaction. For example, feedback and assistance which emphasise understanding, problem-solving and learning may be preferable to feedback which focuses on praise for ability. There is also the potential to use goal orientation as an additional level of fading within a more general scaffolding structure. A method which has the possibility of integrating personal goals (detected through a learner model) and contextual goals (provided by the software scaffolding).

Conclusions

Interactive learning environments are increasingly being designed with an emphasis on the motivational and affective state of the learner (du Boulay & Luckin, 2001). In this case-study we have presented evidence which suggests that children's achievement goals are crucial to consider within this broader framework. Mastery and performance goals influence collaborative behaviour in distinct ways; mastery goals promote collaborative discussion and argumentation while performance goals encourage an individualistic style of interaction. We also highlight the role of the learning context in shaping the types of goals children pursue. This suggests that the motivational aspects of context can also be designed to support the development of children's collaborative skills.

Addressing affect with empathetic learning companions

The next case-study shows how the learning companions can address frustration and have students persevere at a task. This is done by physically mirroring students' posture, and with dialogs that emphasize the malleability of intelligence.

Learning Companions: Strategies for Real-time Affective Support

Winslow Burleson[1], Kasia Muldner[1], Rosalind Picard[2]
[1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Phoenix, USA
[2]MIT Media Lab, Cambridge, USA

Traditionally, intelligent tutoring systems were designed to support learning by providing domain-related advice, such as hints on formulas needed to solve a match or physics problems (e.g. Anderson, Farrell, & Sauers, 1984; Conati, Gertner, & VanLehn, 2002; Heffernan & Koedinger, 2002). Consequently, while many of these tutors were "intelligent" about a particular domain, they were mainly ignorant when it came to supporting users' *affect*, i.e., the emotional tone a person expresses. Such tutors run the risk of hindering optimal learning outcomes by being intrusive or inappropriate, for instance by failing to detect and so appropriately respond to users' frustration.

To overcome the limitations of purely task-based tutors, we are working on devising automated learning companions that provide affective support tailored to a given student's needs - Figure 1 shows one such companion, which we refer to as *Casey*. The motivation for our work is rooted in research showing that affective support plays a key role in human tutoring: approximately 50% of expert tutors' interactions with their students are affective in nature (Lepper, Woolverton, Mumme, & Gurtner, 1993). There is also research demonstrating that affective support and social bonding between teachers and students have considerable impact on learners' performance and motivation. For instance, caring relationships between middle-school children and their teachers are predictive of learners' performance (Wentzel, 1997). Given that learning is facilitated when tutors know when to provide affect- vs. task-based support (e.g. Higgins, 2001), it is paramount that pedagogical agents are capable of detecting users' affect, as well as other states such as expertise, and tailoring the interaction accordingly.

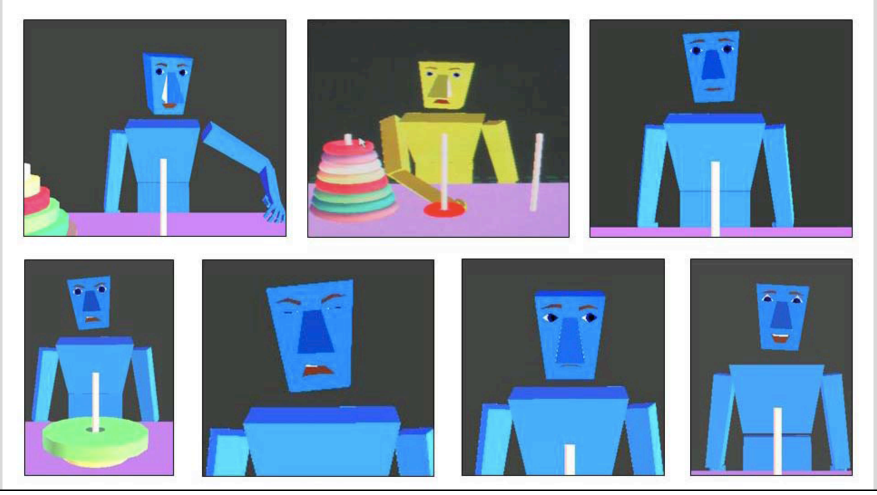


Figure 1. the Casey Learning Companion

The first step in proving tailored affective support is detection of users' affective states. To maintain as much as possible a natural interaction with the system, detection requires the use of non-obtrusive sensing devices. Such devices can obtain information about a user that is a natural by-product of that user's interaction with the computational system. Accordingly, the *Casey* agent is embedded in the Affective Learning Companion (ALC) platform (Burleson & Picard, 2004), which incorporates a variety of sensors to leverage as much as possible data on a user's affective states (see Figure 2). These sensors have been developed and validated by the Affective Computing Group, and include a pressure mouse, skin conductance sensor, posture chair and camera. The pressure mouse detects the intensity of the user's grip on the mouse, which has been shown to correlate to frustration (Dennerlein, Becker, Johnson, Reynolds, & Picard, 2003). The skin conductance sensor is a well-established indicator of user arousal (Boucsein, 1992). The posture chair sensor can be used to classify motivational states such as engagement and boredom (Mota & Picard, 2003). The facial-expression camera can measure head nod/shake, mouth fidgets, smiles, blink events, and pupil dilations (Kapoor, Qi, & Picard, 2003). The ALC platform relies on these sensors' readings to classify and subsequently respond to users' affect via the embedded Casey agent, as we will describe shortly.

While unobtrusive, sensing devices do come with their own set of challenges, such as they produce vast quantities of data that is noisy and difficult to map to users' affective states. To address this challenge, we rely on sophisticated machine learning techniques to classify users' affective states. For instance, one of our projects focused on investigating detection of user frustration (Kapoor, et al., 2007). Our results show that the system reached 79% accuracy in classifying user frustration.

Given data on a user's affective states, the final step includes actually delivering the affective support. As this support will shape the user's experience with the system, including how much she learns from it and how she feels while doing so, this support needs to be designed and evaluated in a principled fashion. Here, we will focus on describing our experience in designing and evaluating affective support that we have embedded into the ALC framework.

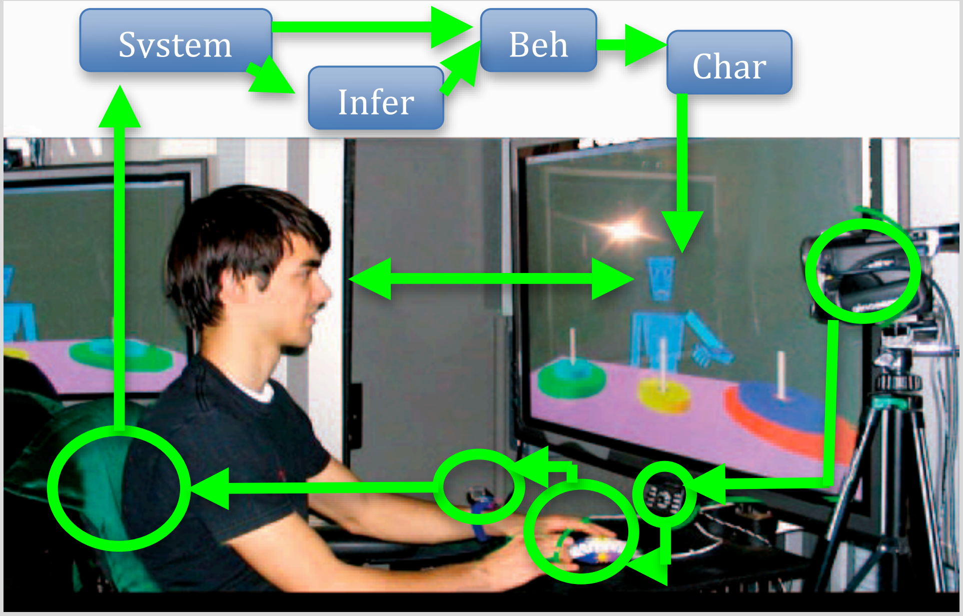


Figure 2. the ALC Sensor Framework

Affective support in the ALC framework: sensor-driven mirroring & affect-based interventions

No automated system today can reliably detect or respond to all the emotions that occur during learning. As we describe below, the Casey agent currently provides two kinds of affective support. While this support is only a few of the myriad possibilities, given that ours is the first experiment to implement real-time electronic-character responses to users' affective states, it is important to analyze the affective support's impact prior to incorporating additional behaviors. Specifically, the affective support delivered by Casey includes sensor-driven *non-verbal mirroring* and *affective interventions*.

Non-verbal mirroring is the behavioral mimicry of others' behaviors, and is an integral part of human interactions. In the ALC platform, mirroring is realized by having the agent (1) mimic aspects of the learner's facial expression according to data provided by the camera, (2) sway in an agitated manner in response to proportional to the pressure a user applies to the mouse, (3) skin tone adjust relative to skin conductance values, such as redden when the value is high, and (4) lean forward or back according to the chair sensor, to mirror a user's posture.

The other behavior we incorporated into the ALC platform corresponds to affect-based interventions that are delivered verbally by the pedagogical agent. One of these interventions is directly rooted in Dweck's work (Carol S. Dweck, 1999b) on helping learners develop meta-affective skills, i.e., the ability to coordinate one's affect via various strategies. In particular, Dweck has shown that one strategy that improves meta-affective skills involves thinking of the mind as being like a muscle and believing that this can increase one's intelligence through effort. Accordingly, one type of message the agent delivers is directly based on Dweck's "the mind is like a muscle" message (e.g. "*Remember, the mind is like a muscle that when exercised may not feel good, but it is getting stronger through exercise*"). The agent also delivers empathetic messages with respect to the learners' levels of frustration (e.g., "*It sounds like you are extremely frustrated with this activity").*

It is important to note that the ALC framework generates both types of affective behaviors adaptively in real time based on the sensor data, making these behaviors tailored to a given learner's affective states. The intention behind the design of both types of behaviors is to provide support for users' *affective* states, instead of task-specific support, as is the case with many of today's tutors. While in the future we do anticipate that an approach integrating both types of support will be beneficial, we first need to gain a better understanding of computer-based support for affective interventions.

How does Affect Support Impact Learners' Experience?

To assess the impact of affective support, we conducted a controlled evaluation with children aged from 11 to 13 from three semirural schools in western Massachusetts. The data we present here is based on 61 of these children who interacted with the Casey agent in the ALC platform in the context of a challenging problem-solving activity: the Tower of Hanoi game (with seven disks; see Fig. 1). To evaluate the impact of mirroring and affective interventions, during the study we compared them with the non-affective counterparts, as follows.

Non-verbal movements: mirroring vs. pre-recorded behaviors. The agent followed one of two strategies for its nonverbal movements: (1) mirroring (described above), or (2) pre-recorded behaviors, generated from one of five recorded files of the "average" user interactions (determined from an earlier pilot study using the standard deviation of each behavioral channel). Consequently, both cases involved the character generating nonverbal movements, but in only the first case were these synchronized to the learner's current sensor outputs.

Tutorial interventions: affective vs. task-based. The agent generated one of two types of verbal interventions (1) *affect-based intervention* (e.g., delivery of Dweck's "mind-is-like-a-muscle" message) or (2) *task-based intervention*, e.g., "*Another way to think about this is to think about the small disks that are in the way. If you move these out of the way, you can move the disk that you want to move.*"

The evaluation corresponded to a 2x2 design with a total of four conditions (mirroring/affect-based intervention, mirroring/task-based intervention, pre-recorded behaviors/affect-based intervention, pre-recorded behaviors/task-based intervention), with participants assigned randomly to the conditions. We now present the methodology and results – here, we describe only the key aspects, full details may be found in (Burleson & Picard, 2007).

Methodology. During the study, the participants first completed a pretest to obtain information on their theories of self-intelligence and goal mastery orientation. Next, the Casey agent showed a slide show that introduced each participant to the study and to Dweck's meta-affective message, using a script that Dweck has shown shifts children's beliefs about their own intelligence toward incremental self-theories. During the slide show, the agent either mirrored users' behaviors or relied on pre-recorded behaviors, depending on which condition the participant was assigned to.

Following the slide show, the agent presented the Towers of Hanoi game, and instructed the participant to "*click on a disk to start whenever you want, I'll just watch and help if I can*." Each participant was given four minutes to engage with the activity before the character intervened with either an affect- or task-based intervention. During the intervention, the agent also asked the participants to self-report on their affective and motivational states (e.g., "*On a scale from 1 to 7, how frustrated are you feeling right now?*"). At this point, the agent informed the learner that it had to go, allowing the learner to respond via one of three buttons: "*OK, bye*"; "*OK, bye I was glad to have you here*"; or "*OK, I'm glad you are finally going*." After the learner selected a bye response or 20 seconds elapsed, which ever came first, the companion disappeared and three quit buttons appeared, offering the learner an opportunity to end the activity. The three buttons were labeled "*I want to stop because I'm too frustrated to continue*," "*I've put in all the effort that I can and want to stop*" and "*I want to stop for some other reason*." The participants could quit at any point, or continue with the activity for up to 15 minutes, at which point the ALC framework ended the activity.

After the interaction with the ALC framework, each participant completed some post-tests to assess their affective experience, including impressions of the character (via a modified Working Alliance Inventory), as well as theories of self-intelligence. As a final step, participants were given the opportunity to re-engage with the activity for two more minutes, to assess the user's intrinsic motivation.

In the following, we refer to an affective agent as being "more emotionally intelligent" when it engages in sensor-driven nonverbal mirroring or when it provides affect-based interventions than when it provided neither. We considered these interventions and the mirroring condition additive; in other words, an agent that provided both mirroring and affective interventions was more "emotionally intelligent" than one that provided either one separately, or neither.

Results

Given that our goal was to evaluate the impact of sensor-driven mirroring and affective interventions, we compared these against their counterparts (pre-recorded behaviors and task-based interventions). The measures we used focused on affect rather than learning, and were obtained from self-report surveys, learners' responses during dialogue with the companion, and learners' behavior, such as duration of engagement and of reengagement in the activity.

We did not find any significant differences between the conditions, indicating that affective support *on the whole* did not have an impact. We did find, however, significant interactions between the conditions with respect to gender, as well as interesting gender differences, which we now describe.

Perseverance on Task & Motivation. We found a trend indicating a gender difference in terms of the strength of the mirroring effect (p=.07). Girls with mirroring reported *slightly* higher willingness to continue working on task than girls without mirroring. In contrast, boys with mirroring reported being *much more* willing to continue working on the task than boys without mirroring. In fact, for boys, the impact of sensor-driven mirroring on willingness to persevere on task showed a trend towards significance (p=0.065). When we checked time on task, however, we did not find that mirroring had an impact on either gender, although overall girls did persevere more than boys (p=0. 016).

We also checked how user motivation was influenced by the *congruence* of the intervention that the agent generated mid-activity, i.e., how appropriate the intervention provided was with respect to a user's actual self-reported frustration level. For girls, the degree of an intervention's congruence significantly influenced their intrinsic motivation, while for boys, there was no significant difference.

Impressions of the Agent. One way to measure the children's impressions of the agent is via their bye-bye button responses (which ranged from "*OK, bye*" to "*OK, bye I was glad to have you here*" to "*OK, I'm glad you are finally going*"). The agent's intervention had opposite effects for boys and girls with respect to the *bye-button* responses: boys responded more positively with task support than boys with affect support (i.e., tended towards the "*glad to have had you around*" response), while girls had the opposite reaction (i.e., tended towards the "*glad you are going*" response when given task support). These results are confirmed by the modified Working Alliance Inventory: boys had more positive impressions of the task-support character than the affective support character, while girls trended toward the opposite response.

We also found some other interesting gender differences. Boys were less likely to agree that knowing the mind is like a muscle could be helpful than girls (p=0.03). Girls were more likely to believe they would be able to use the strategies presented by Casey (P=0.003).

User Frustration. We found that affective interventions had an impact on participants' frustration, as assessed by self reports and post-activity questionnaire. For girls, the interaction between intervention and mirroring was highly significant (p<0.01), indicating that girls who received mirroring and affect-based interventions had lower post-activity frustration than girls who received mirroring with task-based interventions. Girls without mirroring, i.e., who experienced pre-recorded behaviors, had the opposite relationship with the intervention: girls who received affect-based interventions had higher levels of post-activity frustration than girls who received task-based interventions.

Boys also showed significant differences when grouped by intervention (p=0.009): they showed twice as much post-activity frustration when they received affect-based interventions, as compared to task-based intervention. There was also a trend toward significance for mirroring: boys who received mirroring reported a third less frustration than boys who didn't (p=0.061).

Regardless of the type support (affective, non-affective), overall girls reported being less frustrated than boys at the end of the activity (p=0.01).

Meta-Affective Reports. Regardless of the level of frustration, we found a trend indicating that participants of both genders who received *affect support* also self-reported higher levels of *Flow* and lower levels of *Stuck (*p=0.065). Furthermore, for girls only, affect-based intervention correlated positively with meta-affective skill (p = .040, r = .37), and with more *Flow* (p = .006, r = .52).

Overall, girls self-reported as experiencing less *Flow* and more *Stuck* than boys *(p=.017)*.

Another gender difference relates to the relationship between *Flow*/*Stuck* and meta-affective skill. While we did not find that these two measures correlated significantly when assessed across both genders, they did correlate when assessed with only girls (p = .010, r = .49). Assessment of these measures for boys also shows a significant but negative correlation (p = .021, r = –.40). Boys' meta-affective skill correlated significantly with perseverance (p = .048, r = .34), whereas there was no significance for girls.

Implications of the Findings

While we showed that mirroring had a positive impact on both boys and girls, in terms of their self-reported willingness to persevere on task, we also found that affective support in some instances had complimentary impact that depended on gender. For instance, the social bond that girls and boys developed with the learning companion depended both on the type of intervention and gender. Boys responded more positively to and had more positive impressions of the learning companion providing task interventions than the character providing affect-based interventions, while girls showed the opposite pattern.

The interventions that influenced students' social bond with the agent likely also influenced their frustration. Reflecting the finding that boys' bond was stronger with the agent that delivered task-based interventions, boys reported significantly lower frustration levels with task-based interventions, as compared to affect-based interventions. Although girls showed no major difference in the frustration level based on the type of intervention, further analysis indicated a more complex relationship and significant differences due to the interaction of the type of intervention and the presence of mirroring. We can explain these differences in terms of the "coordination" of the different elements of the character's emotional intelligence. Girls who experienced an affect-based intervention *in conjunction* with mirroring (case A) had lower frustration than girls who received *either* (1) an affect-based intervention without mirroring (case B) or (2) task-based intervention with mirroring (case C). This result may be explained by the observation that in contrast to cases B and C, in case A the mirroring and intervention are coordinated so that the character displays higher levels of emotional intelligence than in the other two cases.

We found one of the key gender differences occurred in the relationship between meta-affective skill and *Flow*/*Stuck*. There was no significant correlation across both genders, but a strong correlation between girls' meta-affective skill and more *Flow*, and a strong correlation in the opposite direction for boys. Here, grouping the genders clearly mixes different gender effects, yielding no significance when assessed together. One possible hypothesis for this discrepancy between genders is that girls at this age might be better able to assess their own emotions than boys. If girls are better at assessing their emotions, they might be better able to use their meta-affective skill to lead themselves to more flow, less stuck. On the other hand, although boys might report that they have meta-affective skill, they might actually be less able to recognize their own emotions and thus apply them to their own experiential benefit.

A final word

As intelligent tutoring systems (and other systems using relational-agent strategies) advance to incorporate greater emotional intelligence, developers and researchers should be able to enhance their systems and learners' experiences by incorporating elements of emotional intelligence. At the same time, developers and researchers must be careful to appropriately coordinate the diverse elements of emotional intelligence and be well aware of the differences in the impact of these elements on 11- to 13-year-old boys and girls.

Our investigation has highlighted how various factors related to affective support impact users' experience. In particular, the type of intervention, its level of congruence with respect to a learner's frustration and the presence or absence of sensor-driven non-verbal mirroring influenced the participants’ frustration levels, meta-affective skill, amount of *Flow* and *Stuck*, and intrinsic motivation. Our findings highlight that the various elements of an adaptive learning companion’s emotional intelligence should be presented in a coordinated way, as inconsistencies increased both girls’ and boys’ frustration.

Toward addressing affective states with production rules

The next case-study looks at ways of both detecting and responding to students' affective states such as boredom, confusion and frustration as part of the family of tutors known as AutoTutor.

Case-study: Affect-Sensitive AutoTutor

Sidney D'Mello
Institute for Intelligent Systems, University of Memphis, Memphis, USA

Introducing AutoTutor

AutoTutor is an intelligent tutoring system that helps learners construct explanations by interacting with them in natural language and helping them use simulation environments (A.C. Graesser, Chipman, Haynes, & Olney, 2005; Arthur C. Graesser, et al., 2001). AutoTutor helps students learn Newtonian physics, computer literacy, and critical thinking skills by presenting challenging problems (or questions) from a curriculum script and engaging in a mixed-initiative dialog while the learner constructs an answer. AutoTutor has different classes of dialogue moves that manage the interaction systematically. AutoTutor provides feedback on what the student types in (positive, neutral, or negative feedback), pumps the student for more information, prompts the student to fill in missing words, gives hints, fills in missing information with assertions, identifies and corrects misconceptions and erroneous ideas, answers the student's questions, and summarizes topics. During the tutorial dialogue, AutoTutor attempts to elicit information from the learner by first providing hints, then prompts and finally states the missing information to the learning via assertions. A full answer to a question is eventually constructed during this dialogue, which normally takes between 30 and 100 turns between the student and tutor.

The impact of AutoTutor in facilitating the learning of deep conceptual knowledge has been validated in over a dozen experiments on college students as learners for topics in introductory computer literacy (Arthur C. Graesser et al., 2004), conceptual physics (VanLehn et al., 2007), and critical reasoning on scientific methods (Storey, Kopp, Wiemer, Chipman, & Graesser, in press). Tests of AutoTutor have produced gains of .4 to 1.5 sigma (a mean of .8), depending on the learning measure, the comparison condition, the subject matter, and version of AutoTutor.

While the current versions of AutoTutor adapt to the cognitive states of the learner, the adaptivity bandwidth of the affect-sensitive AutoTutor would be expanded to be responsive to the learners' affective states. We envision a micro-adaptive strategy where AutoTutor would detect and intelligently respond to the affective states of boredom, engagement/flow, confusion, and frustration. These are the affective states that are most prominent during tutorial sessions with AutoTutor (S. Craig, et al., 2004; S. K. D'Mello, et al., 2006; A. Graesser, et al., 2007; A. Graesser, et al., 2006).

Detecting Learners' Affective States

Our affect detection system monitors conversational cues, gross body language, and facial features. The system uses supervised machine learning methods for affect classification. Training data were collected in a learning session with AutoTutor (N = 28), after which the affective states of the learner were rated by the learner, a peer, and two trained judges.

Conversational cues (dialogue features). A one-on-one tutoring session with AutoTutor yields a rich trace of contextual information, characteristics of the learner, episodes during the coverage of the topic, and social dynamics between the tutor and learner. These dialogue features are computed in real time for each student-tutor turn (i.e. student submits response, tutor provides feedback, tutor presents next question) and are used to predict the affective states of the learner. For example, boredom occurs later in the session, after multiple attempts to answer a question, and when AutoTutor gives more direct dialogue moves (i.e. assertions instead of hints). Alternatively, confusion occurs earlier in the session, within the first few attempts to answer a question, with slower and shorter responses, low quality answers, when the tutor is less direct in providing information (i.e. with hints instead of assertions), and when the tutor provides negative feedback (D'Mello, Craig, Witherspoon, McDaniel, & Graesser, 2008).

Gross body language (posture features). We use the Body Posture Measurement System (BPMS), developed by Tekscan™, to monitor the gross body language of a learner during a tutorial session with AutoTutor (see left panel of Figure 1). The BPMS consists of a thin-film pressure pads that are mounted on the seat and back of the learner's chair. The system provides a real time pressure map of the spatial distribution of pressure exerted on the pads. The learners' affective states are then tracked via representative configurations that the body adopts during affective experiences. For example, engaged learners tend to lean forward while bored learners tend to lean back. Confused and frustrated learners adopt an alert position, where they sit upright in a tense state (S. D'Mello, Taylor, & Graesser, 2007).

Facial feature tracking. We used the IBM BlueEyes system developed by Picard and colleagues (Kapoor & Picard, 2005) to monitor facial expressions for affect recognition (see top panel of Figure 1). The system locates and tracks the pupils of the eye and brows in real time. It also labels facial action units (muscle movements in the face, Ekman & Friesen, (1978). The action units are then linked to the different emotions. For example, confusion is expressed by a lowered brow and the tightening of the eye lids (S. D. Craig, D'Mello, Witherspoon, & Graesser, 2008; McDaniel et al., 2007).

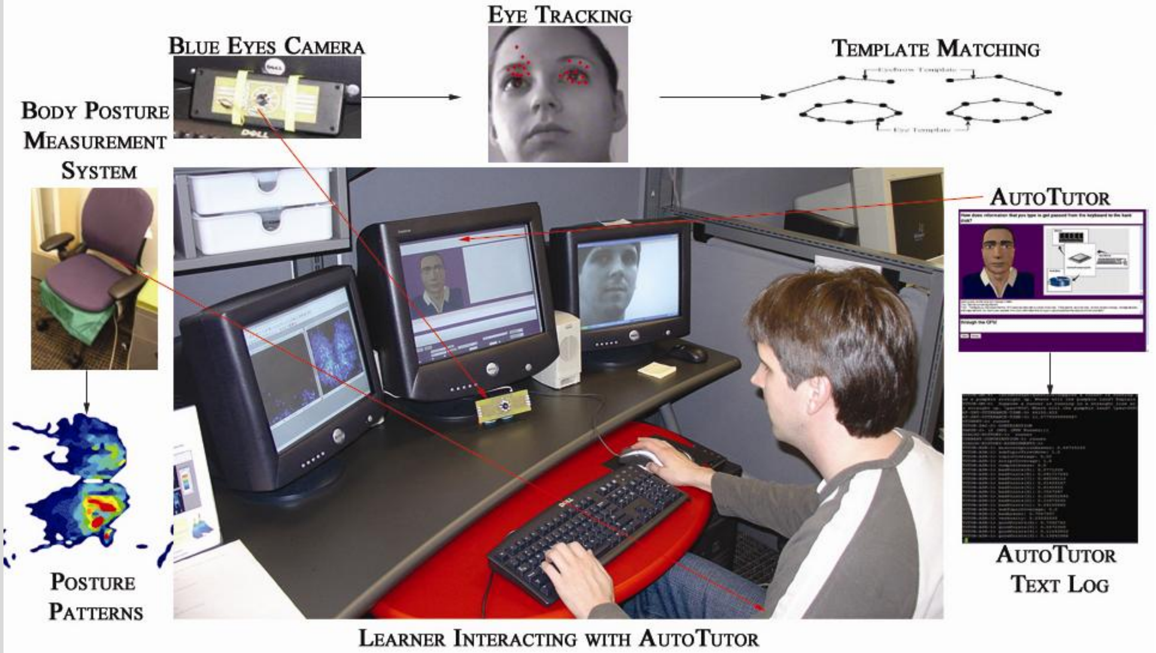


Figure 1. Sensors used for affect detection as learner interacts with AutoTutor

Classification accuracy. We are currently exploring some of the technical challenges associated with the automated detection of the facial expressions. However, experimental simulations with standard classifiers (e.g. logistic regressions, Bayesian models, neutral networks) indicate that conversational cues and gross body language are viable channels for affect detection. Conversational cues alone yielded accuracies of 69%, 68%, 71%, and 78%, in individually detecting boredom, confusion, flow, and frustration from neutral (chance=50%) (S. D. Craig, et al., 2008). Classification accuracies obtained from gross body language were 70%, 65%, 74%, and 72% in detecting boredom, confusion, flow, and frustration versus the neutral baseline (baserate = 50%) (S. D'Mello, Graesser, & Picard, 2007). Taken together, classification accuracies are 73% when each affective state is aligned with the optimal sensory channel.

Responding to Learners' Affective States

We are in the process of fortifying AutoTutor with the necessary pedagogical and motivational strategies to address the cognitive and the affective states of the learner. We are implementing a set of production rules that addressed the presence of boredom, confusion, and frustration (i.e. the negative states) by amalgamating perspectives from goal theory (Carol S. Dweck, 2002; Stein & Levine, 1991), attribution theory (Batson, Turk, Shaw, & Klein, 1995; Heider, 1958; Weiner, 1986) and cognitive disequilibrium theory (Festinger, 1957; Arthur C. Graesser & Olde, 2003). In addition to theoretical considerations, the assistance of experts was enlisted to help create the set of tutor responses. Two experts in pedagogy, with approximately a decade of related experience each, were provided with excerpts from real AutoTutor dialogues (including both the tutor and student dialogue content, screen capture of the learning environment, and video of the student's face as illustrated in Figure 1). The experts were instructed to view each of the excerpts and provide an appropriate follow-up response by the tutor.

The production rules are designed to map dynamic assessments of the students' cognitive and affective states with tutor actions to address the presence of the negative emotions. There are five parameters in the student model and 5 parameters in the tutor model. The parameters in the student model include, (a) the current affective state detected, (b) the confidence level of that emotion classification, (c) the previous affective state detected, (d) a measure of student ability (dynamically updated throughout the session), (e) the conceptual quality of the student's immediate response (computed via Latent Semantic Analysis (Landauer & T.Dumais, 1997). AutoTutor incorporates this 5 dimensional assessment of the student and responds with: (a) feedback for the current answer, (b) an empathetic and motivational statement, (c) the next dialogue move, (d) an emotional display on the face of the AutoTutor embodied pedagogical agent, and (e) an emotionally modulated voice produced by AutoTutor's text to speech engine.

As a complete example, consider a student has been performing well overall (high ability), but the most recent contribution was not very good (low current contribution quality). If the current emotion was classified as boredom, with a high probability, and the previous emotion was frustration then AutoTutor might say the following, "Maybe this topic is getting old. I'll help you finish so we can try something new". This is a randomly chosen phrase from a list that was designed to indirectly address the student's boredom and to try and shift the topic a bit before the student becomes disengaged from the learning experience. This rule fires on several different occasions, and each time it is activated, AutoTutor selects a dialogue move from a list of associated moves. In this fashion, the rules are context sensitive and are dynamically adaptive to each individual learner.

The empathetic and motivational statement represents AutoTutor's attempt to address the presence of negative emotional states. The AutoTutor animated conversational agent also expresses different emotional states which convey an emotion to the user. These states include surprise, delight, disappointment, compassion, and skepticism. For example, surprise is displayed when the tutor detects a degree of novelty such as a low domain knowledge student providing an exceptionally good answer. Expressions of delight parallel positive feedback statements while disappointment accompanies expressions of negative feedback. The tutor displays compassion when a learner has been making a serious attempt but is having trouble grasping the material. Finally, in addition to the aforementioned affective facial expressions, AutoTutor produces affective speech by modulating its pitch range, pitch level, and speech rate via the Speech Synthesis Markup Language (SSML).

Evaluating the Affect-Sensitive AutoTutor

The affect-sensitive AutoTutor aspires to keep students engaged, boost self-confidence, and presumably maximize learning by narrowing the communicative gap between the highly emotional human and the emotionally challenged computer. In order to test whether an affect-sensitive cognitive tutor is effective, we will compare two different versions of AutoTutor: one that is sensitive to learner emotions and one that is not. The original AutoTutor has a conventional set of fuzzy production rules that are sensitive to cognitive states of the learner, but not to the emotional states of the learner. Our improved AutoTutor is sensitive to these affective states. The obvious prediction is that learning gains and the learner's impressions should be superior for the affect-sensitive AutoTutor.

Enhancing Affect with Meta-cognitive Progress Charts and Tips

The next case-study shows how the affective dimension has been enhanced by showing students their progress as they learn with tutoring software for mathematics, Wayang Outpost.

Case-study: Progress Charts and Tips

Ivon Arroyo and Beverly P. Woolf
University of Massachusetts, Amherst, USA

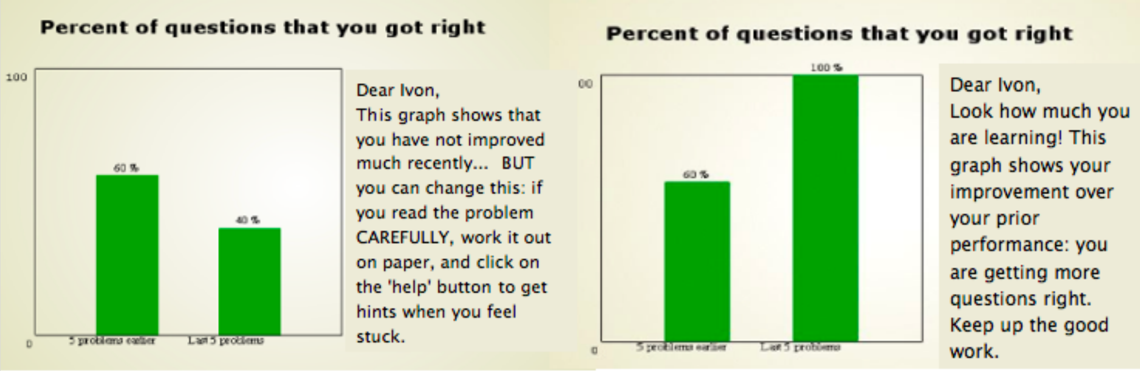
One possibility for software to respond to students' affect is to introduce metacognitive tools that help students reflect about the learning experience. This case-study introduces Progress Charts that display student's learning, accompanied by metacognitive tips in a mathematics tutoring system (Arroyo, Beal, Murray, Walles, & Woolf, 2004). Giving detailed meta-cognitive messages between problems, it was hypothesized, could guide students not only to improve learning, but especially to improve affective outcomes: improve students' attitudes for learning (mastery-learning), perceptions of the tutoring system, self-efficacy in mathematics and value attributed to mathematics. In addition, we hypothesized that these interventions could potentially reduce student gaming.

Progress Tips in Wayang Outpost

Wayang Outpost is a web-based multimedia tutoring system for geometry (Arroyo, et al., 2004) that helps students solve challenging standardized tests problems. This system is used in real public school settings. Each day in math class, students log on and are directed towards the different modules for pre/post-testing and tutoring.

Wayang tracks levels of disengagement (or gaming) along time using a Hidden Markov Model (Johns & Woolf, 2006). Student engagement is defined as a discrete, dynamic variable. Engagement on the current problem depends on student behavior on that problem, and on the level of engagement during the previous problem. The engagement variable can take on three values: 1. gaming by exhausting the hints to reach the final hint that gives the correct answer in a brief period of time (hint abuse); 2. gaming by quickly guessing answers to find the correct answer guess(rapid guessing); 3. not-gaming ('engaged'). These disengagement values match the gaming categories in Baker et al's (Ryan S. J. d. Baker, Corbett, Roll, & Koedinger, 2008) PSLC Gaming Detector, though without the distinction of harmful/non-harmful gaming. These levels of engagement were used for posterior analyses about gaming and engagement during problem-solving, as impacted by the interventions.

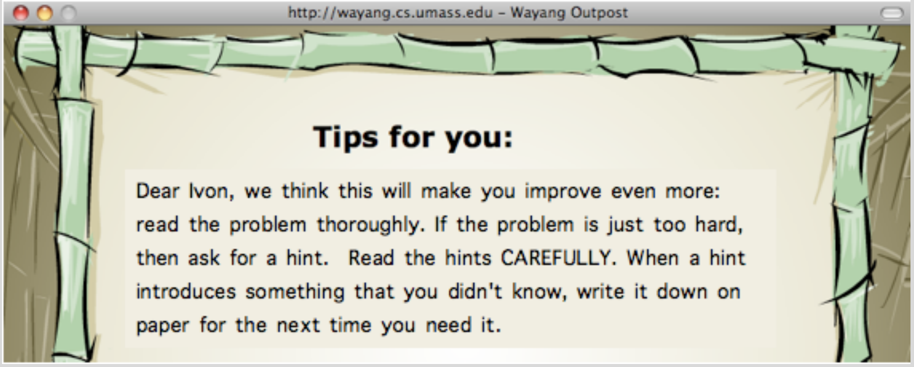
Wayang was enhanced with two kinds of "intervention screens" that appear at fixed intervals of 6 problems (i.e., after clicking the 'next problem' button on the sixth problem). In the experiment presented here, intervention screens were shown to all students, but their contents were driven by the student's behavior within the system. Interventions are of two kinds: either i) a performance graph with an accompanying message, similar to Figure 1 (students received a negative graph or a positive graph depending on their recent performance and their past performance) or ii) a tip (a message) that encouraged some productive learning behavior. The tutoring software provided two kinds of tips in-between problems: Tip-read-carefully and Tip-make-guess. Tip-read-carefully encouraged students to slow down, read the problem and hints carefully, shown in Figure 2 ("Dear Ivon, We think this will make you improve even more: Read the problem thoroughly. If the problem is just too hard, then ask for a hint. Read the hints CAREFULLY. When a hint introduces something that you didn't know, write it down on paper for the next time you need it"). Tip-make-guess encouraged the student to think about the problem, make a guess and, if the guess was wrong, ask for hints ("Dear Ivon, Think the problem thoroughly and make a guess. If your guess is wrong, no problem, just ask for a hint. If you need more hints, keep clicking on help".) As can be seen in the text, tip-make-guess encouraged the use of guessing as part of a sophisticated meta-cognitive strategy, as a way of guiding guessing students to switch from gaming the system to more appropriate approaches. Students were addressed by their first name both in the messages accompanying the charts and the tips. Whether a student saw a progress chart or a tip, and which one, was a randomly-made decision.



Dear Ivon,
This graph shows that you have not improved much recently... BUT you can change this: if you read the problem CAREFULLY, work it out on paper, and click on the 'help' button to get hints when you feel stuck.

Dear Ivon,
Look how much you are learning! This graph shows your improvement over your prior performance: you are getting more questions right. Keep up the good work.

Figure 1. Progress Charts show students their accuracy of responses from earlier in the session to recently



Tips for you:

Dear Ivon, we think this will make you improve even more: read the problem thoroughly. If the problem is just too hard, then ask for a hint. Read the hints CAREFULLY. When a hint introduces something that you didn't know, write it down on paper for the next time you need it.

Figure 2. Tips in Wayang encourage good learning habits

Progress Tips: Study

Participants

Eighty eight (88) students from four different classes (10th grade students and some 11th graders) from an urban-area school in Massachusetts used Wayang Outpost for one week. Wayang Outpost was used in 4 class periods for about 2.5 hours of total tutoring (the rest of the time was spent doing pre-testing and post-testing). A second control group (called no-tutor control) consisted of matched classes of students who did not use the Wayang software at all, but were of the same grade level, equivalent proficiency level, and taught by the same teachers.

Intervention

When students logged on to the software, they were randomly assigned to either an experimental or a control condition. Students in the experimental condition used Wayang with Interventions (Progress charts and Tips). The control group used the traditional Wayang without Interventions every six problems.

Even though Wayang has an adaptive component to tailor the sequencing of problems depending on students' performance at past problems, for this study, the sequencing of problems in Wayang Outpost was fixed (i.e. the same for all students). This decision was made thinking that, if the interventions had an impact, students' engagement in problems would affect problem-solving behavior, and make the software provide harder problems; this would not only add variation to our data but leave us uncertain whether we were really measuring the impact of Progress Charts and Tips, or the impact of getting harder problems, or both. Thus, in an effort to keep the study as clean as possible to measure only the impact of progress charts and tips, Wayang provided a fixed sequence of problems for both experimental and control conditions. Problems were grouped by mathematics topic, so that problems sharing similar skills were close to each other (facilitating transfer from problem to problem); problems were sequenced from ‘easy’ to ‘hard’ overall difficulty, according to empirical measures of student effort that have been estimated throughout the years.

Measures of Effectiveness

Mathematics performance was evaluated with two tests of 43 items extracted from standardized tests: the SAT (Scholastic Aptitude Test) and MCAS (Massachusetts Comprehensive Assessment System). These tests were provided before and after using Wayang, and were counterbalanced. A second measure of math achievement collected was the MCAS standardized test scores from 10th graders who took this exam days after the experiment finished, including also the scores of a set of students of the same level who did not use Wayang at all, but were part of a parallel class, and had the same teachers.

A post-tutor survey was provided, including measures of a student’s performance/learning orientation, how human-like the tutor was, and a student’s liking of mathematics. Most of these metrics came from instruments used by Baker et al (2008) to study the characteristics of students who game the system. In addition, items that measured self-concept in mathematics (Wigfield and Karpathian, 1991) and a 10-question self-efficacy instrument were also part of the survey. Last, we included questions that measured student perceptions of the software and the help provided by the software. All items were in a 6-Likert-type scale, except for 2 learning vs. performance orientation items (cf. Mueller & Dweck, 1998).

As measured before, another measure of effectiveness within the tutoring session was engagement vs. disengagement (gaming) estimations in each problem, according to our Hidden Markov Model. If the interventions were effective, this would be reflected in students’ gaming behavior and gaming frequency while using the software.

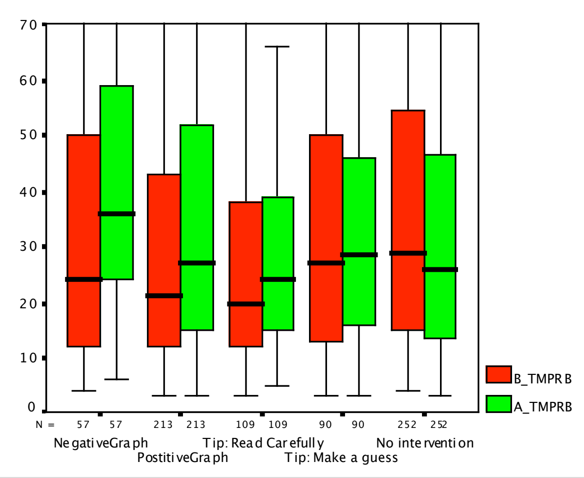
Within Tutor Behavior Results: Gaming and Time-On-Task

At first analysis, there did not appear to be a difference in *overall* frequency of gaming. There was a 17% chance of gaming in a problem, for both the control and experimental groups. After taking a closer look, we realized that guessing was the most frequent kind of gaming behavior, and a significant difference in guessing was found favoring the interventions group (16% for the control group, 12% for the experimental group, independent samples t-test, t=1.97, p=0.05). This means that students in the experimental group slightly changed from one kind of gaming behavior to another, increasing their degree of hint abuse, which add up to 5% in the experimental group instead of 3% for the control group. Progress charts and tips encouraged students to seek more help, but maybe not in the ideal way we may have wanted.

We analyzed whether there was a *local* advantage for the impact of the interventions at reducing gaming. 433 engagement assessments were extracted for specific problems after an intervention was seen, for students in the experimental condition, and 347 for students in the control condition. When analyzing specific forms of gaming (guessing, hint abuse or problem-skipping) we found that students guessed significantly less in problems immediately after the intervention was seen: 12% guessed in the experimental condition vs. 18% guesses in the control condition, a significant difference of 6% more guesses in the control condition, $c^2(1,N=780)= 6.3$, $p=0.01$, which is the most frequent form of gaming in our tutoring system.

Another question explored was whether students’ guessing behavior *changed* after they saw an intervention. Students who guessed in the problem before the intervention (58 cases in the experimental condition, 61 cases in the control condition) had a 12% higher chance to re-engage (not game) in the problem after the intervention than the control group, who did not see the intervention in-between those two problems. Students who abused help in the problem before the intervention (20 in the experimental group, 12 in the control group) had 10% higher chance of being engaged (not gaming) in the following problem after the intervention than the control group. Students who skipped the problem before the intervention (only 7 cases in the experimental condition and 5 cases in the control condition) had a 51% higher chance of being engaged in the following problem after the intervention than after seeing nothing in between those two problems. In general, there is a trend for recovering the "engagement" state after seeing an intervention than students who did not see an intervention. Further analysis indicated that the reduced number of guesses was lowest particularly after the Progress Charts (10% guesses), less than after the Tips alone (14% guesses), in turn less than after seeing no intervention for the control group (18%). A Chi-square test indicated that significantly fewer guesses were observed after seeing a progress chart than after seeing nothing, c^2(1,N=589)= 7.2, p=0.007.

Figure 3.
Seconds spent in problems before and after Progress Charts and Tips



Last, we analyzed how different interventions affected the time spent in the following problem after they were shown. What is the difference between the time spent during the two problems (before and after the charts/tips)? Students who did not see an intervention tended to decrease their time spent in subsequent problems, as shown in figure 3, while students who saw an intervention tended to increase the time spent per problem when there was an intervention in between them. The last two boxes in the Box-Plot of Figure 3 show the median and quartile seconds spent per problem for 252 random pairs of subsequent problems, for students in the tutor-control group, who did not see any progress charts/tips. These two boxes suggest that students generally get more disengaged as the session progresses (median and quartiles are lower in the second problem), by a median 5 seconds less in the following problem. The eight boxes to the right correspond to the seconds spent in 469 problems immediately before and immediately after each intervention, for students in the Intervention group. Clearly, the effect of decreasing time per problem reverses after an intervention: students increase the median time spent in the problem (time on task) after seeing progress charts or tips. A repeated measures ANOVA confirmed that there is a significant difference in time change within (F(721,1)=8.79, p=.003) and between the experimental and tutor-control groups (F(721,1)=7.3, p=0.007). However, note that the shift in time is more pronounced for the graph interventions than for the tip interventions. In fact, for the particular case of Tip-Make-guess, there is not a clear change at all. A paired-samples t-test gave a significant difference for time spent from the problem before to the problem after a Graph Intervention (t(270,1)=-2.9, p=0.004), but not for before and after the tips (t(199,1)=.69, p=.49). Thus, Graph Interventions appear to be the cause of students spending more time in the following problem, while the tips did not change students' time-on-task.

Differences in Math Learning

Table 1 shows the results for pre- and post-test scores for the three groups, i) Intervention Group (used the tutor with interventions every six problems), ii) Tutor Control Group (used the tutor without the interventions) and iii) No-Tutor Control (matched students who did not use the software).

Group	Math Pretest	Math Posttest	MCAS Passing Rate
No Tutor Control			76% (N=38)
Tutor Control	40% (20) (N=40)	40% (28)* (N=40)	79% (N=34)
Tutor Intervention	33% (19) (N=36)	42% (22)* (N=36)	92% (N=24)

Table 1. Means and standard deviations in performance measures before and after tutoring

The raw learning gain (Posttest-Pretest) for the experimental group was 7%, while the Tutor Control group showed no improvement (Table 1 and Figure 2). Across conditions, the learning gains are smaller than what we had observed in previous cohorts of students using Wayang Outpost (15% in about the same amount of time). We think that the fixed sequencing might have affected learning gains in contrast to our standard adaptive sequencing of problems that tailored problems depending on students past performance. Low posttest scores do not prevent us from carrying out a between-subjects comparison. An ANCOVA was used to analyze the difference between the learning gains between the two groups (tutor Intervention and tutor-control). The dependent variable was posttest score (percent correct), with group as an independent variable, and pretest as a covariate. The test of between subjects indicated a significant difference in posttest score between the tutor-control and tutor-Intervention groups (F(1, 76)= 4.23, p=.04), suggesting that there is a significant difference in learning gains favoring the interventions-enhanced group.

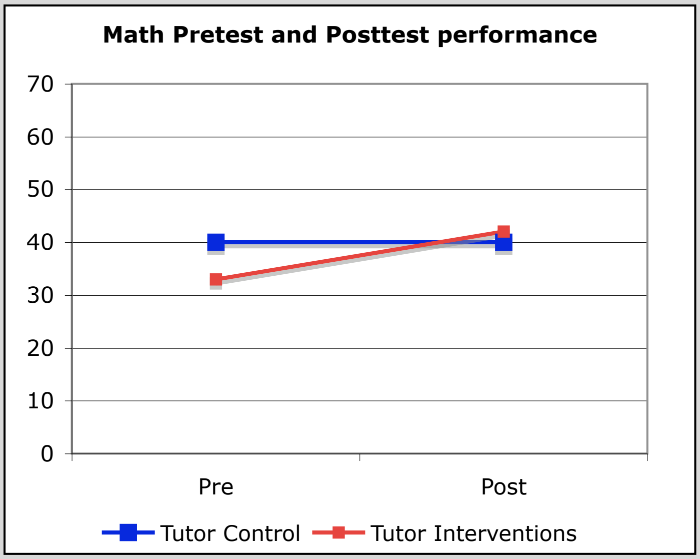


Figure 4. Learning Gains Associated With Receiving Progress Charts and Tips or not receiving them

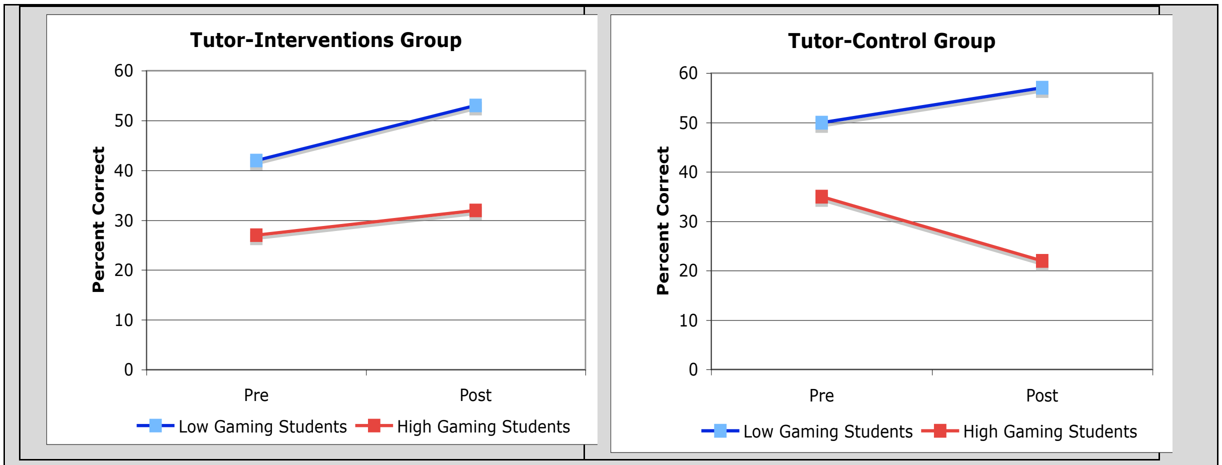


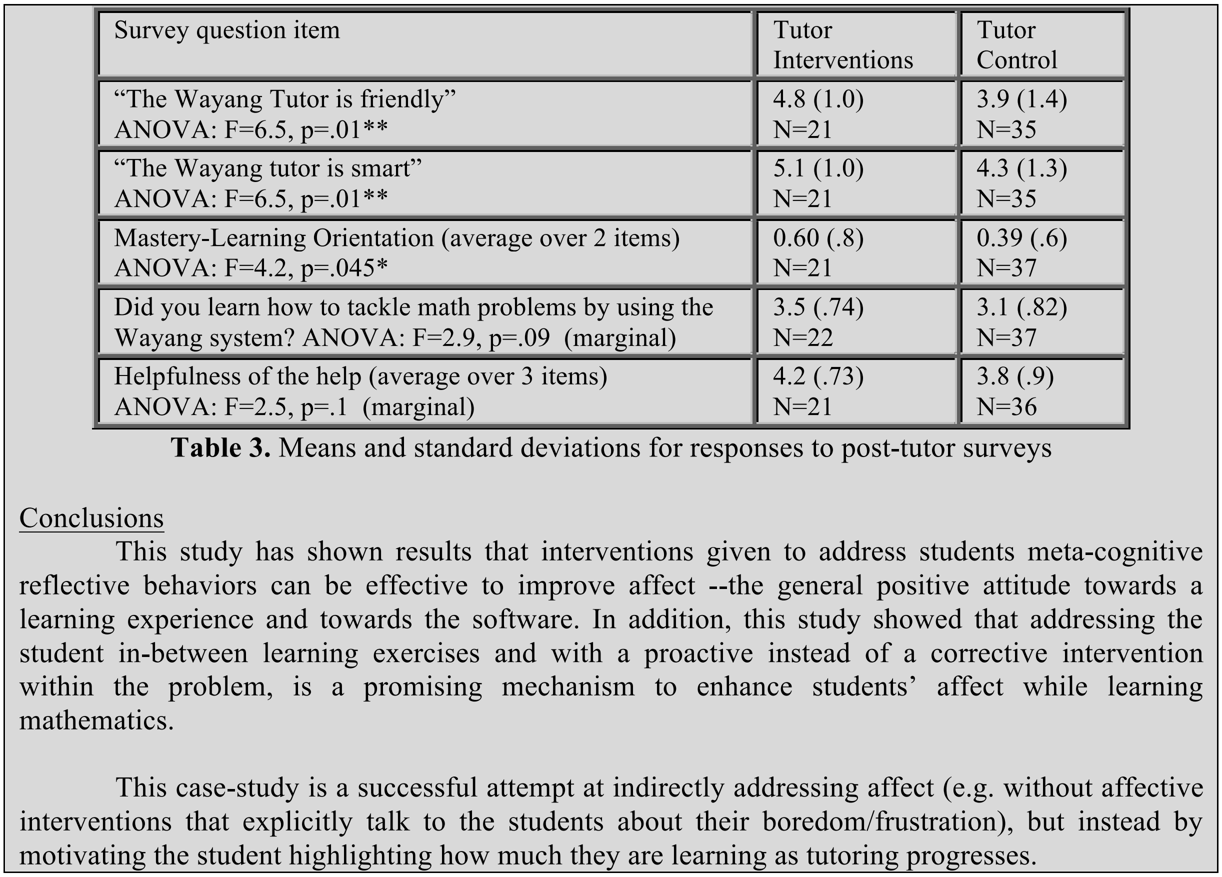
Figure 5. High gamers improve when they receive interventions (left) but not when they don't receive them (right)

The difference in learning gains for high vs. low gamers was also analyzed. Students were split at the median to classify them as low vs. high gamers. An ANCOVA for posttest score, with group and low/high-gamer as independent variable, and pretest as a covariate was analyzed. The result was a marginally significant interaction effect for group x low/high-gamer (F(1,76)=3.4, p=0.07), and a significant higher posttest score for students in the interventions group F(1,76)=3.4, p=0.02. This means that, similarly to Scooter, progress charts/tips interventions especially benefit high gamers: when students who game a lot are shown interventions, their posttest score increases instead of decreasing. Figure 3 shows this interaction effect.

Because the experiment was carried out days before a statewide-standardized test exam (MCAS), we collected standardized scores for students who took the exam, including a matched group of students (same level, same teachers) who did not use the tutor. Students in the Interventions group had a higher average learning gain (7% more than the control group) and higher passing rate at the MCAS exam (92% vs. 79%) than their counterparts in the control group, and higher than the no tutor control group (92% vs. 76%). A Chi-square test indicated that students in the Tutor Intervention group had marginally higher passing rate (c^{2}(1,N=62)= 2.4, p=0.12) than the No Tutor Control group.

Affective Variables Results

Table 2 shows only results of the survey items that showed at least a marginally significant difference. Students in the Interventions group agreed more with the statements such as "the Wayang tutor was smart and friendly". They also had significantly higher mastery-learning orientation scores in the two items that measured performance vs. learning orientation. Marginally significant differences were observed for students thinking they have learned with the Wayang tutor and beliefs about the helpfulness of the help, all favoring the Interventions group. No significant differences were encountered for questions about 'computers caring about myself', 'Wayang is genuinely concerned about my learning', feeling of control over computers, mathematics liking, 'the tutor is concerned about my learning', self-concept about mathematics ability, or self-efficacy. These last items indicate deeper perceptions about oneself than the earlier items., The Interventions do not seem to have impacted these deeper self-perceptions, but did lead to significant effects, in terms of perceptions of the system, its helpfulness, and the students'willingness to learn.



Survey question item	Tutor Interventions	Tutor Control
"The Wayang Tutor is friendly" ANOVA: F=6.5, p=.01**	4.8 (1.0) N=21	3.9 (1.4) N=35
"The Wayang tutor is smart" ANOVA: F=6.5, p=.01**	5.1 (1.0) N=21	4.3 (1.3) N=35
Mastery-Learning Orientation (average over 2 items) ANOVA: F=4.2, p=.045*	0.60 (.8) N=21	0.39 (.6) N=37
Did you learn how to tackle math problems by using the Wayang system? ANOVA: F=2.9, p=.09 (marginal)	3.5 (.74) N=22	3.1 (.82) N=37
Helpfulness of the help (average over 3 items) ANOVA: F=2.5, p=.1 (marginal)	4.2 (.73) N=21	3.8 (.9) N=36

Table 3. Means and standard deviations for responses to post-tutor surveys

Conclusions

This study has shown results that interventions given to address students meta-cognitive reflective behaviors can be effective to improve affect --the general positive attitude towards a learning experience and towards the software. In addition, this study showed that addressing the student in-between learning exercises and with a proactive instead of a corrective intervention within the problem, is a promising mechanism to enhance students' affect while learning mathematics.

This case-study is a successful attempt at indirectly addressing affect (e.g. without affective interventions that explicitly talk to the students about their boredom/frustration), but instead by motivating the student highlighting how much they are learning as tutoring progresses.

Designing to counter negative behaviours associated to boredom

This case-study describes a system that detects a specific learning behaviour – gaming the system – that is indicative of a non-optimal learning strategy that has been shown to have its roots in the affective dimension, being preceded by boredom (Ryan S. J. d. Baker, D'Mello, Rodrigo, & Graesser, 2010).

Case-study: Scooter the Tutor

Ryan Baker
Department of Social Science and Policy Studies, Worcester Polytechnic Institute, Worcester, USA

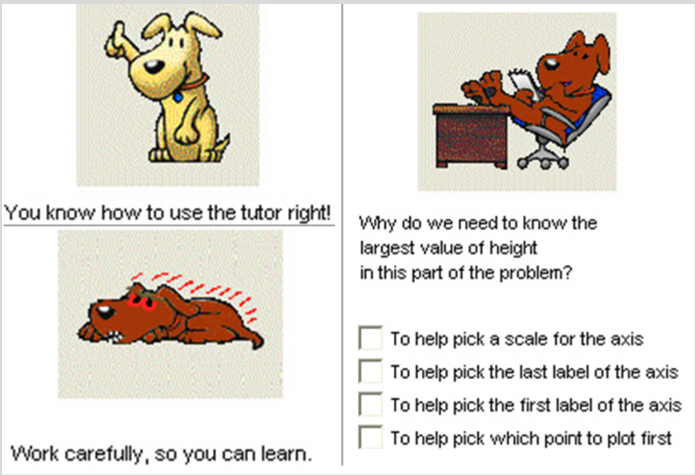
One key possibility for how software might respond to differences in student affect is to respond to the behaviors that emerge from affect, rather than the affect itself. This approach may be easier to deploy in the short term, as there are already several accurate and validated detectors of student behavior, which require no data beyond data of the interactions between the student and the software interface (e.g. Aleven, McLaren, Roll, & Koedinger, 2004; Ryan S.J.d. Baker, 2007; R. S. J. d. Baker, et al., 2008; Beal, Mitra, & Cohen, 2007; Beck, 2005; Walonoski & Heffernan, 2006). Some students (particularly computer-savvy students) may also be more willing to believe that software can accurately detect their behaviors than their emotions, leading students to respond to the software differently. Finally, the link between a student's behavior and the software's response may be easier to understand than the link between affect and a response. (These last two possibilities are, of course, hypotheses rather than proven findings, and may turn out to be wrong).

However, responding directly to the student behaviors which emerge from a specific affect state has the potential to positively impact behaviors and even learning, while not improving the affect that those behaviors emerge from (and in some cases possibly even worsening students' affect). As such, this type of design has some risk to engender long-term negative consequences (in terms of students' desire for future learning within the software environment), even as it has positive effects at other levels.

We offer a case-study showing an example of how this may occur, involving an affective learning companion, Scooter the Tutor (Ryan S. J. d. Baker et al., 2006). Scooter the Tutor is an affective learning companion designed specifically to respond to gaming the system. Gaming the System, as discussed in chapter 3, has been shown to emerge from students' negative affect. In particular, students who experience boredom are significantly more likely to game the system shortly after experiencing those affective states (Ryan S. J. d. Baker, et al., 2010; Ryan S. J. d. Baker, Rodrigo, & Xolocotzin, 2007). Gaming also leads to future boredom. The fact that gaming the system both emerges from boredom and leads to future boredom creates what (S. D'Mello, Taylor, et al., 2007) refer to as a "vicious cycle".

Scooter the Tutor, deployed in a Cognitive Tutor (Anderson, Corbett, & Koedinger, 1995), is an affective learning companion who responds to gaming in two fashions. First, Scooter responds to gaming with expressions of negative emotion. When the student is not gaming, Scooter looks happy and occasionally gives the student positive messages (see the top-left of Figure ALPHA).

Scooter's behavior changes when the student is detected to be gaming harmfully. If the detector assesses that the student has been gaming harmfully, but the student has not yet obtained the answer, Scooter displays increasing levels of displeasure (culminating in the expression shown on the bottom-left of Figure ALPHA), to signal to the student that he or she should now stop gaming, and try to get the answer in a more appropriate fashion. Scooter's displeasure is combined with changes in color to make his emotion visible to a casual glance from a teacher, even at a distance. Second, Scooter gives students supplementary exercises which involve the same skills and concepts as material bypassed by gaming. The supplementary exercises both remind the student of the importance of the gamed step for the overall learning goal, and give the student an alternate opportunity to learn the material. The goal of Scooter's design is to both reduce gaming, and give students an alternate method to learn the knowledge missed by gaming. As such, Scooter does not respond directly to student affect, but instead responds to the behaviors emerging from that affect. Images of Scooter's behavior are shown in Figure 1.



You know how to use the tutor right!

Work carefully, so you can learn.

Why do we need to know the largest value of height in this part of the problem?

□ To help pick a scale for the axis

□ To help pick the last label of the axis

□ To help pick the first label of the axis

□ To help pick which point to plot first

Figure 1. Scooter the Tutor – looking happy when the student has not been gaming harmfully (top-left), giving a supplementary exercise to a gaming student (right), and looking angry when the student is believed to have been gaming heavily, or attempted to game Scooter during a supplementary exercise (bottom-left).

Scooter's behavior is driven by the PSLC Gaming Detector (R. S. J. d. Baker, et al., 2008). The PSLC Gaming Detector distinguishes between different types of gaming the system; Scooter focuses on a type of gaming behavior shown to be associated with significantly worse learning.

Scooter was evaluated in an in-vivo experiment in 5 classrooms in the Pittsburgh suburbs, where 102 students who had already used an intelligent tutor for at least 6 months as part of their regular mathematics curriculum were selected to receive a tutor with Scooter (experimental condition) or a tutor without Scooter (control condition). Students' degree of gaming behavior was evaluated in each condition using live observations (see the previous chapter for more details on this method), and students' learning was assessed using counter-balanced pre and post tests. Students' attitudes towards Scooter and the learning situation were also measured, using questionnaires given at pre and post test.

Scooter was associated with a sizeable, though only marginally significant, reduction in the frequency of observed gaming. 33% of students were seen gaming in the control condition (using quantitative field observations), while 18% of students were seen gaming in the experimental condition. However, although fewer students gamed, those students who did game did not appear to game less (14% in the experimental condition, 17% in the control condition).

In terms of domain learning, there was not an overall between-condition effect. However, only a minority of students received a substantial number of supplemental exercise interventions from Scooter (because only a minority of students gamed the system). There is some evidence that Scooter may have had an effect on these specific students.

Scooter's supplemental exercises appeared to be associated with significantly better domain learning. The third of students that received the most supplementary exercises had significantly better learning than the other two thirds, as shown in Figure 2. In particular, students who received the most supplementary exercises started out behind the rest of the class, but caught up by the post-test (see Figure 3 Left). By contrast, in both the control condition (see Figure 3 Right) and in prior studies with the same tutor, frequent harmful gaming is associated with starting out lower than the rest of the class, and falling further behind by the post-test, rather than catching up.

By contrast to the supplementary exercises, Scooter's emotional expressions were not associated with better or worse learning. Students who received more expressions of anger did not have a larger average learning gain than other students.

As well, there was no evidence that receiving an expression of anger or supplementary exercise reduced future gaming behavior. Hence, the observed reduction in gaming may have been from Scooter's simple presence. Students who chose to game knowing that Scooter was there did not appear to reduce their gaming.

Given the connection between receiving Scooter's exercises and learning, it is surprising that there was not an overall learning effect for Scooter. In particular, it is surprising that the students who chose not to game the system did not have better learning; one possibility is that they replaced gaming with other ineffective learning strategies. One possibility is that the reduction in the number of gaming students paradoxically prevented these students from receiving interventions that would have helped them. Potential evidence for this hypothesis is seen among students who received very few supplementary exercises (the middle bar in Figure 2) – these students had poorer learning than the students who received no supplementary exercises. Perhaps if these students had gamed more, and received more supplementary exercises, their learning would have been better (though this approach clearly would not benefit their long-term acquisition of metacognitive and/or affective regulation skill).

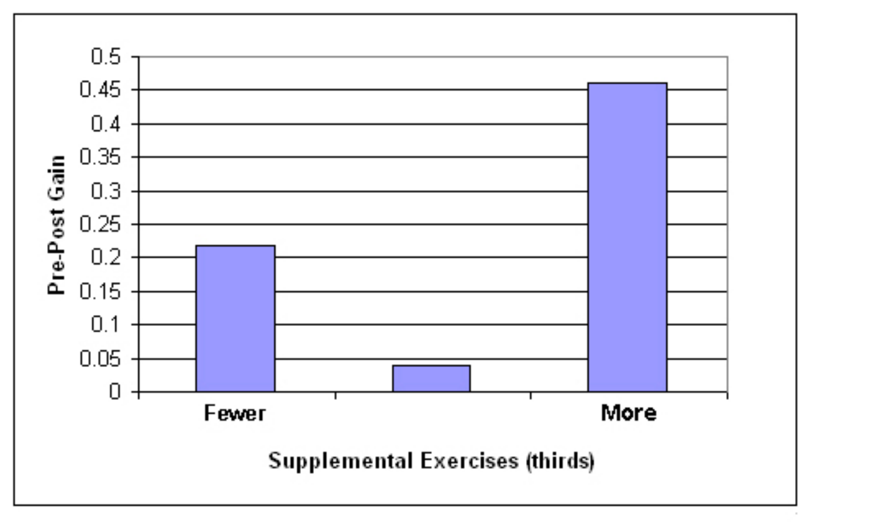


Figure 2. The Learning Gains Associated With Receiving Different Levels of Supplemental Exercises From Scooter

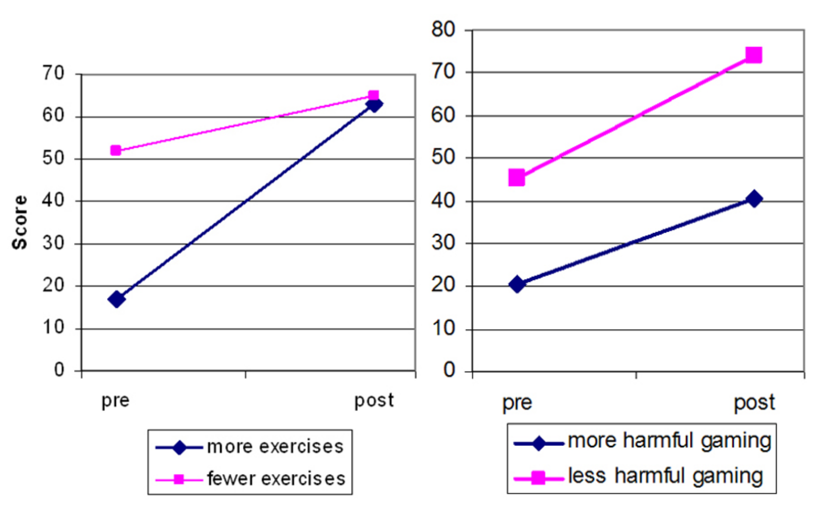


Figure 3. Left: The Learning Gains Associated With Receiving Different Levels of Supplemental Exercises From Scooter (Top Third versus Other Two Thirds).Right: The Learning Gains Associated With Different Levels of Harmful Gaming, in the Control Condition (Top Half of Harmful Gaming Versus Other Students)

However, at the same time that Scooter was effective at reducing gaming, and appeared to increase some students' learning, there was evidence that students who received interventions from Scooter disliked the experience. Table 1 shows students' pre-post responses on questionnaire items relevant to their experience with Scooter (pre-test items are duplicates of post-test items, substituting the words "the tutor" for the word "Scooter"). As can be seen, students generally thought Scooter was less smart than the regular tutor, and students who received more supplementary exercises or expressions of anger also thought that Scooter did not treat people as individuals, was not friendly, and ignored their feelings.

Post-Test Item	More Supp. Ex. (Pre->Post)	More Anger (Pre->Post)	Other Students (Pre->Post)
“Scooter treats people as individuals”	*3.9 -> 2.7*	4.3 -> 3.1	4.3 -> 4.1
“Scooter ignores my feelings”	3.5 -> 3.6	**3.1 -> 4.3**	3.3 -> 3.4
“I feel that Scooter, in his own unique way, is genuinely concerned about my learning.”	3.6 -> 2.9	4.1 -> 3.3	3.9 -> 4.0
“Scooter is friendly”	*3.6 -> 2.3*	*4.7 -> 3.0*	4.4 -> 4.0
“Scooter is smart”	**4.7 -> 2.9**	**5.3 -> 2.9**	**4.9 -> 4.1**
“I would like it if Scooter was a part of my regular tutor”	n/a -> 2.9	n/a -> 3.3	n/a -> 3.6
“Scooter is irritable”	n/a -> 3.9	n/a -> 4.4	n/a -> 4.1
“Scooter wants me to do well in class”	n/a -> 3.9	n/a -> 4.2	n/a -> 4.7

Table 1. Students’ attitudes, at pre and post tests. Statistically significant changes (two-tailed p<0.05) shown in boldface; marginally significant changes shown in italics.

These results illustrate both the benefits and risks of focusing intervention on students’ behavior instead of their affect. Scooter appears to have improved students’ behavior and possibly even some students’ learning. However, students appear to have disliked Scooter considerably. While no direct data on moment-to-moment affect was collected during the Scooter study, it seems reasonable to hypothesize that Scooter may have substituted the boredom, confusion → gaming → boredom pattern reported in (Ryan S. J. d. Baker, et al., 2010; Ryan S. J. d. Baker, et al., 2007) with a boredom, confusion → gaming → Scooter → boredom,anger pattern. Interventions that more directly address students’ affect may potentially be able to improve behavior and learning, as Scooter does, without having negative consequences for affect.

Conclusions

The earlier chapters in this book have set out some of the complexities of the emotional and affective dimensions of learning. These range from epistemological, psychological and neurophysiological issues about the nature of emotion and its relation to cognition, through to more social issues around the perception of self and others in learning situations. While building intelligent learning and teaching environments has a long history going back nearly half a century, taking affect directly into account in such systems has a much shorter pedigree.

This chapter has briefly categorised educational tools systems in terms of the degree of adaptivity that they exhibit. For those that are designed to be either macro-adaptive to affective traits such as attitudes, personality or moods, or micro-adaptive to affective states such as boredom or frustration, an account is given of the possible kinds of response to the affective state of the student. These responses are organised into the emotional, the cognitive, the meta-cognitive and meta-affective and the contextual. Seven case-studies have illustrated the broad range of design considerations that can be taken into account.

While some of the theoretical work on the nature of affect in education is sophisticated and subtle, the demands of system designers for theoretical clarity at a detailed level of granularity in terms of what should be designed and how is as yet unsatisfied. The case-studies indicate that substantial progress is being in terms of input mechanisms to infer the emotional state of the learner(s), as well as processing and reasoning that data to produce pedagogical outcomes. However we are still some way from having the kind of agreed affective design toolkit that is needed, based on reproducibly good educational outcomes across different domains and contexts.

Acknowledgements

We thank the authors of the case-studies for allowing us to include their work and also for their helpful comments on this section.

3. HOW DESIGN INFLUENCES EMOTIONS DURING THE USAGE OF COLLABORATIVE LEARNING TECHNOLOGIES

Ulises Xolocotzin Eligio[1], Shaaron E. Ainsworth[1] and Charles C. Crook[2]
[1]School of Psychology and LSRI, University of Nottingham, UK
[2]School of Education and LSRI, University of Nottingham. UK

Introduction

A study is presented that compares collaborations around the concept-mapping tool *2Connect* and the collaborative educational computer game *Astroversity*. The comparison between learning environments is useful to distinguish how different tasks and design features might influence three affective aspects of the situation: Collaborators' emotions, their understanding of a partner's emotions, and the relationship between collaborators' emotions and their interaction quality.

How concept-mapping tools and collaborative educational games might influence people's emotions

Before describing the specific features of 2Connect and Astroversity, this section describes some of the main features of concept-mapping tools and educational collaborative computer games, in terms of the *tasks* involved in their usage, and the *affective features* of their interaction design.

Tasks and technology during the usage of concept-mapping tools and collaborative educational computer games

Concept mapping tools and collaborative educational computer games intend aim to facilitate learning in different ways. Therefore collaborators around these learning environments perform tasks of different nature, employing different technological resources. This section describes these aspects and their possible implications for collaborators' emotions and their understanding of a partner's emotions.

Tasks and technology in the usage of concept mapping tools

It is thought that concept-mapping tasks are beneficial for learning. A concept map is a graphic structural representation in which nodes (i.e., boxes or circles) represent concepts, and lines that connect nodes represent relationships between concepts. It is though that by making a concept map, people learn by means of 'structuring' their knowledge (Novak & Cañas, 2008; Ruiz-Primo & Shavelson, 1996). Figure 3.presents a concept map that describes concept maps.

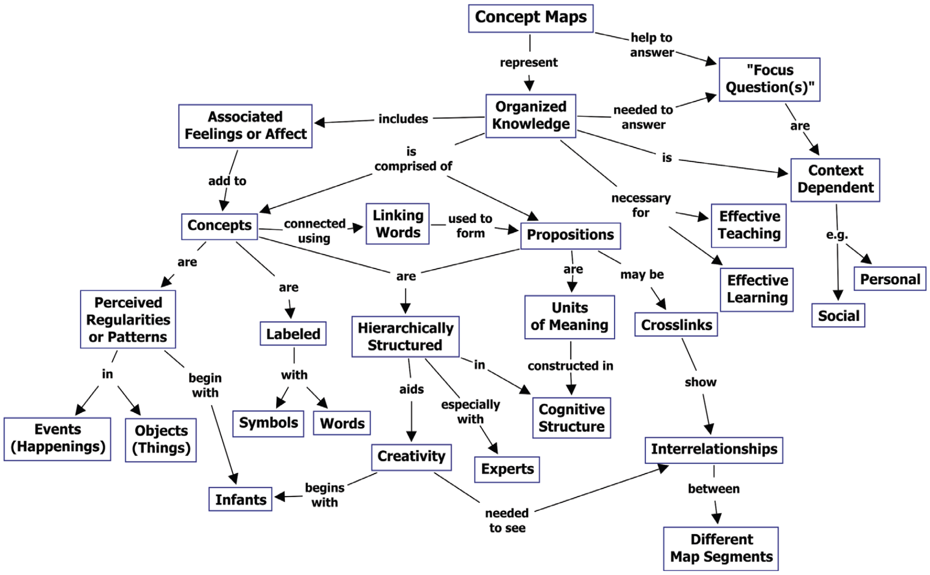


Figure 3. A concept map showing the key features of concept maps (Cañas & Novak, 2006)

Concept-mapping environments are fairly generic, consisting of tools for the user to manipulate text boxes, lines and graphs. Other features might include functionalities to incorporate audio and video, and the possibility to collaborate at-distance. A typical concept-mapping tool can be employed to support learning in different domains (e.g., electricity, genetically modified organism), using with different tasks (e.g., free discussion of a topic).

A collaborative concept-mapping task usually consists of an open-ended discussion. For example, participants may be asked to answer an open question by making a concept map. The final outcome (i.e., the characteristics of the concept map) can be undefined (Novak & Cañas, 2008) or relatively more specific. For example, learners can be asked to fill in a 'skeleton' concept map, or asked not to make more than one relationship per concept (Ruiz-Primo & Shavelson, 1996). In any case, the role of technology is to support the completion of he task, without any direct influence on how the task is set up.

In terms of the interaction between collaborators, learners making a concept-mapping task typically organize their interaction freely. That is, they do not play predefined roles to complete the task. However, some research suggests that structuring the interaction between learners might be beneficial. For example, engagement can be prompted if collaborators are asked to make a concept-map about a controversial topic (e.g., genetically modified organisms) defending opposite positions (Munneke, Andriessen, Kanselaar, & Kirschner, 2007). Moreover, learners' interaction can be structured by means of 'scripts', which are sets of instructions that indicate learners how to interact, for instance, in terms of sequence (e.g., if a learner brings an argument, the script will prompt the partner to state a counter-argument) or role allocation (e.g., after reading information, the script will ask one learner to make a summary, whilst the partner will have to listen) (Kollar, Fischer, & Hesse, 2006). Although scripting is expected to facilitate desirable qualities in collaborators' interaction (e.g., argumentation or mutual learning), it has also been noted that it may reduce the richness and 'fun' of collaboration (Dillenbourg, 2002).

Collaborations around concept mapping tools often imply the usage of a shared interface. In remote interaction, learners might share an interface employing their own terminal each. In co-located interaction, learners usually share one computer and its input device (e.g., a mouse). Recently, there has been increased is an interest to in investigatinge other sorts of interfaces for collaborative concept mapping, such as digital tabletops, that permit learners to simultaneously manipulate the concept-map contents (Do-Lenh, Kaplan, & Dillenbourg, 2009).

Tasks and technology in the usage of collaborative educational computer games

Research about collaborative computer games as educational tools is still rare, especially considering the abundance of studies about educational computer games that focus on individual game play (Kirriemuir & McFarlane, 2004). Briefly, collaborative computer games aim to facilitate learning by supporting the players' coordinated application of skills and knowledge in fixed collaborative tasks.

Clearly, cCollaborative educational computer games are different to concept mapping environments in many ways and in particular in terms of the relationship between technological features and their underlying task. The features of concept-mapping environments are generic across tasks and domains. In contrast, although collaborative educational computer games might be categorized according to genres (e.g., adventure, quest, role playing or first-person shooter), their specific core features (e.g., interface, storyline) are all different because they are designed according to the specific tasks that learners will be asked to perform. In turn, the design of these tasks is directed to promote learning in a specific domain.

Although educational computer games can be all different in terms of its tasks and technological features, they typically present specific goals and often allocate players to complementary roles. Presenting a specific goal to achieve, namely the 'win' state, is key for players of computer games to feel a satisfactory experience (Malone, 1981). Joiner et al. (2006) investigated the effect of having a goal in a ubiquitous game. They found that playing the game with a goal was more interesting than the same game without a goal. Furthermore, the goal version of the game also favoured cooperation between players, even when the game was not explicitly designed to do so.
The allocation of complementary roles is a feature of cooperative computer games that might be beneficial for collaborative learning. Salen & Zimmerman (2004, pp. 253-254) examined *Gaunlet,* a cooperative multiplayer co-located game in which players' task is to find and reach the exit of a series of mazes, employing the complementary skills of their characters (e.g., one character is physically strong whilst the other is a wizard). The complementary nature of the task prompts cooperative work, discussions about fairness in the distribution of resources and a sort of 'useful' conflict between players while discussing strategies to escape the mazes. This is in line with Infante et al. (2009), who reported that the allocation of complementary roles was useful to prompt productive interaction amongst children playing a co-located educational collaborative computer game in the classroom.

Collaborative educational computer games might involve various sorts of tasks, such as going through quests, resolving puzzles, escaping mazes or collecting and analysing data. The tasks, environments, and tools provided, are usually designed with specific learning intentions. For example, in a game intended to support ecology teaching, players go through a number of quests, in which they have to complete different tasks such as acting cooperatively (e.g., approach the same object simultaneously) to eradicate dangerous species or heal infected animals (Susaeta et al., in press). Another example is *Prime Climb;* a game intended to support the learning of factorization in mathematics. Players have to cooperate to climb a series of mountains divided by numbered sections. The task of a player is to jump onto sections that share a factor with the section of her partner. This situation prompts the discussion of players to find out where to jump, in which the topic of factorization is ideally implicit (Dai, Wu, Cohen, & Klawe, 2003; Scott, Mandryk, & Inkpen, 2003).

Collaborative educational computer games can also take advantage of different sorts of interfaces. For example, there have been explorations in the employment of mobile devices in collaborative game-based learning outside the classroom (Facer et al., 2004), the usage of shared interfaces with multiple mice for multiplayer co-located game play (Infante, et al., 2009) and handheld devices (Margolis, Nussbaum, Rodriguez, & Rosas, 2006).

Lastly, it is important to mention that some characteristics of computer games may also be counterproductive for collaboration. For example, the tempo of the game can be so fast that learners don't have time enough to discuss their strategies, or the amount and difficulty of the tasks to perform is so large that learners focus on playing a functional role rather than collaborating (Kiili, 2007).

How the task and technology might influence people's emotions during the usage of concept-mapping tools and collaborative educational games

It is possible to speculate that, during the usage of a concept-mapping tool or an educational collaborative computer game, learners' emotions might be influenced by the nature of the tasks that they are asked to perform.

During a typical collaborative concept-mapping activity, learners' task is to discuss a given topic, following an open-ended goal and interacting 'freely' (i.e., without normative rules such as role allocation). Learners use the concept-mapping tool to make a visual representation of a jointly developed idea. In doing so, they are required to understand one another's perspective. Referring back to Schwartz (1998), the effort to do so should be a major source of motivation (and affect) during collaborative learning. Thus, collaborators' emotions and their understanding of one another's emotions might be influenced by the qualities of their interaction. If their interaction is rich and productive, partners might feel positive affect, and think about the emotions of one another positively. The technology would have a secondary participation in this scenario because its role is limited to support the making of the concept map, without any influence on either the set up of the task or the interaction between learners.

In contrast, collaborative computer games have tasks with specific goals (e.g., resolve a quest, solve puzzles or climb mountains) and allocate players to complementary roles. One could speculate that, since the win state can only be achieved with an effective participation of both collaborators, the partner's actions might be a major influence in the emotions of one another. For example, one collaborator might feel positive emotions if she and her partner are effectively playing their roles. In contrast, negative emotions might result if the partner does not play her role effectively. This context also permits speculations about the way that partners might understand the emotions of one another. For example, they could attribute more positive emotions to the partner who plays her role effectively, and negative emotions to a partner who is not. The technology would have a 'primary' participation in this scenario because the task determines features of the game such as the mechanics and the tools accessible to the players, as well as the dependency on each other.

Affective features in the design of concept-mapping tools and collaborative computer games

Features such as *appearance*, *functionality* and *anthropomorphism* are potential sources of affective reactions during the usage of interactive technologies. This section outlines how these features are implemented in concept-mapping tools and educational computer games. Norman's concept of Emotional Design (Norman, 2004; Norman & Ortony, 2006) explains how the appearance and functionality of interactive technologies influence user's emotions. This concept postulates three levels of 'processing' that provoke affective reactions to the user: *visceral, behavioural,* and *reflective.* The visceral level refers to the perceptual properties of a product, which depends on its appearance. The behavioural level refers to the feelings of control and understanding of a product, which depends on its functionality. The reflective level refers to aspects such as users' personal history and self-image. This level is not covered because is beyond the scope of this chapter. Finally, anthropomorphism refers to people's tendency to treat computers as if they were humans, which is a potential source of affective reactions.

Appearance

The first level in Norman's (2004) model of emotional design is the visceral one. This level refers to how the perceptual properties of a product generate basic and automatic affective reactions in the user. This includes, for example, basic judgements such as good or bad, ugly and pretty. The perceptual properties of a product are mostly concentrated in its appearance. *Expressivity* and *aesthetics* are features that differentiate the appearance of concept mapping tools and educational computer games.

Expressivity refers to the use of graphics and animations that make direct reference to emotion conventions (e.g., emoticons) (Sharp, Rogers, & Preece, 2007). Concept-mapping tools can present emotionally loaded icons. However, expressivity is not a main characteristic of concept-mapping tools. In comparison, it is a common feature in the GUI of educational computer games.
Aesthetics are a common concern in GUI design. Cosmetic properties such as colour and elaborated fonts are elements that can be exploited in order to *embellish* an interface. Although the content included by users in a concept map might have aesthetic properties or the concept map itself might have a nice layout, the GUI of a concept-mapping tool in itself is typically not 'pretty'.

In contrast with concept-mapping tools, appealing aesthetics are a common component in the GUI of computer games. It has been explained that graphics are important for affective aspects such as players' immersion and 'sensory appeal'(Padilla, Gonzalez, Gutierrez, Cabrera, & Paderewski, 2009; Wages, Grünvogel, & Grützmacher, 2004). For example, changes in the light tone of a game's GUI might increase or decrease the quality of players' experience and performance (Knez & Niedenthal, 2008). However, sophisticated graphics might also have negative effects. Animations and eye-catching colours engage players, but this might affect the functionality of the game menus (Johnson & Wiles, 2003). Moreover, Clarke & Duimering (2006) reported that regular players of computer games regard eye candy/good graphics as irrelevant if the game is not enjoyable, suggesting that the enjoyment of the game is relatively independent to the graphics quality. Figure 8 displays examples of expressivity and aesthetics in concept-mapping tools and educational games.

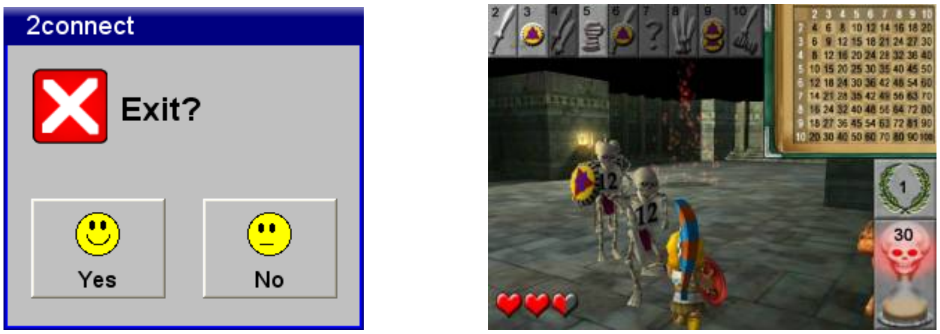


Figure 3, Expressivity and aesthetics in the concept- mapping tool 2Connect and the educational game Zombie Division (Habgood, Ainsworth, & Benford, 2005)

Functionality

The second level in Norman's (2004) concept of emotional design is the behavioural one. At this level, user's emotional reactions towards an interactive product are linked to the functionality. Positive affect emerges when the user feels control and understanding of a product. *Feedback* and *complexity* differentiate the functionality of concept-mapping tools and educational computer games.

Concept-mapping tools offer very basic feedback (e.g., editing or saving functions). In contrast, educational computer games offer feedback about the learner's performance, which is known to have beneficial effects (Rosas et al., 2003). For example, feedback that indicates progress in a game provokes positive emotions (Jones & Issroff, 2007).

The complexity of concept-mapping tools is not a potential source of emotions. This kind of learning technology provides learners with simple to use tools that often resemble more mundane technology such as paper and pencil. In comparison, the interaction design of educational computer games is more complex. The number and complexity of tools and rules is often abundant, which might facilitate engagement and motivation, but can also overburden learners. For example, players of computer games like having multiple levels and a varied story line; but dislike having to make precise manoeuvres in large and slow to learn scenarios (Clarke & Duimering, 2006).

Anthropomorphism

Anthropomorphism refers to the tendency of people to regard computers as persons. That is, attributing to them qualities like motivations, emotions or personality. Anthropomorphism can be the consequence of both the perceptual and functional properties of software. Animated agents or messages in first person can intentionally trigger affective reactions derived from anthropomorpism. In fact, the implementation of human-like features such as animated companions has proven to be beneficial in computer-supported learning. For example, Maldonado et al (2005) found that the implementation of a co-learner that expresses emotions helped e-learners to have a better performance and feel better (i.e., not alone, praised and supported). Similarly, Morishima et al (2004) found that cooperating with an agent that expresses emotions increases positive affect (e.g., cooperativeness, trustworthiness and warmth).

It can be said that concept-mapping tools do not have interface elements nor do they interact with users in a way that can be regarded as 'human-like'. In comparison, the use of animated characters and messages that imitate humans are common features of computer games. Moreover, anthropomorphic features are beneficial for the player experience. For example, AI animated characters that exhibit realistic human-like behaviour such as unpredictability and mistakes are well appreciated (Clarke & Duimering, 2006). Prendinger, Mayer, Mori, & Ishizuka (2003) reported that the implementation of an empathic character decreased the stress of users in a mathematical game; although this had no impact on players' performance.

Comparative summary of concept-mapping tools and collaborative educational computer games

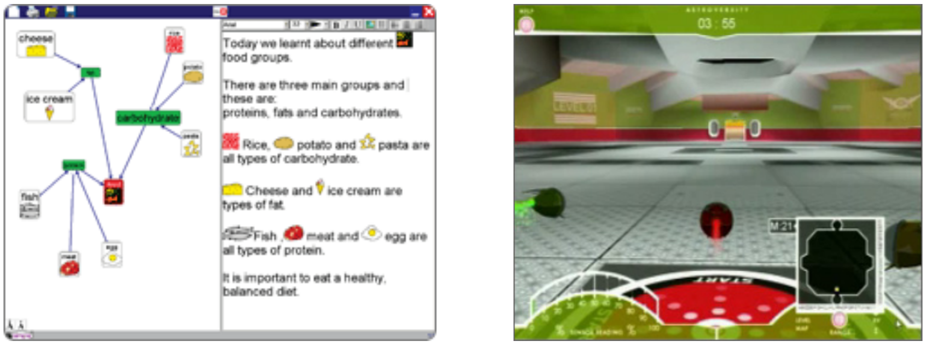
Table 4 compares the typical underlying tasks and affective features of concept-mapping tools and educational collaborative computer games. During collaborations around concept-mapping tools, learners' task is to discuss a topic and elaborate a graphical representation to structure the results of such discussion. The outcome is open ended and learners interact 'freely'. In contrast, collaborative educational computer games may imply a wide range of tasks with specific goals that allocate partners to complementary roles. Clearly, computer games outnumber concept-mapping tools in terms of elements that can directly generate emotional reactions.

Table 4. Underlying tasks and affective features of concept-mapping tools and collaborative educational computer games

	Concept-mapping tools	**Collaborative computer games**
	Underlying tasks	
Intended learning outcomes	Learning by means of argumentation, reflection, and structuring of jointly constructed knowledge	Learning by applying knowledge and/or practicing skills. This recruits cognitive demands such as planning and reasoning.
Specific tasks	The typical task is to discuss a topic and make a graphical representation of the jointly elaborated knowledge (i.e. a concept map).	Task depending on the learning intentions (e.g., to solve quests to learn about ecology, to resolve puzzles to learn mathematics)
Goals	Open ended (e.g. answer an open question)	The specific goal is to achieve the 'win-state' (e.g., finishing a mission)
Interaction with a partner	Learners interact 'freely' (i.e., without pre-defined roles)	Players are allocated to play pre-defined complementary roles
	Affective features	
Appearance	Colourful and aesthetically pleasing interfaces are not common	Aesthetics are important for engagement
	Expressivity can be implemented in icons and menus	Expressivity is a common feature in different aspects of the interface, e.g., avatars and characters
Functionality	The interfaces of concept-mapping tools tend to be fairly simple and most of them work upon the same principles. They can be used in a very intuitive manner.	Educative games are quite diverse in terms of interface and mechanics. Their functions can be complex and often demand effort.
	Concept-mapping tools do not give feedback	In educative games, feedback is common and often used as a resource to promote reflection
Anthropomorphism	Animated characters or first person messages are not a common feature	The use of animated agents or 'persona' is common
	Concept-mapping tools are unlikely to be regarded as with human-like attributes	The characters of the game can provoke emotional reactions

Underlying tasks and affective features in the usage of 2Connect and Astroversity

This section makes an overall description of 2Connect and Astroversity, also explaining how the tasks around these learning environments were set for the presented study. To do that is necessary to mention that participants were organized in pairs. The characteristics of the participants and their organization in the experimental design of the study will be fully described in the Methods section. Figure 4 shows screenshots of Astroversity and 2connect.



Today we learnt about different food groups.

There are three main groups and these are:
proteins, fats and carbohydrates.

Rice, potato and pasta are all types of carbohydrate.

Cheese and ice cream are types of fat.

Fish, meat and egg are all types of protein.

It is important to eat a healthy, balanced diet.

Figure 4. Screenshots of Astroversity and 2Connect

2Connect is a typical concept-mapping tool developed for children. Its interface offers only basic functionalities such as tools to draw nodes and lines with different colours. Other functionalities include a predefined collection of images and the capacity to record and insert audio. In the study, participants were introduced to the functionalities of 2Connect and were asked to make a concept-map to outline the pros and cons of the student life in the University of Nottingham No predefined features for the concept map were requested, and participants were free to include any kind of content other than text (e.g., images or audio). Participants used 2Connect sharing one computer with one mouse as input device, and autonomously (i.e., without direction of the experimenter) determined who was going to be in charge of the computer to control the mouse and type with the keyboard.

Astroversity was designed to support the learning of data gathering and interpretation skills. Astroversity is the name of a spaceship academy invaded by invisible aliens. The first part of the game explains the mission: to find an injured student and trace the safest route for a rescue vehicle to save her. In the second part, a series of animated characters give advice to approach the mission. In the third part, players use a rover to find the injured student and detect invisible aliens using sensors. They plot the areas with alien presence in a paper grid. This is the data collection task. In the fourth part, players use the data plotted in the paper grid to trace the route with less alien presence in a map. Then, the rescue vehicle follows the route traced by the players and rescues the injured student. In the fifth part, players receive feedback on the amount of damage the student received due to alien presence. This is the data interpretation task. Players pass to the next level or not depending on how much damage the student received. There are three levels of increasing difficulty.

In its networked implementation, Astroversity allows up to three players, each one in a different terminal. However, in the presented study, dyads shared one computer using the one player mode. In this configuration, players shared the control over the computer and autonomously adopted the role of controlling the rover or plotting in the paper grid. This configuration conserves the shared goal of Astroversity and the complementary nature of the collaborators' roles. The configuration was used to have dyadic data, easier to handle than data from triads. **Table 5** compares the tasks (as defined in the presented study) and the affective features of 2Connect and Astroversity. Although these learning environments might support collaborative learning effectively, they might influence collaborators' emotions in different ways. Astroversity is more likely to produce intense emotions related to the achievement of goals, and the interaction with various affective features. 2Connect might generate emotions related to the content of the partners' discussion, but there are almost no technological features that can potentially generate emotions.

Table 5. Affective features and collaboration support the design of 2Connect and Astroversiy

	2Connect	**Astroversity**
	Underlying task	
Intended learning outcomes	2connect is a typical concept map tool designed to support leaners as they acquire and structure their knowledge .	Astroversity supports the acquisition of skills for collective data gathering and interpretation.
Setting	Participants shared one computer running 2Connect, with one mouse as input device	Participants shared one compute to play Astroversity, with one mouse as input device. They were also given a paper grid complementary to the game, in which they had to plot some data during the game play.
Specific task	Participants' task was to discuss the pros and cons of their student experiences and make a concept map.	Whilst playing Astroversity, participants performed number of specific tasks such as data search and collection, strategy planarization, and data interpretation.
Goals	Open ended. Participants were not told which aspects of their student experience to discuss, and no predefined characteristics for the concept map were requested.	Participants' goal was to get as far as they could, and were told there were 3 levels of Astroversity.
Interaction with a partner	Partners were not allocated specific roles. However, they organized themselves to control the computer	Partners were not allocated specific roles. However, they organized themselves to either control the computer or to use the paper grid.
	Affective features	
Appearance	2connect was designed for children. Therefore, its interface uses bright colours and emoticons in the menus	Astroversity has a sophisticated 3-D design and presents characters with emotional expressions
Functionality	The interface of 2connect is straightforward. It presents only a few basic commands represented by easy to recognize icons and provides with no feedback.	In Astroversity, players are embodied in a 'rover' controlled with keyboard arrows. Users have to use a paper-pencil plot to help themselves.
Anthropomorphism	In 2 connect, animated characters are implemented in menus, but their participation is limited and they do not interact with users.	Astroversity's interface is populated by characters with anthropomorphic features

Overview of the study

The study answers exploratory research questions, rather than testing specific hypotheses. There are also methodological investigations, such as the selection of emotions that could be more relevant to the CSCL experience and an assessment of whether using both intensity and frequency in self-report of emotions could bring valuable information.

This study focused on co-located collaborations in order to make a fair comparison between learning environments. A comparison of remote vs co-located collaborations around 2Connect and Astroversity would have required a complicated setting to guarantee comparability across conditions; which was beyond the exploratory ambitions of the study.

Research questions

RQ1: How does using a concept-mapping tool and a collaborative educational game influence people's emotions?

Two aspects about the usage of 2Connect and Astroversity are studied to answer this question. First, the effects of using 2Connect and Astroversity on collaborators' emotions are compared. Different CSCL activities might generate different sorts of emotions, which is clear with the comparison between the tasks and technological affective features of computer games and concept-mapping tools. It is expected that using Astroversity would generate more intense emotions than 2Connect, but no specific predictions are made about which emotions could differentiate these collaborative learning environments.

Second, the sources of collaborators' emotions during the usage of 2Connect and Astroversity are explored. This is to distinguish the relevance of three main 'actors' in the scenarios of dyadic CSCL: the activity, the self, and the partner. No specific predictions are made, but it is expected that collaborators will refer differently to these aspects as sources of their emotions whilst using Astroversiy and 2Connect.

RQ2: What do collaborators understand about their partners' emotions while using a concept-mapping tool and a collaborative educational game?

Two aspects are studied to answer this question. First, if collaborators feel similar emotions during the use of 2Connect and Astroversity. Second, whether partners 'think about' the emotions of one another differently while using these learning environments. Two mechanisms are investigated: the collaborators' accuracy at judging the emotions of their partners (affective awareness) and the extent to what they judge their partner's emotions on the basis of their own emotions (affective projection). No specific predictions are made, but differences are expected because the tasks of 2Connect and Astroversity may influence the interpersonal understanding of emotions in different ways.

RQ3: What is the relationship between collaborators' emotions and the qualities of their interaction with a partner?

This question was formulated to address the importance of collaborators' emotions beyond the individual level, looking at how it affects the quality of the interaction with a partner. An assessment was made of the relationship between partners' emotions and their judgements about the quality of their interaction. Additionally some illustrative cases are presented to explore the relationship between emotional similarity and some qualities of collaborators' interaction.

Method

Participants

50 unacquainted native English speakers participated in this study. Their mean age was 21.6 years. Recruitment and organization of participants was balanced to control for gender, although gender analysis was not an objective of the study. 60% of the participants were female and 40% male. They were randomly assigned to dyads in three configurations: female (11), male (6) and mixed (8).

Design

A within-participants experimental design was used to explore whether the differences in the design of 2Connect and Astroversity influenced collaborators' understanding of their own emotions and the emotions of their partners. *Learning Environment* was the experimental factor with two counterbalanced conditions. In the condition '2Connect', participants outlined the pros and cons of their student experience in the University of Nottingham using 2connect for 20 minutes. In the condition 'Astroversity', participants played Astroversity for 20 minutes, with the instruction to reach as far as they could.

Questionnaires

Participants answered four questionnaires during the study: *Prior Emotions, Own Emotions, Partner emotions* and *Interaction Quality.* The first three questionnaires collected data about the participants' emotions and their judgements of their partners' emotions. These questionnaires included a list of 15 emotion words: *happy, angry, sad, fearful, angry, bored, challenged, interested, hopeful, frustrated, contempt, disgusted, surprised, proud, ashamed* and *guilty.* These emotions were selected because they have two properties relevant for this study. First, their underlying dimensions of cognitive appraisal make them clearly different from each other. That is, people makes different associations between these emotions and aspects such as the *pleasantness, certainty* or *control* they perceive in the context (Reisenzein & Hofmann, 1993; Smith & Ellsworth, 1985) Therefore, these words are adequate to describe affect in terms of discrete emotions. Second, the selected words include emotions that have been previously studied in the study of learning technologies e.g., *anger, boredom, frustration, contempt, disgust* and *surprise* (S. K. D'Mello, et al., 2006; A. Graesser, et al., 2006).
The questionnaires were employed as follows:

- *Prior Emotions.* Participants answered this questionnaire to report the emotions they felt during the 20 minutes prior to their arrival for the study.

- *Own Emotions.* Participants answered this questionnaire to report the intensity and frequency of their own emotions during the usage of 2Connect and Astroversity, and the extent to which their emotions had to do with *the activity, themselves* and *their partners.*

- *Partner Emotions.* This questionnaire was identical to the Own Emotions questionnaire, except that participants reported the emotions of their partners. Unlike in the Own

Emotions questionnaire, participants did not report whether the emotions of their partners had to do with *the activity, themselves* and *their partners.*

- *Interaction quality.* This questionnaire presented three questions for the participants to assess the quality of their interaction. The questionnaire presented the assertion *while my partner and I were using 2Connect/Playing Astroversity, we…* followed by the sentences: 1) *understood each others' ideas and opinions,* 2) *were thinking alike* and 3) *were cooperating equally.* Participants used a 4-point Likert scale anchored in *not at all* and *to a great extent,* to rate their agreement or disagreement with these sentences.

Procedure

Participants were recruited by advertisement and mailing list in several schools of the University of Nottingham. Once registered, participants were paired with no knowledge of who would be his/her partner. Each dyad was randomly assigned to one of two orders of counterbalance: 2Connect-Astroversity or Astroversity-2Connect. Once in the experimental room, participants confirmed they did not know each other. Then they were introduced to the study, read the information, signed the consent form and answered the Prior Emotions questionnaire. Also, all participants agreed to be videoed. The recording captured partners' interaction (frontal view).

Then, participants were asked to do their first activity, either play Astroversity or use 2Connect, depending on their assigned counterbalance order. After completion of their first activity, participants answered the Own Emotions questionnaire, the Partner Emotions questionnaire and the Interaction Quality questionnaire. After a 5 minutes break, participants were asked to perform their second activity. After completion, they answered again the Own Emotions questionnaire, the Partner Emotions questionnaire and the Interaction Quality questionnaire. Collaboration partners answered the questionnaires in the same room, sitting back to back so they could not see each other's answers. The sessions lasted approximately 70 minutes and finished with a debriefing.

Data screening

Overall distributions

A histogram screening was carried out to examine the data collected with the Interaction quality questionnaire, the Own Emotions questionnaire and the Partner Emotions questionnaire.

The screening of the Own Emotions questionnaire data focused on the scales of frequency and intensity, excluding the scale where participants reported the extent to what their feelings had to do with their own actions, the actions of their partners and the activity. The histograms of the scales of intensity and frequency are virtually identical (**Error! Reference source not found.**). In both cases, there are floor effects in the reports of the emotions *contempt, sad, guilty, fearful, angry, ashamed* and *disgusted*. Other emotions with less extreme skew were *hopeful, frustrated, bored, surprised* and *proud.* Emotions with no skew are *happy, interested* and *challenged.*
The histograms of the frequency and intensity scales of the Partner Emotions questionnaire were virtually identical to the histograms in the Own Emotions questionnaire.

There are at least three explanations for the floor effects and extreme skews in the Own Emotions questionnaire and the Partner Emotions questionnaire. One is that the questionnaire was not sensitive enough to capture a subtle feeling of these emotions. A second explanation is that participants were reluctant to report these emotions. This would be consistent with other studies showing that 'negative' emotions, e.g., anger, sadness, fear, guilt and shame, are reported with less frequency and intensity than other, more 'positive' emotions (Carstensten, Pasupathi, Mayr, & Nesselroade, 2000; Nezlek, Vansteelandt, Van Mechelen, & Kupens, 2008; Tong, Bishop, Enkelmann, Why, & Diong, 2007). A third explanation is that some of the emotions represented by these words are simply not part of the average CSCL experience (e.g. disgust), or occur with low frequency (e.g., surprise).

Consequently, further analysis focused on the emotions *happy, interested, challenged, hopeful, frustrated* and *bored.* The rationale for this selection was to have two positive emotions (happy and challenged), two emotions that imply both positive and negative attributes (challenged and hopeful) and two negative emotions (frustration and bored). Although bored showed a noticeably skew, it was included to maintain the balance. Other emotions with similar distributions as bored, such as *surprised* and *proud,* were not included because they were positive.

The histogram of the interaction quality questionnaire shows no extreme skews. Therefore, all the data collected with this questionnaire was retained for analysis.

Further analysis located missing values and outliers. In the Own Emotions questionnaire, 4 data points were missing and 6 participants gave outlier scores in at least three emotions, and were excluded or substituted in further analyses.

Correlations between frequency and intensity

The questionnaires Prior Emotions, Own Emotions and Partner Emotions, included scales of frequency and intensity. If the data showed a clear differentiation between intensity and frequency, this would indicate that participants could distinguish these 'dimensions' when reporting their emotions.

Correlation and difference were the criteria to determine whether frequency and intensity were distinguishable. Table 6 shows that the correlations between the scores of frequency and intensity in the Own Emotions questionnaire and the Partner Emotions questionnaire, in relation to the use of 2Connect and the use of Astroversity, are all positive and strong (*r* > .70). This suggests that participants increased or decreased their scores of intensity and frequency at the same time almost all the time.

Table 6. Correlations between frequency and intensity in the Own Emotions questionnaire and the Partner Emotions questionnaire, during the use of 2Connect and Astroversity

	Own Emotions questionnaire		Partner Emotions questionnaire	
	2Connect (*n*=44*)*	Astroversity (*n*=44*)*	2Connect (*n*=44*)*	Astroversity (*n*=44*)*
Happy	.88**	.79**	.84**	.90**
Interested	.72*	.88**	.70**	.76**
Challenged	.92**	.77*	.77**	.73**
Hopeful	.96**	.87**	.65**	.92**
Frustration	.83**	.79**	.92**	.88**
Boredom	.96**	.89**	.94**	.95**

**p* <0.05, ***p* < 0.01

Figure 5 shows that the differences between the frequency scores and the intensity scores in the Own Emotions questionnaire and the Partner Emotions questionnaire, during the use of 2Connect and Astroversity. Overall, are all very small (< 1.0).

Table 7. Means and SD of the average differences between frequency and intensity scores in the Own Emotions questionnaire and the Partner Emotions questionnaire

	Own Emotions questionnaire		Partner Emotions questionnaire	
	M (*n*=44*)*	SD (*n*=44*)*	M (*n*=44*)*	SD (*n*=44*)*
2Connect	.39	.28	.41	.35
Astroversity	.53	.23	.44	.24

The strong and positive correlations between frequency and intensity and the small differences between their means indicate that these attributes were not differentiated by participants. Therefore, the analysis of frequency and intensity would be redundant. Consequently, the analysis focuses upon the scales of intensity since this is more commonly used than frequency in emotion research (see for example, Feldman, 2004; Gray & Watson, 2007; Smith & Ellsworth, 1985; Tong, et al., 2007).

Results

The results are sorted in three sections, one for each research question.

Effects of using 2Connect and Astroversity on collaborators' emotions

To answer RQ1: *How does using a concept-mapping tool and a collaborative educational game influence people's emotions?,* Analyses were made to test the emotional intensity of participants in 2Connect and Astroversity, as well as the differences in what participants attributed as sources of their emotions.

Effects on emotional intensity

The analysis consisted of a *doubly* MANCOVA including the learning environment as within-participants factor with two levels (2Connect/Astroversity), counterbalance order as between participants factor with two levels (2Connect-Astroversity/Astroversity-2Connect) and the prior emotion scores as covariates.

Table 8 shows the Mean and SD of the participants' intensity scores for the emotions *happy, intensity, challenged, hopeful, frustrated* and *bored*. There were no effects of the participants' emotions previous to the study [F (18, 108) = .42, *ns*) or the counterbalance order [F (6, 34)= 1.53, *ns*]. The main effects of learning environment indicated that participants felt more intense emotions whilst using Astroversity than whilst using 2Connect [F (6, 34)= 12.93, *p*< .001, h_p^2= .69). This effect was not general to all the analysed emotions. Participants felt equally, happy, interested and bored across learning environments, but using Astroversity made them feel more challenged [F (1, 39) = 41.29, *MSE*=.65, *p*< .001, h_p^2=.51), hopeful (F [1, 39] = 33.60, *MSE*=.55 *p*< .001, h_p^2=.46) and frustrated (F [1, 39] = 41.73, *MSE*= .98, *p*< .001, h_p^2=.52). The effects are illustrated in Figure 5.

Table 8. Means and SD of the intensity scores of the emotions happy, interested, challenged, hopeful, frustrated and bored

	2Connect (*n=44)*		Astroversity (*n=44)*	
Emotion	*M*	*SD*	*M*	*SD*
Happy	1.61	1.06	1.48	1.00
Interested	1.74	1.08	1.77	1.05
Challenged	0.86	1.03	1.98	1.05
Hopeful	0.25	0.58	1.18	1.08
Frustrated	0.55	0.82	1.93	1.19
Bored	0.68	1.03	0.70	1.05

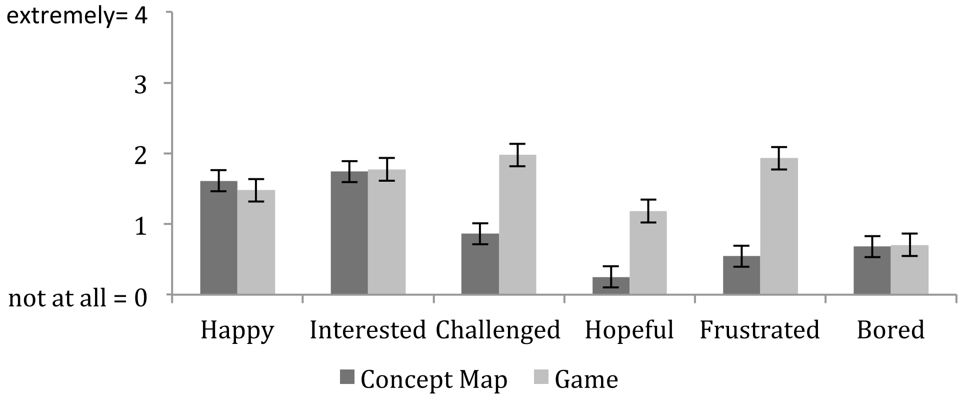


Figure 5. Mean intensity scores of the emotions, happy, interested, challenged, hopeful, frustrated and bored during the usage of 2Connect and Astroversity

Sources of emotions

This analysis investigated how participants referred to their partners, themselves or the activity as sources of their challenge, hope and frustration. The analysis focused on these emotions because they were reported more intensely inGrelation to Astroversity than in relation to 2connect. This helped to reduce data cGomplexity and aCMlso to highlight the differences between learning G G environments. Table 9shows the means and SD of participants' scores on the scalesCM of CM attribution to the *partner, self* and *activity* as sources of their challenge, hope and frustration.

Table 9. Means and SD of the participants' attribution scores for the partner, self and the activity as sources of challenge, hope and frustration, in 2Connect and Astroversity

		2Connect		Astroversity	
Emotion	Source	*M*	*SD*	*M*	*SD*
Challenged (*n*=44)	Partner	.82	.99	1.50	.95
	Self	.93	1.20	1.53	.89
	Activity	1.36	1.63	3.14	1.25
Hopeful (*n*=44)	Partner	.39	.87	1.07	1.04
	Self	.48	1.06	1.16	1.07
	Activity	.48	1.17	1.89	1.61
Frustrated (*n*=44)	Partner	.57	.87	1.32	.98
	Self	.63	.96	1.48	1.13
	Activity	1.02	1.48	2.89	1.52

The attribution scores were analyzed with a 4-way ANOVA including order as between-participants factor with two levels (2Connect-Astroversity/Astroversity-2Connect) and three within-participants factors: source (Partner/Self/Activity), emotion (Challenge/Hope/Frustration) and learning environment (2Connect/Astroversity). Table 10 shows the results, which indicate no significant main effects of order and significant main effects for the three within-participants factors. Further analyses consisted of Bonferroni pairwaised comparisons.

In relation to specific emotions, the sources of challenge and frustration were rated higher than the sources of hope (respectively: *Mean Dif*= .64 *p*< .001, *Mean Dif*= .41 *p*< .001), whilst the sources of challenge and frustration were rated equally (*Mean Dif*= .21, *ns*). In relation to the learning environment, the three emotion sources were rated higher for the usage of Astroversity than for the usage of 2Connect (*Mean Dif*= 1.04 *p*< .001). Lastly, in relation to the source, participants gave higher rates to the activity as an emotion source than to the partner (*Mean Dif*= .76, *p*< .001) or themselves (*Mean Dif*= .85, *p*< .001). The difference between the partner and the self as emotion sources was significant but relatively small (*Mean Dif*= .09 *p*< .05).

Table 10. Effects table for the 2 (Order) x 2 (Source) x 3 (Emotion) x 2 (Learning Environment) mixed ANOVA over the attribution scores

Effect	$df_{(effect,\ error)}$	*MSE*	*F*	h_p^2
	Between-participants main effects			
Order	1,42	7.46	1.69	.03
	Within-participants main effects			
Source (S)	2, 84	.73	75.32***	.64
Emotion (E)	2,84	2.22	12.25***	.22
Learning Environment (L)	1,42	3.04	70.09***	.62
		Interactions[a]		
SxL	2,84	.57	35.45***	.46
SxE	4,168	.30	11.67***	.22
ExL	2,84	2.50	.33	.00
SxExL	4,168	.27	1.77	.03

[a]Interactions with Order were all not significant. For readablity, they are not included in the table

The significant interactions Source x Learning Environment and Source x Emotion, were interpreted with graphs and simple effects analysis. As for the Source x Learning Environment interaction, simple effects of source within each learning environment indicated that the differences between partner, self and the activity as sources of emotions (i.e., attribution scores for each source averaged across emotions) were smaller in 2Connect [$F(2, 86)= 13.70$, $MSE= 1.58$, $p<.001, h_p^2= .24$)] than in Astroversity [$F(2, 86)= 75.27$, $MSE= 24.65$, $p<.001, h_p^2= .63$)]. Hence Figure 6shows that the activity was more emphatically referred as an emotion source during Astroversity than during 2Connect.

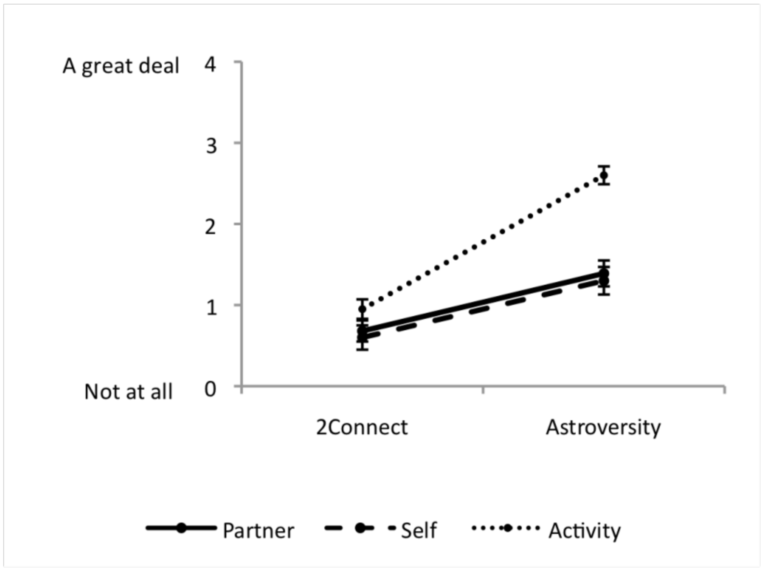


Figure 6. Attribution scores for the partner, self and the activity as sources of emotions as a function of learning environment**.**

As for the Source x Emotion interaction, simple effects of source within emotions (i.e., attribution scores averaged across learning environments) indicated significant but relatively small differences between the participants' attributions to the partner, themselves and the activity as sources of hope [$F(2,86)=19.47$, $MSE=.13$, $p<.001, h_p^2= .31$]. In contrast, participants made clearer differentiations between the sources of challenge [$F(2.86)=60.19$, $MSE=.27$, $p<.001, h_p^2= .58$] and frustration [$F(2.86)=49.18$, $MSE=.27$, $p<.001, h_p^2= .53$]. Figure 7 shows that participants gave a relatively equal importance to the partner, themselves and the activity as sources of hope, whereas they gave clearly more importance to the activity than to the partner or themselves as a source of challenge and frustration.

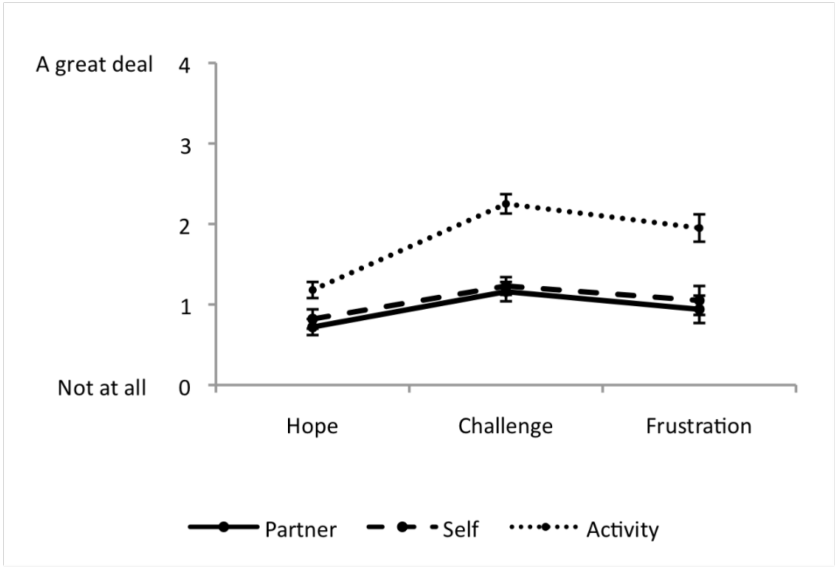


Figure 7. Attribution scores for the partner, self and activity as a function of emotion.

These results indicate that participants identified the sources of their emotions differently depending on the learning environment and specific emotion. Participants made little differentiation between the partner, themselves or the activity as sources of their hope, challenge and frustration whilst using 2Connect. In contrast, the activity was more strongly attributed as an emotion source during the usage of Astroversity, more in relation to challenge and frustration than in relation to hope.

Assessment of emotion understanding

To answer RQ2: *What do collaborators understand about their partners' emotions while using a concept-mapping tool and a collaborative educational game?,* two analysis were made that tested the emotional similarity between partners while using Astroversity and 2Connect, as well as the accuracy at judging the emotions of a partner.

Similarity between partners' emotions

First, the overall similarity between partners' emotions was tested against a by-chance baseline. Secondly, the affective similarity between partners' using 2Connect and Astroversity was assessed separately.

The difference between partners' emotions was measured with a variable labelled *affective similarity index.* In the first analysis, the affective similarity index was defined as the correlation between the dyad members' answers to the Own Emotions questionnaire (considering only the emotions happy, interested, challenged, hopeful, frustrated and bored) across learning environments (n=12 for each dyad, made of 6 emotions x 2 Learning Environments). This index was contrasted with the affective similarity index of a *nominal dyad.* That is, a dyad made out of persons in the same learning environment (using 2Connect or Astroversity) but paired randomly post-hoc. The SD's were as large as the Mean in the affective similarity index of the actual dyads (*M*=.33, *SD=.34)* and actually larger than the mean in the nominal dyads (*M*=.22*, SD*=.34). Therefore, a *Mann-Withney* test was used to compare between these indexes. The results indicated no significant differences between the actual dyads and the nominal dyads (*Z*=-.81, *p=.41*)1.

The second analysis tested the affective similarity on each learning environment. Affective similarity indexes in relation to the use of 2Connect and Astroversity were calculated for each dyad. These indexes were defined as the correlation between the dyad members' answers to the Own Emotions questionnaire (n=6 for each dyad, made of 6 emotions).

[1] It is known that the sampling distribution of the Pearson correlation (*r*) is not normal. Therefore, transformation with Fisher's formula is recommended when using correlation coefficients as data Kenny, D., Kashy, D., & Cook, W. (2006). *Dyadic data analysis*. New York - London: Guilford Press.. However, results obtained with untransformed correlation coefficients are presented. These results do not differ significantly from those obtained with transformed data, and are preferred because its interpretation is more straightforward.

Table 11 shows the *Means, Medians, SD'*s and *Mean ranks* of the affective similarity indexes of actual dyads and nominal dyads. As in the first analysis, the SD's are larger than the means and therefore, the data was analyzed with Mann-Withney tests. There were no differences between the affective similarity indexes of the actual dyads and the affective similarity indexes of nominal dyads, neither in 2Connect, (*Z*=-.98, *p*=.32) or Astroversity (*Z*=-.91, *p*=.36).

Table 11. Means, SD and Mean rank of the affective similarity indexes of actual dyads and nominal dyads in relation to the use of 2Connect and Astroversity

	Actual (n=25)				Nominal (n=25)			
	M	*Med*	*SD*	*Mean Rank*	*M*	*Med*	*SD*	*Mean Rank*
2Connect	.30	.37	.44	27.53	.13	.21	.55	23.48
Astroversity	.28	.31	.49	27.38	.08	.20	.56	23.62

These results indicate that in general, the emotions reported by participants were not more similar to the emotions reported by their partners than would be expected by chance. However, these results are inconclusive given the large variation in the data. Figure 8 shows that actual dyads had a tendency to have higher affective similarity indexes than the nominal dyads, but the large dispersion of the data (i.e., large standard errors) diminished the effect of this tendency, making it not significant.

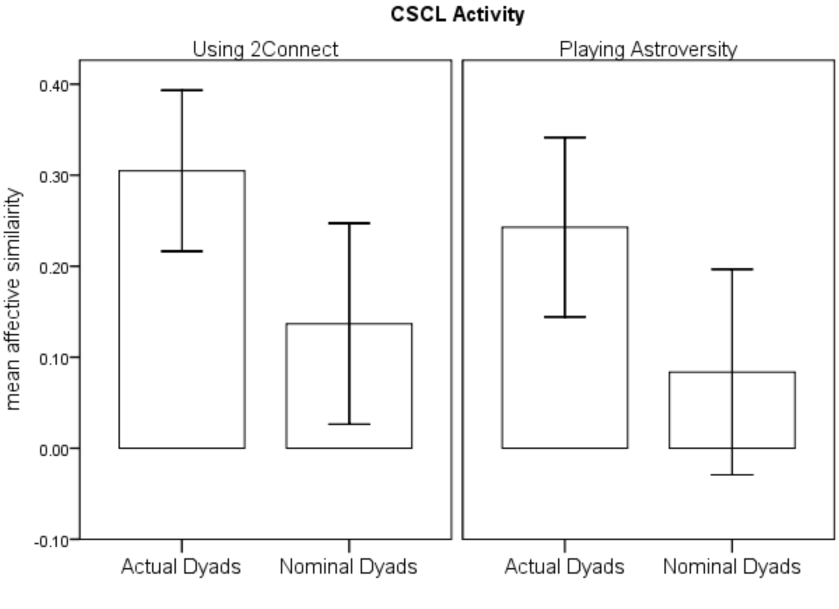


Figure 8. Means and standard errors of the affective similarity index of actual dyads and nominal dyads in the use of 2Connect and Astroversity

Judgements of a partner's emotions

An assessment was made of collaborators' affective projection and affective awareness. The affective awareness referred to the accuracy of collaborators at judging the emotions of their partners. The affective projection referred to the extent to which the collaborators judged their partners' emotions on the basis of their own emotions. The data were the participants' answers to the Own Emotions questionnaire and the Partner Emotions questionnaire. Table 12 shows the Means and SD's of these questionnaires in relation to the use of 2Connect and to the use of Astroversity.

Table 12. Means and SDs of the participants' answers to the Own Emotions Questionnaire and the Partner Emotions questionnaire in 2Connect and Astroversity

	2Connect				Astroversity			
	Own Emotions (n=50)		Partner emotions (n=50)		Own Emotions (n=50)		Partner emotions (n=50)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Happy	1.72	1.1	1.46	.95	1.52	1.01	1.52	.99
Interested	1.86	1.08	1.90	.84	1.82	1.06	1.98	.96
Challenged	1.04	1.08	.92	1.07	2.04	1.04	1.98	.915
Hopeful	.28	.67	.36	.75	1.24	1.08	1.18	1.14
Frustrated	.58	.86	.60	.92	1.88	1.20	1.34	1.20
Bored	1.86	1.08	.80	1.03	1.82	1.06	.82	1.00

The data was arranged in a dataset of n= 50. To analyse all the dyads, the outliers found in the data screening were substituted with the overall mean in each learning environment. There were columns for the participants' answers to the Partner Emotions questionnaire and the Own Emotions questionnaire, as well as columns for their partners' answers to the Own Emotions questionnaire.

Within each learning environment, there were separated analyses for each emotion: happy, interested, challenged, hopeful, frustrated and bored. These analyses consisted of multiple regressions including the following variables.

- Dependent variable: The participants' answers to the Partner Emotions questionnaire (what they thought about their partner's emotions)

- Independent variables:

    - The participants' answers to the Own Emotions questionnaire (what they were feeling). The effect of this variable would be the 'indicator' of the participants' affective projection. If significant, this would suggest that participants based their judgements of their partner's emotions on the basis of their own emotions.

    - The answers to the Own Emotions questionnaire of the partner (what the partner was 'actually' feeling). The effect of this variable would be the indicator of the participants' affective awareness. If significant, this would mean that participants made an accurate judgement of their partners' emotions, suggesting that during

> the use of 2Connect and Astroversity, they paid attention to aspects such as the emotional expressions of the partner, or that they accurately identified the events that affected their partner's emotions.

Appendix A presents the extensive summaries of the analyses. To ease the interpretation, Table 13 includes only the Beta (*B*) coefficient of each independent variable. There were 24 tests and, therefore, a risk of type I errors. A conservative approach would be to make Bonferroni corrections to lower the alpha value employed as criteria to determine significance. However, the n (50) is not vast and the application of Bonferroni corrections would lower the alpha value to a rather harsh level (p<.002), which increases the risk of type II errors. Therefore, an alpha value of p<.01 was used as criterion for significance.

Overall, the participants' affective awareness was very low. In both learning environments and in most emotions, the participants' reports of their partners' emotions (their answers to the Partner's Emotions questionnaire) were not associated with the emotions actually reported by their partners (the partner's answers to the Own Emotions questionnaire), hence the non significant *B* values.
In contrast, participants showed a strong affective projection. They had a strong tendency to judge their partners' emotions on the basis of their own emotions, especially during the use of 2Connect. In this activity, the *B* values of the participants' scores in the Own Emotions questionnaire are significant for all emotions except frustration. Although with less strength, the same tendency is observable in relation to the use of Astroversity. For all emotions, the participants' own emotions were marginally or significantly associated with their judgements of their partners' emotions.

Table 13 Beta values of the participants' own emotion scores and their partners' own emotion scores as predictors of their judgements about each other's emotions

	2Connect		Astroversity	
	Own emotions	Partner's own emotions	Own emotions	Partner's own emotions
Emotion	*B*	*B*	*B*	*B*
Happy	0.52**	0.17	0.36*	0.25
Interested	0.40**	0.04	0.30^1	0.32*
Challenged	0.52**	0.003	0.24^1	0.16
Hop eful	0.56**	-0.13	0.58**	0.18
Frustrated	0.33^1	-0.07	0.52**	-0.02
Bored	0.58**	0.20	0.30^1	0.12

^1p<.05, *p<.01, **p<.005

The relationship between collaborators' emotions and their interaction quality

To answer RQ3: *What is the relationship between collaborators' emotions and the qualities of their interaction with a partner?* One analysis was made that assessed the correlations between joint measures of partners' emotions and their perceived interaction quality. Additionally, the examples of some dyads are presented to illustrate different ways in which the relationship between emotions and interaction quality could occur might look like in some cases.

The relationship between collaborators' emotions and their perceived interaction quality

Dyad level measures of partners' emotions and their perceived interaction quality were composed. Then, the correlations between these measures were calculated in relation to the use of 2Connect and Astroversity. The analysis was made with a dyad-wise dataset (n=25). There were 25 rows, one for each dyad, and columns for each dyad member (e.g., emotion intensity of dyad member *a,* and emotion intensity of dyad member *b*). The measurement of emotion intensity at the dyad level was defined as the sum of the dyad members' answers to the Own Emotions questionnaire. Thus, there was a dyad level measure for each emotion happy, challenged, hopeful and frustrated. The use of sum to make dyad-level variables is common in CSCL studies (e.g., Do-Lenh, et al., 2009).

As for the measurement of interaction quality, a *perceived interaction quality* variable was defined as the averaged items in the interaction quality questionnaire. The reliability of this questionnaire was marginally acceptable in the concept map (*Cronbach's alpha*= .61) and good in the game (*Cronbach's alpha*= .84). The perceived interaction quality of dyad members was summed to make a dyad-level measure.

Table 14 shows the correlations between the dyad level measures of perceived interaction quality and the dyad level measures of happiness, interest, challenge, hope, frustration and boredom. Regardless of the learning environment, those who rated their collaboration quality higher also enjoyed more (i.e., reported more happiness and less boredom). Only difference between learning environments was that only during Astroversity, those who felt more frustrated also made lower ratings of their interaction quality. The rest of the emotions did not significantly correlate with the perceived interaction quality.

Table 14. Correlation coefficients of the dyad level measures of perceived interaction quality and emotions

	Dyad level emotion intensity (n=25)					
Dyad level PIQ	Hap	Int	Cha	Hop	Fru	Bor
2Connect	.40*	.28	-.02	-.12	-.33	-.46*
Astroversity	.47*	.27	-.14	.14	-.40*	-.49*

**p*<.05

Illustrative examples of the relationship between emotions and interaction quality

The results presented above suggest that people's emotions are associated with their interaction quality. When partners reported more happiness and less boredom they also rated their interactions more positively, regardless of the learning environment. Only in Astroversity, partners reported more frustration when they rated their interaction negatively. This complements other results, which indicate that individuals reported positive emotions (e.g., interest, happiness) in 2Connect and Astroversity, but they reported more challenge, hope and frustration in Astroversity (Section 0), and that some collaborative partners reported more similar emotions than others (Section 0).

The video data of partners' interaction was employed to illustrate these results. Some examples are presented of dyads in which the partners reported more similar or more different emotions whilst interacting around 2Connect and Astroversity. The examples show how the emotions of some collaborators might have been associated with their interaction with the technology (e.g., reactions to the interface, execution of a collaborative task) and a partner (e.g., responsiveness). The aim is to aid the interpretation of the quantitative results with some examples. However, although the video data showed the partners' interaction, not recordings of the computer screen were taken and therefore, observations of their performance in the learning environments could not me made. Thus, the features described in the example dyads were selected as mere illustrations instead of systematically determined with inferential analysis and, therefore, generalisation to other dyads is not possible.

Selection

On each learning environment *(2Connect/ Astroversity),* the dyads were classified in terms of high affective similarity (dyads in the 4th quartile of the affective awareness index) and low affective similarity (dyads in the 1st quartile of the affective similarity index). For each activity, two dyads with low affective similarity and two dyads with high affective similarity were randomly selected to make a total of eight dyads.

Table 15 shows the affective awareness indexes and the pseudonyms assigned to the members of the selected dyads.

Table 15. Affective similarity level, affective similarity index and pseudonyms of the partners in the example dyads

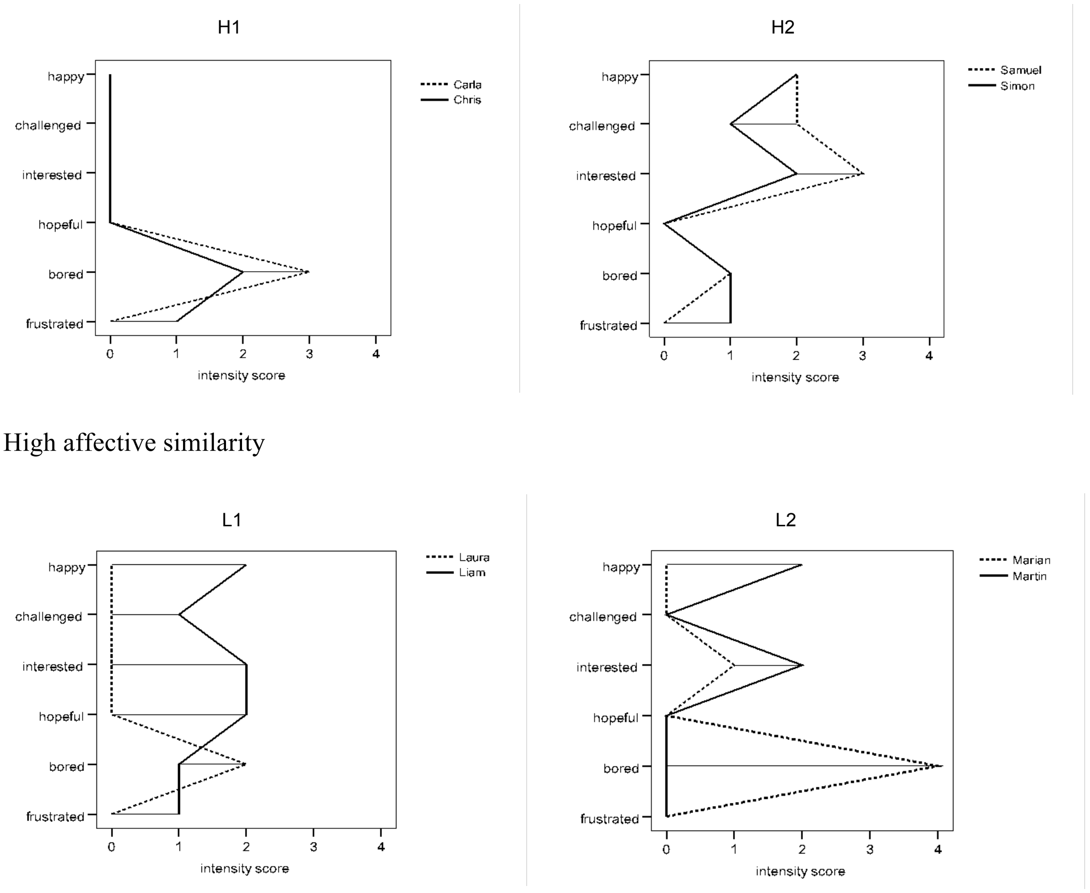
Learning environment	Affective similarity level	Dyad ID	Affective similarity index	Pseudonyms
2connect	High	HS1	.88	Chris and Carla
		HS2	.80	Samuel and Simon
	Low	LS1	-.16	Martin and Marian
		LS2	-.44	Laura and Liam
Astroversity	High	HS1	.60	Paola and Peter
		HS2	.75	Nora and Natalie
	Low	LS1	.21	Arthur and Armand
		LS2	-.72	Sophie and Sarah

Examples of collaborations around 2Connect

Emotion profiles

Figure 9 shows the emotion profiles of the dyads collaborating around 2connect. It shows that collaborators with high affective similarity were either equally bored or equally enjoying the activity. The profile of Chris and Carla (H1) show a 'flat' emotional intensity except for boredom. Conversely, the profile of Samuel and Simon (H2) shows that both of them enjoyed the task -they felt happy, challenged and interested with moderate to a lot of intensity, and felt less bored, frustrated and hopeful.

In dyads with low affective similarity, in the case of L1, Liam reported more happiness, interest, hope and frustration than Laura, who did not report these emotions at all. In the case of L2, the difference was that Martin was moderately happy and not bored, whereas Marian was not happy at all and extremely bored.



High affective similarity

Low affective similarity

Intensity score: 0= not at all, 1 = slightly, 2= moderately, 3= a lot, 4 = extremely

Figure 9. Emotion profiles of dyads with high affective similarity (H1, H2) and low affective similarity (L1, L2) during the use of 2connect

Interaction quality

Affective expressions. Individuals rarely showed positive or negative affective expressions, at least not in a clear manner (e.g., they did not laugh or made jokes). Although partners in dyad L2, Chris and Carla, equally bored, yawned and explicitly commented about the dullness of the activity.

Responsiveness. One key feature of collaborators' interaction, that probably reflected their emotion profiles, was the responsiveness with the partner. For example, in Dyads L1 and L2, the individuals who reported less happiness, challenge and interest (Laura and Marian) responded to the ideas of the partner, but this was not reciprocal. For example, in dyad L2, Laura reported more boredom and less interest than Liam. She proposed ideas and also complemented the ideas proposed by Liam. In contrast, Liam proposed ideas but did not complement or request Laura's ideas. Often, he responded to Laura's questions with elaborated accounts of his own student experiences without asking hers. For example:

Laura: *and then, are you travelling for the summer?*
Liam: *ehm… I am going to [long list of places] and then to Australia*
Laura: [turning to him, smiling] *jesus, all in the summer?*
Liam: *yeah, with my friends and then to Australia with my family*
Laura: *and you are in what year?*
Liam: *first year*
Laura: *Psychology?*
Liam: *Chemistry*

Responsiveness between partners was also relevant in dyads H1 and H2. For example, partners of dyad H1 felt equally bored and not happy or interested at all. In this case, Carla was usually more responsive than Chris. Chris frequently added topics to the concept map without saying anything to Carla or typed whatever she proposed, not asking for more nor complementing her ideas. Sometimes, Carla asked Chris what he was typing but his responses were short. For example:

Chris: [silently typing]
Carla: *what are you putting?*
Chris: *ehm, excellent…* [pause] *I do not know what I am trying to say [pause] how good the teach is*
Carla: *excellent teaching, I guess…*

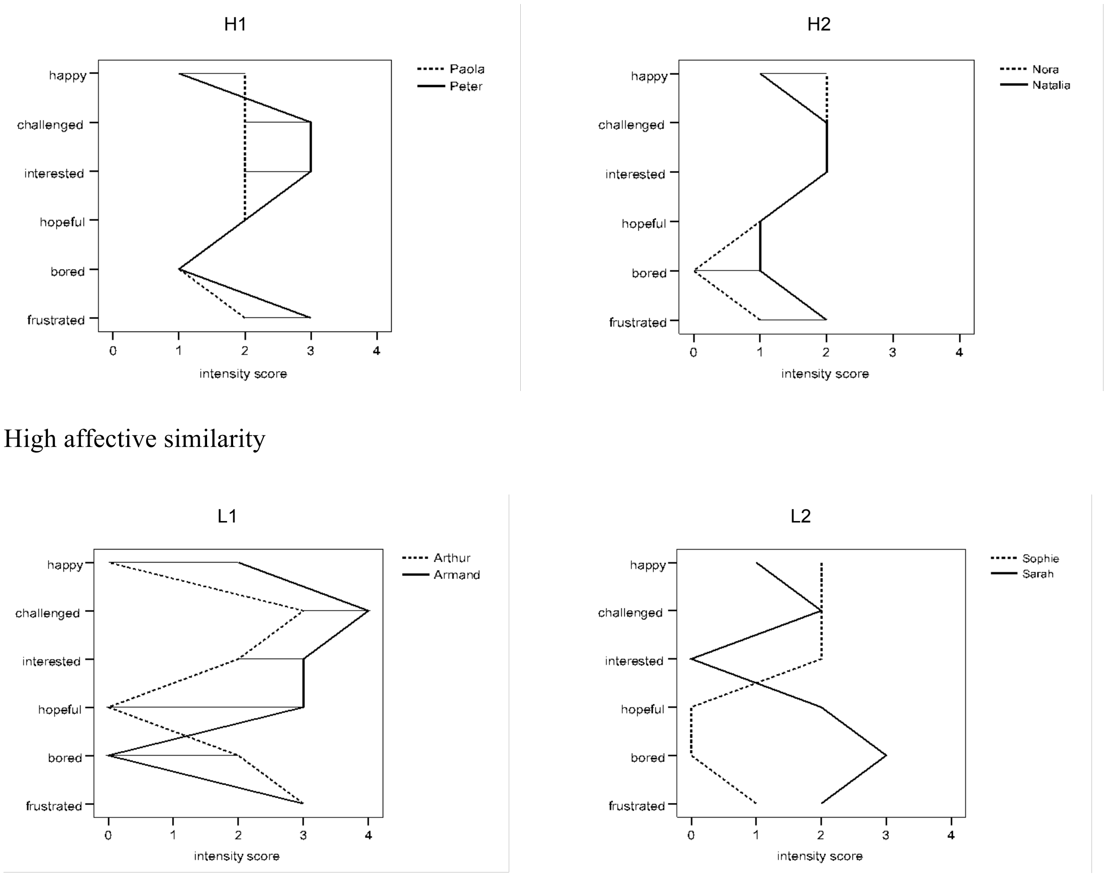
Differently, partners in dyad H2 reported similar happiness and interest. In this case, they were usually responsive to the propositions of one another. For example, in the following extract, Samuel is proposing a topic to include in the concept map. Simon responds with further information on the topic and Samuel responds back:

Samuel: [proposing a pro of their university] *good rank international*
Simon: *what is our rank international?*
Samuel: *73 or something*
Simon: *is not bad... currently we are in the 20 something in the UK*
Samuel: *is sixth*
Simon: *sixth?*
Samuel: *yeah...*

Examples of collaborations around Astroversity

Emotion profiles

Figure 10 shows the emotion profiles of dyads playing Astroversity. Members of dyads with high affective similarity were 'emotionally' engaged with the activity. Paola and Peter (H1) and Nora and Natalie (H2), reported low boredom, moderate happiness, interest, challenge, hope and frustration. In contrast, in dyads with low affective similarity, one partner showed more engagement. In L1 Armand reported more happiness and hope and in L2 Sarah reported more interest, frustration and hope.



High affective similarity

Low affective similarity

Intensity score: 0= not at all, 1 = slightly, 2= moderately, 3= a lot, 4 = extremely

Figure 10. Emotion profiles of dyads with high affective similarity (H1, H2) and low affective similarity (L1, L2) during the use of Astroversity

Interaction quality

Affective expressions. Partners in dyad L2 reported similar intensities of happiness and interest, but Sarah reported less interest and, in general, more negative emotions such as hope, boredom and frustration. Nevertheless, neither Sarah nor Sophie showed affective expressions. In Dyad L1, partners reported similar challenged and interest, but Arthur reported more happiness and negative emotions and showed more affective reactions than Armand.

Partners in dyads H1 and H2, who reported positive emotions similarly (e.g., more challenge and interest than boredom), frequently showed positive affective expressions. Partners displayed simultaneous affective reactions towards Astroversity, especially when animated characters gave them recommendations and feedback with expressive elements. For example, in H1, Peter and Paola made make jokes about the feedback of the game, as in the following extract:

> Peter: [reading the interface and smiling] *ok, brain damage but stable*
> Paola [smiling]: *but he’s all right*
> Peter: [smiling] *what!?*

> Paola [reading the interface and smiling]: *you need more practice as a team!*
> Peter: [laughs]
> Paola: [laughs]

Also, the conversations of Paola and Peter (H1) and Nora and Natalie (H2) usually included jokes and/or affective expressions. In the following extract for example, Peter is listing coordinates while scanning alien presence. Paola is plotting and then, joyfully expressed that they were making progress.

> Peter: *N18 is zero*
> Paola: *awesome, N18 is zero... ok we seem to have some sort of progression*
> Peter: *N17 is also zero*
> Paola: [cheers]

Responsiveness. Partners in dyads H1 and H2 usually asked and proposed ideas to one another. For example, in dyad H1, when Peter was telling the coordinates with highest alien presence for Paola to plot them in the grid, she often proposed other coordinates to look at. Accordingly, Peter responded and scanned these coordinates. Conversely, when Paola listed the more secure coordinates for Peter to trace the safest route, he noticed that some of the coordinates listed by Paola were actually dangerous and proposed alternative routes. Similarly, in dyad H2, Natalie and Nora displayed more responsiveness with the passing of time. Their communication evolved from short exchanges combining silence, monologue and backchannel, to a sort of interaction in which they asked, proposed or elaborated about each other's ideas and opinions. In the following extract, Nora and Natalie show responsiveness while scanning alien presence:

> Natalie: But, [the alien presence] *is also high this side maybe we can try this area*
> Nora: *yeah but this is a long...*
> Natalie: [backchannels] *thus*
> Nora: *It can't be trouble through it, as long as you go round on it, is that a good idea?*
> Natalie: *I am not to sure*
> Nora: *because in that case we can go over here, somewhere* [laughs]
> Natalie: [backchannels] *yeah...*
> Nora: *like… to G15*
> Natalie: [goes to plot]

Partners in dyads L1 and L2 showed less propositions and requests of ideas to one another. For example, Arthur and Armand regularly proposed ideas, but their responses were often limited to a minimal backchannel. It was usually Armand, the one who enjoyed the activity more, who was feeding the conversation while Arthur just complemented. The following extract illustrates some of these forms of response. Armand is scanning alien presence and Arthur is plotting in the paper grid.

> Armand: *R5?*
> Arthur: *mhmm*
> [silence]
> Arthur: we are in...
> Armand: *no...* [mumbling]
> [silence]
> Arthur: *That's high isn't it?*
> Armand: *yeah...*
> Arthur: [so where is the...]
> Armand: [we are at R5.] *which is here... which wasn't that bad*
> Arthur: *yeah...*

Similarly, Sophie and Sarah usually responded with backchannel but with no further elaboration to one another's ideas, as in the following extract:

Sarah: [scanning] *ok, so now this route is no longer safe*
Sophie: [backchannel] *mhh...*
[…]
Sarah: *the entrance is no longer safe*
[…]
Sarah: *where is the middle part?*
Sophie: [backchannel] *erm…*
[…]
Sarah: [monologue, reacting to the interface] *what?*
Sophie: *erm...*

Coordination. Dyads H1 and H2 usually did the collaborative tasks of Astroversity with a clear role assignation, which helped them to coordinate their actions during the collaborative task of Astroversity. For example, in dyad H1, when Peter and Paola were scanning alien presence, Peter controlled the rover telling the coordinates with high or low alien presence to Paola, who plotted them in the grid using a constant backchannel. Conversely, when tracing the safest route for the student to be rescued, Paola told the coordinates she plotted in the grid while Peter traced the route in the computer and used backchannel. Dyad H2 organized similarly. The following extract is taken from a moment when Nora and Natalie were tracing the safest route for the student to be rescued:

Nora: [points to the screen] *right, say...that one?*
Natalie: yeah... I20
Natalie: [backchannel] *I20... I19*
Nora: I19, got it
Natalie: yeah... N19
Nora: *N19?*
Natalie: *M19*
Nora: *M?*
Natalie: *yeah*
Nora: *got it*

Dyads L1 and L2 organised differently to make the collaborative tasks of Astroversity. For example, in Dyad L1, Armand, who reported more positive emotions, controlled the computer for detecting alien presence while Arthur plotted in the paper grid. Conversely, Arthur controlled the computer and Armand suggested coordinates in the part for tracing the safest route. In the case of Dyad L2, partners were more disorganised. For example, in the part of Astroversity where the dyad LS2 was scanning the spaceship to detect alien presence, Sarah scanned the screen while Sophie was supposed to be plotting in the paper grid. The next task consisted of tracing the safest route to escape, avoiding the alien presence plotted in the paper grid. Sophie was supposed to tell the coordinates for Sarah to trace the route, but Sophie did not plot while Sarah was scanning and consequently, they failed the trial.

Integrative summary of dyad examples

The presented examples revealed some features of the interaction with the task environment and with the partner that might be related to the individuals' emotions, and the similarity of emotions between partners in this studys.

It was common for individuals under different circumstances not to show affective expressions, as it occurred in the four dyads using 2Connect, or in dyads who reported dissimilar emotions around Astroverstity. The expression of positive emotions was common in partners who reported similar emotions whilst playing Astroversity (Paola and Peter; and Nora and Natalie). They laughed or made jokes as a response to features of Astroversity such as the characters that give feedback, or as part of their ongoing interaction whilst resolving the collaborative tasks of the game.

The partners' response to one another also seemed to be related to the partners' emotions. Individuals whose partners showed little responsiveness (e.g., did not ask for, or elaborated on, the ideas of the other) also reported less positive emotions, as it occurred to Laura and Marian whilst using 2Connect, or to Armand whilst playing Astroversity. Furthermore, partners in dyads feeling more similarly positive emotions around Astroversity (Paola and Peter; and Nora and Natalie) and 2Connect (Simon and Samuel) were also more responsive with one another.

Finally, the sort of organisation displayed by partners whilst playing Astroversity might have also been related to their emotions. This aspect was not observed in the examples of dyads around 2Connect, presumably because partners did not have to execute any sort of organised interaction (in the sense of action coordination in a specific task). Thus, Paola and Peter; and Nora and Natalie; who similarly enjoyed Astroversity, displayed an effective coordination during collaborative tasks (e.g., searching alien presence), based on a clear role assignation. In contrast, collaborators who felt different emotions showed different organisations, such as Arthur and Armand swapping the control of the computer and the paper grid, or Sarah and Sophie being unable to clarify roles to play.

Discussion

This study compared collaborations around the concept-mapping tool 2Connect and the collaborative educational computer game Astroversity to explore how the usage of these technologies affects collaborators' emotions. This discussion integrates the results. First, the methodological findings are explained. Then, the results are discussed to answer the research questions.

Methodological findings

There were two methodological aims related to the usage of questionnaires for collaborators to report their emotions. The first was to select emotions relevant to the CSCL experience. The second was to assess whether the inclusion of scales for frequency and intensity would help to obtain fruitful data about collaborators' emotions.
No studies were found that reported a list, or a 'survey' of the emotions people feel during CSCL. Therefore, one methodological aim was to select the emotions relevant to the use of 2Connect and Astroversity. Initially, 15 words that refer to emotions differentiable in terms of their underlying cognitive appraisals were included in the questionnaires: *happy, angry, sad, fearful, angry, bored, challenged, interested, hopeful, frustrated, contempt, disgusted, surprised, proud, ashamed* and *guilty.*

There were extreme floor effects for contempt, sad, guilty, fearful, angry, ashamed and disgusted. Presumably participants did not report these emotions because they are negative. It has been found that 'negative' emotions, e.g., anger, sadness, fear, guilt and shame, are reported with less frequency and intensity than other, more 'positive' emotions (Carstensten, et al., 2000; Nezlek, et al., 2008; Tong, et al., 2007). However, other explanation is that some of these emotions are simply not part of the average CSCL experience, (e.g. disgust), or probably occur with low frequency, (e.g., surprise).

Thus, the study focused on the emotions happy, interested, challenged, hopeful, frustrated and bored. The rationale for this selection was to have two positive emotions (happy and challenged), two emotions that imply both positive and negative attributes (challenged and hopeful) and two negative emotions (frustration and bored).

The second methodological aim was to explore whether participants could differentiate between the intensity and the frequency of their emotions. That could have been useful to detail the collaborators' emotions more. However, the high correlations and small differences between the scores of frequency and intensity showed that participants did not differentiate between these aspects of their emotions. Therefore the study focused on intensity, as is common in emotion research.

How does using a concept-mapping tool and a collaborative educational game influence people's emotions?

The Introduction section explained the differences between the collaborative usages of 2Connect and Astroversity in terms of their underlying tasks and the affective features of their designs. This served as a framework for the analyses that assessed the differences between the emotions reported by collaborators around these learning environments and what they referred as the sources of their emotions.

Activities around Astroversity and 2Connect were equally pleasant and engaging, provoking equally mild levels of happiness and interest, as well as low levels of boredom. However, playing Astroversity provoked more intense 'goal-oriented' emotions such as challenge, hope and frustration.

Although with similar intensity, 2Connect and Astroversity might have prompted pleasantness and engagement in different ways. Collaborators used 2Connect to make a concept map about their student experiences. 2Connect has very few affective elements (e.g., expressivity, anthropomorphism) and therefore it should have had little influence on collaborators' emotions. Also, partners were not allocated to predefined roles, no specific characteristics for the concept-map were requested and no other materials (e.g., paper and pencils) were provided.
Probably the interaction with a partner was the key for the pleasantness and engagement in 2Connect. Partners probably felt positive affect (e.g., more interest and happiness) if they showed motivation to understand each other (Schwartz, 1998). Some illustrative examples are in line with this. Partners who were mutually responsive (i.e., asked and elaborated on the ideas of the other) also reported similar happiness and interest (Samuel and Simon), whereas partners who were not reported equal boredom (Chris and Carla) and those who interacted with an unresponsive partner reported low enjoyment (Laura in L1 and Marian in L2).

Pleasantness and engagement might have appeared differently in Astroversity. Partners resolved tasks with specific goals (e.g., scan alien presence), playing complementary roles (e.g., plotting on the paper grid whilst the partner scans alien presence on the computer) and employing supplementary materials (e.g., paper and pencil). These aspects were embedded in an environment with affective features such as expressivity, aesthetics and anthropomorphised feedback. In this context, partners' probably felt positive affect as they were effectively playing their roles, successfully resolving the tasks, and interacting with the affective features of the game (e.g., feedback), as observed in some illustrative examples. For instance, partners who reported similarly positive emotions around Astroversity (H1 and H2) were also more responsive to one another and coordinated effectively to resolve the tasks, and laughed and joked about the animated characters that gave them feedback.

However, even when both learning environments provoked equal pleasantness and engagement, participants did felt more intense feelings of challenge, hope and frustration whilst playing Astroversity. To complement this result, the analysis turned to what collaborators identified as sources of these emotions.

Participants made a significant but small differentiation between the partner, themselves and the activity as sources of their challenge, hope and frustration whilst using 2Connect. The small differentiation is probably related to the fact that partners felt these emotions with low intensity, (lower than other emotions such as happiness and interest). In contrast, during Astroversity, participants made clear differentiations when rating the sources of these emotions. They attributed more emphatically to the activity than to themselves or their partner as an emotional source, more in relation to challenge and frustration than in relation to hope.

This suggests that making effort to achieve the specific goals of the tasks, playing of complementary roles, and interacting with the affectively loaded GUI of Astroversity, were main sources of challenge and frustration. Moreover, all these aspects are absent in 2Connect, which probably explains the difference in the intensities of these emotions across learning environments. Moreover, in the illustrative examples, partners around Astroversity showed coordinated action and played complementary roles to resolve tasks like scanning alien presence, a sort of interaction not observed in the cases around 2Connect.

Research on cognitive appraisal and emotions is helpful to further interpret these results. It is not surprising that whilst playing Astroversity, the activity was the main source of challenge, hope and frustration, and that these emotions were more intensely reported than whilst using 2Connect. Challenge, hope and frustration are emotions associated with situations that demand high *effort* and *attentional activity* (Reisenzein & Hofmann, 1993; Smith & Ellsworth, 1985), as it was required to resolve the tasks of Astroversity. For example, whilst scanning alien presence, collaborators have to act quickly and collect as much data as they can. However, the task requirements might not be the only source of challenge, hope and frustration. Probably the affective features also played a key role. For example, frustration could have been triggered when collaborators received the feedback about their performance from expressive characters, denoting some form of anthropomorphism.

However, although challenge, frustration and hope are similarly associated with the appraisals of effort and attention, there are other appraisals that differentiate them. Challenge and hope are opposites in the appraisal dimension of *certainty.* Challenge is experienced when a person is certain about the outcome of a situation, whereas hope is associated with situations with an uncertain outcome. Some parts of Astroversity are ‘intentionally’ designed to generate uncertainty. For example, when players have to scan invisible alien presence or wait for feedback.

Challenge and frustration are opposites in the appraisal dimension of *perceived situational control*, which ranges from self-agency to other–agency. Challenge is associated with the sense of self-agency, which implies the control over the situation. In contrast, frustration is associated with the sense of other-agency, which implies that the situation is controlled by another ‘agent’, a person or, as in the case of this study, a computer program (i.e., Astroversity). On the one hand, collaborators had control of their own actions whilst playing Astroversity. This could have generated a sense of self-agency and the subsequent feeling of challenge. In contrast, the feelings of frustration could have been generated by the fact that collaborators could not control neither the actions of the partner or the structure of the game. A sense of ‘other-agency’ in relation to the partner might have aroused because in order to achieve the win state, collaborators largely depended on each other. A sense of ‘other-agency’ in relation to the computer might have aroused because collaborators had to follow the rules ‘imposed’ by Astroversity with no control over its mechanics.

What do collaborators understand about their partners' emotions while using a concept-mapping tool and a collaborative educational game?

The study also explored what collaborators understand about the emotions of a partner. Two aspects were analysed: the similarity between partners' emotions and their judgements about the emotions of one another.

Similarity between partners' emotions

The affective similarity between collaborators was not higher than would be expected by chance, neither in the overall assessment across learning environments, nor in the separate tests of 2Connect and Astroversity. One interpretation is that collaborators' emotions were unrelated because they depended on the emotions they felt before the study. But this interpretation is not supported since the multivariate relationship between collaborators' prior emotion scores and their emotion scores during 2Connect and Astroversity was not significant [F (18, 108) = .42, *ns*)].

Apparently, the statistical test of the emotional similarity between partners was not significant because some partners reported very similar emotions, and others did so very differently, as indicated by the large SD's of the affective similarity index. At least two explanations are possible for this.

One explanation is that the 5-point Likert scale used by collaborators to report their emotions was too short and collaborators made a 'forced choice'. This could have increased both the measurement error and the probability that collaboration partners selected different options to report their emotions.

The other explanation is that the large differences between dyads' affective similarities were due to within-dyad factors. That is, some dyads were different to others in a way that affected their affective similarities. The illustrative cases showed that some partners showed different interaction qualities, probably related with the similarity of their reported emotions. For example, partners who reported similar emotions also showed more responsiveness, around 2Connect and Astroversity, and only around Astroversity also showed more coordination and positive affective expressions.

Accuracy at judging the emotions of a partner

Affective awareness was measured as the accuracy of collaborators at judging the emotions of their partners; taking the emotions reported by the partner as 'true' reference. The measurement of affective projection referred to the degree in which collaborators based their judgements of their partners' emotions on their own emotions.

In general, collaborators' affective awareness was rather low, both during the use of 2Connect and during the use of Astroversity. One interpretation is that whilst interacting around these learning environments, collaborators did not pay attention to the emotional expressions of their partners, or to the particular events that could have affected their emotions. However, it is also possible that the inaccuracy at reporting the emotions of the partner is the result of having only one opportunity to do so at the end of each activity. Probably the participants could have been more accurate if they had more occasions to report the emotions of their partners.

It was found that in general, collaborators showed a tendency to 'project' their own emotion onto their partners. This suggests that collaborators based their understanding of their partner's emotions on their own emotions. Probably collaborators did so to avoid effort at reporting the emotions of the partner. Or probably they did put effort at reporting the emotions of the partner but felt they lacked information to do so.

This is consistent with the bias in interpersonal perception known as the 'false consensus effect'. That is, a person will think that others in the same situation have the same opinion (for a review, see Marks & Miller, 1987). This tendency was lower during the use of Astroversity than during the use of 2Connect. Probably, when using Astroversity, collaborators projected less because the partner expressed emotions more emphatically, or because the game mechanics and narrative generated situations that could be clearly identified as sources of the partner's emotions. However, even when collaborators showed less affective projection during Astroversity, this did not imply that they were more aware about the emotions of the partner.

Importantly, these results were calculated with control of the emotional similarity between partners. A number of multiple regression analyses were employed to predict' the collaborators' judgements about the emotions of their partners. These included both the participants' own emotions and the emotions felt by their partners as dependent variables. In this way, the individual predictive power (i.e., the beta coefficients) of these variables has been calculated controlling for their relationship (i.e., the emotional similarity between partners). A high relationship could have provoked 'collinearity', reflected on very large standard errors and, consequently, a reduced significance for most of the beta coefficients. However, the results show that the beta coefficients of the participants' own emotions were fairly high (≅ .50) and that overall, the standard errors are not extremely large (≅ .10). Thus, the analysis is reliable and 'conceptually' valid. It is recognised that feeling in the same way as the partner is a potential influence on the understanding of a partners' emotions. However, the results suggest that the similarity with the partner is relatively independent of both the collaborators' tendency to project their emotions onto their partners and of their low affective awareness.

What is the relationship between collaborators' emotion and the qualities of their interaction with a partner?

The relationship between collaborators' emotions and their perceived interaction quality was assessed with joint measures of emotion intensity and perceived interaction quality. In the case of collaborators' emotions, the joint measure was the sum of the emotion intensity scores of both partners. In the case of perceived interaction quality, the joint measure was the sum of the perceived interaction quality scores of both partners.

Both in 2Connect and Astroversity, the joint measure of interaction quality was positively correlated with the joint measure of happiness, and negatively correlated with the joint measure of boredom. Recall that collaborators reported similarly mild levels of happiness and low levels of boredom around the two learning environments. Altogether, these results suggest that collaborators' enjoyed more if they thought they were thinking alike, understanding each other and cooperating equally; irrespective of the tasks and affective features of the technology. The qualitative examples illustrated this. Individuals who felt more positive emotions also had partners who asked and responded more to her ideas. Moreover, in the cases of partners who reported similarly positive emotions around Astroversity, the two partners were responsive to one another. Furthermore, the partners around 2Connect that only reported boredom rarely responded to one another.

In Astroversity, but not in 2Connect, the dyads' joint measure of perceived interaction quality was negatively correlated with their joint measure of frustration. Recall that frustration was one of the emotions that collaborators reported more intensely during Astroversity than during 2Connect. Probably the collaborative tasks of Astroversity prompted, but did not guaranteed, a sort of interaction quality that helped partners to increase their control of the situation and reduce uncertainty, leading to a reduced frustration. Again, this is illustrated in the dyad examples. Those who reported similarly positive emotions (e.g., happiness) and less negative emotions (e.g., frustration) around Astroversity, often took clearly defined roles to solve the collaborative tasks. In comparison, those dyads in which at least one partner felt frustration played less fluently, and the role assignation between partners was unclear and/or one of the partners did not play her role effectively. This sort of relationship between interaction quality and emotions was not observed around 2Connect, presumably because partners did not have to execute fixed task that required coordinated action.

Finally, the dyads' joint measure of perceived interaction quality was not correlated with their joint measures of interest, challenge and hope. This suggest that collaborators' assessed the quality of their interaction with a partner independently of the attention they put in the activity (interest), the positive aspects of their engagement (challenge) or their expectations about the activity (hope).

Conclusions

This study compared the affective effects of collaborations around clearly different collaborative learning environments such as the game Astroversity and the concept-mapping tool 2Connect. Partners reported more goal-oriented emotions (e.g., challenge and frustration) in relation to the collaborative tasks of Astroversity. But there were no differences in terms of emotion understanding since partners showed a tendency to assume similarity, which was not necessarily true. This indicated that partners had little awareness about the emotions of the partner. In turn, emotions seemed to be associated with interaction quality, which probably suggests that partners influence each other's emotions during collaborative interaction; and idea illustrated by dyad examples (e.g., partners who showed mutual responsiveness also reported more positive emotions).

These findings suggest that the resolution of collaborative tasks with fixed goals and complementary roles are features of collaborative games that might facilitate joyful and productive collaborations. However, whether this occurs or not, seemeds to depend on the interaction with a partner. It is also important to recognise the limitations of the study. Partners reported their emotions and the emotions of their partners in one occasion, at the end of each activity. Therefore, moments of especially intense emotions could not be identified. Moreover, because there were no screen recordings, it was not possible to link emotions or emotion understanding with performance. Also, participants' collaborative tasks were fairly short, leaving uncertain whether collaborators' emotions might change over time.

Appendix A. Regression analyses testing accuracy at reporting the partner’s emotions

		Own				Partner			
		B	*SE*	*b*	*t*	*B*	*SE*	*b*	*t*
2Connect	Happy	0.52**	0.09	0.61	5.5	0.17	0.09	0.2	1.8
	Interested	0.40**	0.1	0.51	4.03	0.04	0.1	0.06	0.45
	Challenged	0.52**	0.11	0.55	4.54	0.003	0.115	0.004	0.03
	Hopeful	0.56**	0.14	0.5	4.04	-0.13	0.14	-0.11	0.92
	Frustration	0.33[1]	0.15	0.3	2.17	-0.07	0.15	-0.063	-0.45
	Bored	0.58**	0.12	0.55	4.7	0.20	0.12	0.2	1.68
Astroversity	Happy	0.36*	0.13	0.37	2.82	0.25	0.13	0.36	1.96
	Interested	0.30[1]	0.12	0.34	2.63	0.32*	0.11	0.36	2.8
	Challenged	0.24[1]	0.12	0.28	2.03	0.16	0.12	0.18	1.33
	Hopeful	0.58**	0.13	0.55	4.6	0.18	0.13	0.17	1.44
	Frustration	0.52**	0.125	0.52	4.15	-0.02	0.12	-0.02	-0.17
	Bored	0.30[1]	0.13	0.32	2.3	0.12	0.13	0.13	0.9

[1] p<.05, *p<.01, **p<.005

ENDNOTE GENERATED REFERENCES FOR ALL CHAPTERS

Ainsworth, S. (2007). Using a Single Authoring Environment across the Lifespan of Learning. *Journal of Educational Technology & Society, 10*(3), 22-31.

Aleven, V., Koedinger, K., & Cross, K. (1999). Tutoring Answer Explanation Fosters Learning with Understanding. In S. P. Lajoie & M. Vivet (Eds.), *Artificial Intelligence in Education. Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration* (Vol. 50, pp. 199-206): IOS Press.

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward Tutoring Help Seeking; Applying Cognitive Modeling to Meta-Cognitive Skills. In J. C. Lester, R. M. Vicari & F. Paraguacu (Eds.), *7th International Conference on Intelligent Tutoring Systems, ITS 2004, Maceio, Brazil, Proceedings* (Vol. Lecture Notes in Computer Science 3220, pp. 227-239): Springer.

Alsmeyer, M., Luckin, R., & Good, J. (2007). Getting Under the Skin of Learners: Tools for Evaluating Emotional Experience. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (Vol. Frontiers in Artificial Intelligence and Applications 158, pp. 153-160). Amsterdam: IOS Press.

Ames, C. (1992). Classrooms: Goals, Structures, and Student Motivation. *Journal of educational psychology, 84*(3), 261-271.

Anderson, J. R., Corbett, A. T., & Koedinger, K. R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences, 4*(2), 167-207.

Anderson, J. R., Farrell, R., & Sauers, R. (1984). Learning to program in LISP. *Cognitive Science, 8*(2), 87-129.

Ang, S. O., Chen, H., Hirota, K., Gordeuk, V. R., Jelinek, J., Guan, Y., et al. (2002). Disruption of oxygen homeostasis underlies congenital Chuvash polycythemia. *Nature Genetics, 32*, 614-621.

Arroyo, I., Beal, C., Murray, T., Walles, R., & Woolf, B. P. (2004). Web-Based Intelligent Multimedia Tutoring for High Stakes Achievement Tests In J. C. Lester, R. M. Vicari & F. Paraguacu (Eds.), *7th International Conference on Intelligent Tutoring Systems, ITS 2004, Maceio, Brazil, Proceedings* (Vol. Lecture Notes in Computer Science 3220, pp. 469-477). Berlin: Springer.

Arroyo, I., Beck, J. E., Woolf, B. P., Beal, C. R., & Schultz, K. (2000). Macroadapting Animalwatch to Gender and Cognitive Differences with Respect to Hint Interactivity and Symbolism In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *5th International Conference on Intelligent Tutoring Systems, ITS 2000* (Vol. Lecture Notes in Computer Science 1839, pp. 574-583). Montreal, Canada: Springer.

Avramides, K., & du Boulay, B. (2009). Motivational Diagnosis in ITSs: Collaborative, Reflective Self-Report. In V. Dimitrova, R. Nizoguchi, B. du Boulay & A. Graesser (Eds.), *Artificial Intelligence in Education. Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* (Vol. Frontiers in Artificial Intelligence and Applications 200, pp. 587-589): IOS Press.

Baines, E., Blatchford, P., & Kutnick, P. (2004). Changes in grouping practices over primary and secondary school *International Journal of Educational Research, 39*(1-2), 9-34.

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why Students Engage in "Gaming the System" Behaviours in Interactive Learning Environments. *Journal of Interactive Learning Research, 19*(2), 185-224.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). *Off-task behavior in the cognitive tutor classroom: when students "game the system"*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Baker, R. S. J. d. (2007). *Modeling and understanding students' off-task behavior in intelligent tutoring systems*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S., Roll, I., Wagner, A. Z., et al. (2006). Adapting to When Students Game an Intelligent Tutoring System. In M. Ikeda, K. D. Ashley & T.-W. Chan (Eds.), *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, Proceedings* (Vol. Lecture Notes in Computer Science 4053, pp. 392-401): Springer.

Baker, R. S. J. d., Corbett, A. T., Roll, I., & Koedinger, K. R. (2008). Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*(3), 287-314.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments *International Journal of Human-Computer Studies, 68*(4), 223-241.

Baker, R. S. J. d., Rodrigo, M. M. T., & Xolocotzin, U. E. (2007). The Dynamics of Affective Transitions in Simulation Problem-Solving Environments In A. Paiva, R. Prada & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007, Lisbon, Portugal, Proceedings* (Vol. Lecture Notes in Computer Science 4738, pp. 666-677): Springer.

Banse, R., & Scherer, K. R. (1996). *Journal of Personality and Social Psychology, 70*(3), 614-636.

Barrett, F. L. (2006). Are Emotions Natural Kinds? . *Perspectives on Psychological SCience, 1*(1), 28-58.

Batson, C. D., Turk, C. L., Shaw, L. L., & Klein, T. R. (1995). Information function of empathic emotion: Learning that we value the other's welfare. *Journal of Personality and Social Psychology, 68*(2), 300-313.

Beal, C. R., Mitra, S., & Cohen, P. R. (2007). Modeling learning patterns of students with a tutoring system using Hidden Markov Models. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (Vol. Frontiers in Artificial Intelligence and Applications 158). Amsterdam: IOS Press.

Beck, J. E. (2005). Engagement tracing: using response times to model student disengagemen. In C.-K. Looi, G. McCalla, B. Bredeweg & J. Breuker (Eds.), *Artificial Intelligence in Education, Supporting Learning through Intelligent and Socially Informed Technology* (Vol. Frontiers in Artificial Intelligence and Applications 125, pp. 88-95). Amsterdam: IOS Press.

Bonanno, G. A., & Keltner, D. (2004). The coherence of emotion systems: Comparing "on-line" measures of appraisal and facial expressions, and self-report *Cognition and Emotion, 18*(3), 431-444.

Boucsein, W. (1992). *Electrodermal Activity*. New York, NY: Plenum Press.

Burleson, W. (2006). *Affective Learning Companions: strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance.* Massachusetts Institute of Technology, Cambridge, MA.

Burleson, W., & Picard, R. (2004). *Affective agents: sustaining motivation to learn through failure and a state of stuck*. Paper presented at the Workshop on social and emotional intelligence in learning environments, 7th International Conference on Intelligent Tutoring Systems.

Burleson, W., & Picard, R. (2007). Evidence for gender specific approaches to the development of emotionally intelligent learning companions. *IEEE Intelligent Systems, Special issue on Intelligent Educational Systems, 22*(4), 62-69.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin, 56*(2), 81-105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin Compamy.

Cañas, A. J., & Novak, J. D. (2006). *Re-examining the foundations for effective use of concept maps*.

Carstensten, L., Pasupathi, M., Mayr, U., & Nesselroade, J. R. (2000). Emotional experience in everyday life across the adult life span. *Journal of Personality and Social Psychology*(79), 644-655.

Chanel, G., Ansari-Asl, K., & Pun, T. (2007). *Valence-arousal evaluation using physiological signals in an emotion recall paradigm*. Paper presented at the International Conference on Systems, Man and Cybernetics.

Chen, J., & Lemon, O. (2009). Robust Facial Feature Detection and Tracking for Head Pose Estimation in a Novel Multimodal Interface for Social Skills Learning *Proceedings of Advances in Visual Computing. 5th International Symposium, ISVC 2009, Part II* (Vol. 5876, pp. 588-597). Berlin: Springer-Verlag.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*(2), 145-182.

Clarke, D., & Duimering, P. R. (2006). How computer gamers experience the game situation: a behavioral study. *Computers in Entertainment, 4*(3).

Coan, J. A., & Allen, J. J. B. (Eds.). (2007). *Handbook of emotion elicitation and assessment*: Oxford University Press.

Conati, C., Chabbal, R., & Maclaren, H. (2003). *A Study on Using Biometric Sensors for Monitoring User Emotions in Educational Games*. Paper presented at the Proceedings of the Workshop "Assessing and Adapting to User Attitude and

Affects: Why, When and How? In UM '03, 9th International Conference on User Modeling.

Conati, C., Gertner, A., & VanLehn, K. (2002). Using Bayesian Networks to Manage Uncertainty in Student Modeling *User Modeling and User-Adapted Interaction, 12*(4), 371-417.

Conati, C., & Maclaren, H. (2009). Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction, 19*(3), 267-303.

Conati, C., & Zhou, X. (2002). Modeling Students' Emotions from Cognitive Appraisal in Educational Games. In S. A. Cerri, G. Guy & F. Paraguacu (Eds.), *Intelligent Tutoring Systems. 6th International Conference, ITS2002, Biarritz, France and San Sebastian, Spain, Proceedings* (Vol. Lecture Notes in Computer Science 2363, pp. 944-954). Berlin: Springer.

Cowie, R. (2005). What are people doing when they assign everyday emotion terms? . *Psychological Inquiry, 16*(1), 11-48.

Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication, 40*(1-2), 5-32.

Craig, S., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor *Learning, Media and Technology, 29*(3), 241-250.

Craig, S. D., D'Mello, S., Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning *Cognition & Emotion, 22*(5), 777-788.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal performance*. New York: Cambridge University Press.

D'Mello, S., Graesser, A., & Picard, R. W. (2007). Toward an affect-sensitive AutoTutor. *IEEE Intelligent Systems, 22*(4), 53-61.

D'Mello, S., Taylor, R., Davidson, K., & Graesser, A. (2008). Self Versus Teacher Judgments of Learner Emotions During a Tutoring Session with AutoTutor In B. P. Woolf, E. Aïmeur, R. Nkambou & S. Lajoie (Eds.), *Intelligent Tutoring Systems, 9th International Conference, ITS 2008, Montreal, Canada, Proceedings* (Vol. Lecture Notes in Computer Science 5091, pp. 9-18): Springer.

D'Mello, S., Taylor, R. S., & Graesser, A. (2007). *Monitoring Affective Trajectories during Complex Learning*. Paper presented at the Proceedings of the 29th Annual Meeting of the Cognitive Science Society.

D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. *International Journal of Artificial Intelligence in Education, 16*(1), 3-28.

D'Mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction, 18*(1-2), 45-80.

Dai, J., Wu, M., Cohen, J., & Klawe, M. (2003). *Primeclimb: designing to facilitate mediated collaborative inquiry*. Paper presented at the CSCL 2003.

Darnon, C., Muller, D., Schrager, S. M., & Pannuzzo, N. (2006). Mastery and Performance Goals Predict Epistemic and Relational Conflict Regulation *Journal of educational psychology, 94*(4), 766-776.

de Vicente, A. (2003). *Towards Tutoring Systems That Detect Students' Motivation: An Investigation.* University of Edinburgh, Edinburgh.

de Vicente, A., & Pain, H. (2002). Informing the detection of the students' motivational state: an empirical study. In S. A. Cerri, G. Guy & F. Paraguacu (Eds.), *Intelligent Tutoring Systems. 6th International Conference, ITS2002, Biarritz, France and San Sebastian, Spain, Proceedings* (Vol. Lecture Notes in Computer Science 2363, pp. 933-943): Springer.

del Soldato, T., & du Boulay, B. (1995). Implementation of Motivational Tactics in Tutoring Systems. *International Journal of Artificial Intelligence in Education, 6*(4), 337-378.

Dennerlein, P., Becker, T., Johnson, P., Reynolds, C., & Picard, R. W. (2003). *Frustrating Computers Users Increases Exposure to Physical Factors*. Paper presented at the Proceedings of the International Ergonomics Association.

Devillers, L., & Vidrascu, L. (2006). *Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs*. Paper presented at the INTERSPEECH 2006 - ICSLP. Ninth International Conference on Spoken Language Processing.

Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure *Journal of Personality and Social Psychology, 36*(5), 451-462.

Diener, C. I., & Dweck, C. S. (1980). An analysis of learned helplessness: II. The processing of success. *Journal of Personality and Social Psychology, 39*(5), 940-952.

Dillenbourg, P. (2002). Over-scripting CSCL: The risks of blending collaborative learning with instructional design. *Three worlds of CSCL. Can we support CSCL*, 61-91.

Do-Lenh, S., Kaplan, F., & Dillenbourg, P. (2009). *Paper-based Concept Map: the Effects of Tabletop on an Expressive Collaborative Learning Task.* Paper presented at the The 23rd BCS conference on Human Computer Interaction (HCI 2009), Cambridge, UK.

du Boulay, B., & Luckin, R. (2001). Modelling human teaching tactics and strategies for tutoring systems. *International Journal of Artificial Intelligence in Education, 12*(3), 235-256.

du Boulay, B., Luckin, R., Martinez-Miron, E., Rebolledo-Mendez, G. R., & Harris, A. (2008). Motivationally Intelligent Educational Systems: Three Questions *Second International Conference on Innovations in Learning for the Future, Future e-Learning 2008* (Vol. 4793, pp. 1-10). Istanbul: University Rectorate Publication.

du Plessis, S. A. (1998). *A Conceptual Framework for the Development of Intelligent, Learning Style and Computer-based Educational Software for topics from Operations Research.* University of Stellenbosch.

Dweck, C. S. (1999a). Caution--Praise Can Be Dangerous. *American Educator, 23*(1), 4-9.

Dweck, C. S. (1999b). *Self-theories: Their role in personality, motivation, and development*. Philadelphia, PA: Psychology Press.

Dweck, C. S. (2002). Messages that motivate: How praise molds students' beliefs, motivation, and performance (in surprising ways). In J. M. Aronson (Ed.), *Improving academic achievement: impact of psychological factors on education* (pp. 37-60). New York: Acadmic Press.

Dweck, C. S., Chiu, C.-y., & Hong, Y.-y. (1995). Implicit Theories and Their Role in Judgments and Reactions: A Word From Two Perspectives *Psychological Inquiry, 6*(4), 267-285.

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological review, 95*(2), 256-273.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). Duchenne's smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology, 58*(2), 342-353.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., et al. (1987). Universals and cultural differences in the judgements of facial expressions of emotion. *Journal of Personality and Social Psychology, 53*(4), 712-717.

Elfenbein, H. A., & Ambady, N. (2003). Universals and Cultural Differences in Recognizing Emotions. *Current Directions in Psychological Science, 12*(5), 159-164.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol Analysis. Verbal Reports as Data*. Cambridge, MA: MIT Press.

Facer, K., Joiner, R., Stanton, D., Reidz, J., Hullz, R., & Kirk, D. (2004). Savannah: mobile gaming and learning? *Journal of Computer Assisted Learning, 20*, 399-409.

Feldman, B. L. (2004). Feelings or Words? Understanding the Content in Self-Report Ratings of Experienced Emotion. *Journal of Personality and Social Psychologyi, 87*(2), 266-281.

Festinger, L. (1957). *A Theory of Cognitive Dissonnace*. Stanford, CA: Stanford University Press.

Findlay, J. M., & Gilchrist, I. D. (2003). *Active vision: The psychology of looking and seeing*. Oxford: Oxford University Press.

Fletcher, R. R., Poh, M.-Z., & Eydgahi, H. (2010). *Wearable sensors: Opportunities and challenges for low-cost health care* Paper presented at the Annual International Conference of the Engineering in Medicine and Biology Society (EMBC).

Forbes-Riley, K., Rotaru, M., & Litman, D. J. (2008). The relative impact of student affect on performance models in a spoken dialogue tutoring system. *User Modeling and User-Adapted Interaction, 18*(1-2), 11-43.

Forgas, J. P. (2007). When sad is better than happy: Negative affect can improve the quality and effectiveness of persuasive messages and social influence strategies. *Journal of Experimental Social Psychology, 43*(4), 513-528.

Forgas, J. P. (2008). Affect and Cognition. *Perspectives on Psychological SCience, 3*(2), 94-101.

Gardner, H. (1983). *Frames Of Mind: The Theory Of Multiple Intelligences*. New York: Basic Books.

Goleman, D. (1995). *Emotional Intelligence: Why It Can Matter More Than IQ for Character, Health and Lifelong Achievement*. New York: Bantam Books.

Goleman, D. (2006a). *Emotional Intelliegence: Why it can matter more than IQ*. New York: Bantam.

Goleman, D. (2006b). *Social Intelligence: The New Science of Human Relationships*. London: Hutchinson.

Graesser, A., Chipman, P., King, B., McDaniel, B., & D'Mello, S. (2007). Emotions and Learning with AutoTutor. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (Vol. Frontiers in Artificial Intelligence and Applications 158, pp. 569-571). Amsterdam: IOS Press.

Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). *Detection of Emotions during Learning with AutoTutor* Paper presented at the Proceedings of the 28th Annual Conference of the Cognitive Science Society.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education, 48*(4), 612-618.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., et al. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments & Compurers, 36*(2), 180-192.

Graesser, A. C., & Olde, B. A. (2003). How Does One Know Whether a Person Understands a Device? The Quality of the Questions the Person Asks When the Device Breaks Down. *Journal of educational psychology, 95*(3), 524-536.

Graesser, A. C., Person, N. K., Harter, D., & TRG. (2001). Teaching Tactics and Dialog in AutoTutor. *International Journal of Artificial Intelligence and Education, 12*, 257-279.

Gray, E., K., & Watson, D. (2007). Assessing positive and negative affect via self-report. In J. Coan & J. Allen (Eds.), *Handbook of Emotion Ellicitation and Assessment* (pp. 171-183). New York: Oxford University Press.

Grimm, M., Mower, E., Kroschel, K., & Narayanan, S. (2006). *COMBINING CATEGORICAL AND PRIMITIVES-BASED EMOTION RECOGNITION*. Paper presented at the 14th European Signal Processing Conference (EUSIPCO 2006).

Gunes, H., & Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions, 1*(1), 68-99.

Habgood, J., Ainsworth, S., & Benford, S. (2005). Endogenous fantasy and learning in digital games. *Simulation & Gaming, 36*(4), 483.

Hancock, C., & Osterweil, S. (1996). Zoombinis and the Art of Mathematical Play. Retrieved July 7th 2010, 2010

Harris, A., Yuill, N., & Luckin, R. (2007). Creating Contexts for Productive Peer Collaboration: Some Design Considerations. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (pp. 391-398). Amsterdam: IOS.

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1994). *Emotional Contagion*. Cambridge: Cambridge University Press.

Heffernan, N., & Koedinger, K. R. (2002). An Intelligent Tutoring System Incorporating a Model of an Experienced Human Tutor. In S. A. Cerri, G. Guy & F. Paraguacu (Eds.), *Intelligent Tutoring Systems. 6th International Conference, ITS2002, Biarritz, France and San Sebastian, Spain, Proceedings* (Vol. Lecture Notes in Computer Science 2363, pp. 596-608). Berlin: Springer.

Heider, F. (1958). *The Psychology of Interpersonal Relations*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Higgins, M. C. (2001). When is Helping Helpful? Effects of Evaluation and Intervention Timing on Basketball Performance. *Journal of Applied Behavioral Science, 37*(3), 280-298.

Infante, C., Weitz, J., Reyes, T., Nussbaum, M., Gómez, F., & Radovic, D. (2009). Co-located collaborative learning video game with single display groupware. *Interactive Learning Environments*(1), 1--18.

Izard, C. E. (1991). *The Psychology of Emotions*. New York: Plenum.

Johns, J., & Woolf, B. P. (2006). *A Dynamic Mixture Model to Detect Student Motivation and Proficiency*. Paper presented at the Proceedings of the 21st national Conference on Articifial Intelligence (AAAI-06).

Johnson, D., & Wiles, J. (2003). Effective affective user interface design in games. *Ergonomics, 46, 13*(14), 1332-1345.

Joiner, R., Nethercott, J., Hull, R., & Reid, J. (2006). Designing educational experiences using ubiquitous technology. *Computers in Human Behavior, 22*(1), 67-76.

Jones, A., & Issroff, K. (2007). Learning Technologies. Affective and Social Issues. In G. Conole & M. Oliver (Eds.), *Contemporary Perspectives in E-Learning Research. Themes, Methods and Impact on Practice* (pp. 191-201). London and New York: Routledge.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies, 65*(8), 724-736.

Kapoor, A., & Picard, R. W. (2005). *Multimodal affect recognition in learning environments*. Paper presented at the Proceedings of the 13th annual ACM international conference on Multimedia

Kapoor, A., Qi, Y., & Picard, R. W. (2003). *Fully Automatic Upper Facial Action Recognition*. Paper presented at the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003).

Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition *Journal of educational psychology, 74*(6), 844-851.

Kenny, D., Kashy, D., & Cook, W. (2006). *Dyadic data analysis*. New York - London: Guilford Press.

Kerawalla, L., Pearce, D., Yuill, N., & Harris, A. (2008). “I’m keeping those there, are you?” The role of a new user interface paradigm – Separate Control of Shared Space (SCOSS) – in the collaborative decision-making process. *Computers & Education, 50*(1), 193-206.

Kiili, K. (2007). Foundation for problem-based gaming. *British Journal of Educational Technology, 38*(3), 394-404.

Kim, Y. (2007). Desirable Characteristics of Learning Companions. *International Journal of Artificial Intelligence in Education, 17*(4), 371-388.

Kirriemuir, J., & McFarlane, A. (2004). *Literature Review in Games and learning*: Futurelab.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual Fixation Patterns During Viewing of Naturalistic Social Situations as Predictors of Social Competence in Individuals With Autism *Archives of General Psychiatry, 59*(9), 809-816.

Knez, I., & Niedenthal, S. (2008). Lighting in digital game worlds: effects on affect and play performance. *CyberPsychology & Behavior, 11*(2), 129-137.

Kollar, I., Fischer, F., & Hesse, F. W. (2006). Collaboration scripts–a conceptual analysis. *Educational Psychology Review, 18*(2), 159-185.

Lahaderne, H. M. (1968). Attitudinal and intellectual correlates of attention: A study of four sixth-grade classrooms. *Journal of educational psychology, 59*(5), 320-324.

Landauer, T. K., & T.Dumais, S. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Pscychological Review, 104*(2), 211-240.

Larsen, J. T., McGraw, A. P., Mellers, B. A., & Cacioppo, J. T. (2004). The Agony of Victory and Thrill of Defeat Mixed Emotional Reactions to Disappointing Wins and Relieving Losses. *Psychological Science, 15*(5), 325-330.

Lee, S. W., Kelly, K. E., & Nyre, J. E. (1999). Preliminary Report on the Relation of Students' On-Task Behavior With Completion of School Work. *Psychological Reports, 84*(1), 267-272.

Lehman, B., Matthews, M., D'Mello, S., & Person, N. (2008). What are you feeling? Investigating Student Affective States During Expert Human Tutoring Sessions. In B. P. Woolf, E. Aïmeur, R. Nkambou & S. L. Lajoie (Eds.), *Intelligent Tutoring Systems, 9th International Conference, ITS 2008, Montreal, Canada, Proceedings* (Vol. Lecture Notes in Computer Science 5091, pp. 50-59): Springer.

Lepper, M. R., & Hodell, M. (1989). intrinsic motivation in the classroom. In C. Ames & R. E. Ames (Eds.), *Research on motivation in education* (Vol. 3, pp. 73-105). New York: Academic Press.

Lepper, M. R., & Woolverton, M. (2002). The wisdom of practice: Lessons learned from the study of highly effective tutors. . In J. Aronson (Ed.), *Improving Academic Achievement: Impact of Psychological Factors on Education* (pp. 135-158). New York: Academic Press.

Lepper, M. R., Woolverton, M., Mumme, D. L., & Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In S. Lajoie & S. Derry (Eds.), *Computers as Cognitive Tools* (pp. 75-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Light, P., & Littleton, K. (1999). Introduction: Getting IT together. In K. Littleton & P. Light (Eds.), *Learning with computers: analysing productive interaction* (pp. 1-9). London: Routeledge.

Lindström, M., Ståhl, A., Höök, K., Sundström, P., Laaksolathi, J., Combetto, M., et al. (2006). *Affective diary: designing for bodily expressiveness and self-reflection*. Paper presented at the SIGCHI '06 extended abstracts on Human factors in computing systems.

Litman, D. J., & Forbes-Riley, K. (2004). *Predicting student emotions in computer-human tutoring dialogues*. Paper presented at the ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics

Lloyd, J. W., & Loper, A. B. (1986). Measurement and evaluation of task-related learning behaviors: Attention to task and metacognition. *School Psychology Review, 15*(3), 336-345.

Luckin, R., & Hammerton, L. (2002). Getting to know me: helping learners understand their own learning needs through metacognitive scaffolding *Intelligent Tutoring Systems. 6th International Conference, ITS2002, Biarritz, France and San Sebastian, Spain, Proceedings* (Vol. Lecture Notes in Computer Science 2363, pp. 759-771): Springer.

Maldonado, H., Lee, J.-E. R., Brave, S., Nass, C., Nakajima, H., Yamada, R., et al. (2005). *We learn better together: enhancing eLearning with emotional characters*. Paper presented at the Conference on Computer Supported Collaborative learning 2005: the next 10 years!

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 5*(4), 333-369.

Margolis, J. L., Nussbaum, M., Rodriguez, P., & Rosas, R. (2006). Methodology for evaluating a novel education technology: a case study of handheld video games in Chile. *Computers & Education, 46*(2), 174-191.

Marks, G., & Miller, N. (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin, 102*(1), 72-90.

Masthoff, J., & Gatt, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: affective state in group recommender systems *User Modeling and User-Adapted Interaction, 16*(3-4), 281-319.

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and autonomic physiology *Emotion, 5*(2), 175-190.

Mavrikis, M. (2008). *Modelling Students' Behaviour and Affective States in ILEs through Educational Data Mining.* Unpublished PhD Thesis, University of Edinburgh.

Mavrikis, M., D'Mello, S., Porayska-Pomsta, K., Cocea, M., & Graesser, A. (2010). Modeling Affect by Mining Students' Interactions within Learning Environments, . In C. Romero, S. Ventura, M. Pechenizkiy & R. S. J. d. Baker (Eds.), *Handbook of Educational Data Mining*: Chapman & Hall/CRC.

Mavrikis, M., Maciocia, A., & Lee, J. (2007). Towards Predictive Modelling of Student Affect from Web-Based Interactions. In R. Luckin, K. R. Koedinger & J. Greer

(Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (Vol. Frontiers in Artificial Intelligence and Applications 158, pp. 169-176). Amsterdam: IOS Press.

McDaniel, B., D'Mello, S., King, B., Chipman, P., Tapp, K., & Graesser, A. (2007). *Facial Features for Affective State Detection in Learning Environments*. Paper presented at the Proceedings of the 29th Annual Meeting of the Cognitive Science Society.

Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). Manual for the Patterns of Adaptive Learning Scales (PALS). In U. o. Michigan (Ed.). Ann Arbor.

Mitrovic, A., & Martin, B. (2002). Evaluating the Effects of Open Student Models on Learning. In P. De Bra, P. Brusilovsky & R. Conejo (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems: Second International Conference, AH 2002 Proceedings* (pp. 296-305). Berlin: Springer-Veralg.

Morishima, Y., Nakajima, H., Brave, S., Yamada, R., Maldonado, H., Nass, C., et al. (2004). The role of affect and sociality in the agent-based collaborative learning system *Affective Dialogue Systems* (pp. 265-257). Berlin/Heidelberg: Springer.

Mota, S., & Picard, R. W. (2003). *Automated Posture Analysis for Detecting Learner's Interest Level*. Paper presented at the Conference on Computer Vision and Pattern Recognition Workshop.

Munneke, L., Andriessen, J., Kanselaar, G., & Kirschner, P. (2007). Supporting interactive argumentation: Influence of representational tools on discussing a wicked problem. *Computers in Human Behavior, 23*, 1072-1088.

Murray, T., & Arroyo, I. (2002). Toward Measuring and Maintaining the Zone of Proximal Development in Adaptive Instructional Systems. In S. A. Cerri, G. Gouarderes & F. Paraguacu (Eds.), *Intelligent Tutoring Systems, 6th Internatiuonal Conference, ITS 2002, Proceedings* (Vol. Lecture Notes in Computer Science 2363). Berlin: Springer.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*, 669-678.

Nezlek, J. B., Vansteelandt, K., Van Mechelen, I., & Kupens, P. (2008). Appraisal-Emotion relationships in daily life. *Emotion, 8*(1), 145-150.

Norman, D. A. (2004). Emotional Design: Why we love (or hate) everyday things: New York: Basic Books.

Norman, D. A., & Ortony, A. (2006). Designers and users: Two perspectives on emotion and design. In S. Bagnara & G. Crampton Smith (Eds.), *Theories and practice in interaction design* (pp. 91 - 106). London: Lawrence Erlbaum Associates.

Novak, J. D., & Cañas, A. J. (2008). *The theory underlying concept maps and how to construct and use them*. Pensacola Fl, : Florida Institute for Human and Machine Cognition

O'Bryen, P. (1996). *Using questionnaires to assess motivation in second language classrooms*: University of Hawaii.

Op ’t Eynde, P., & Turner, J. E. (2006). Focusing on the Complexity of Emotion Issues in Academic Learning: A Dynamical Component Systems Approach. *Educational Psychology Review, 18*(4), 361-376.

Padilla, N., Gonzalez, J., Gutierrez, F. L., Cabrera, M. J., & Paderewski, P. (2009). Design of Educational Multiplayer Videogames. A Vision From Collaborative Learning. *Advances in Engineering Software, 40*(12).

Pavlidis, I., Levine, J., & Baukol, P. (2001). *Thermal image analysis for anxiety detection*. Paper presented at the International Conference on Image Processing, 2001. .

Pekrun, R. (2006). The Control-Value Theory of Achievement Emotions: Assumptions, Corollaries, and Implications for Educational Research and Practice. *Educational Psychology Review, 18*(4), 315-341.

Pekrun, R., Goetz, T., Perry, R., & Titz, W. (2002). Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research. *Educational psychologist, 37*(2), 91 - 105.

Picard, R. W. (2010). Emotion Research by the People, for the People. *Emotion Review, 2*(3), 250-254.

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2011). Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam *IEEE Transactions on Biomedical Engineering,, 58*(1), 7-11.

Porayska-Pomsta, K., Bernardini, S., & Rajendran, G. (2009). *Embodiment as a means for Scaffolding Young Children’s Social Skill Acquisition*. Paper presented at the 8th International Conference on Interaction Design and Children.

Porayska-Pomsta, K., Frauenberger, C., Pain, H., Rajendran, G., Smith, T., Menzies, R., et al. (2011). Developing technology for autism: an interdisciplinary approach *International Journal of Personal and Ubiquitous Computing*.

Porayska-Pomsta, K., Mavrikis, M., & Pain, H. (2008). Diagnosing and acting on student affect: the tutor’s perspective. *User Modeling and User-Adapted Interaction, 18*(1-2), 125-173.

Porayska-Pomsta, K., & Pain, H. (2004). *Exploring Methodologies for Building Socially and Emotionally Intelligent Learning Environments*. Paper presented at the Workshop on Social and Emotional Intelligence in Learning Environments (SEILE), at ITS2004.

Prendinger, H., Mayer, S., Mori, J., & Ishizuka, M. (2003). *Persona effect revisited. Using bio-signals to measure and reflect the impact of character-based interfaces*. Paper presented at the IVA 2003 : Intelligent Virtual Agents

Read, J., MacFarlane, S., & Casey, C. (2002). *Endurability, Engagement and Expectations: Measuring Children’s Fun*. Paper presented at the International Workshop on Interaction Design and Children.

Read, J. C., & MacFarlane, S. (2006). *Using the fun toolkit and other survey methods to gather opinions in child computer interaction*. Paper presented at the IDC '06 Proceedings of the 2006 conference on Interaction design and children

Reidsma, D., Hofs, D. H. W., & Jovanovic, N. (2005). *A presentation of a set of new annotation tools based on the NXT API*. Paper presented at the Measuring Behaviour 2005.

Reisenzein, R., & Hofmann, T. (1993). Discriminating Emotions from Appraisal-relevant Situational Information: Baseline Data for Structural Models of Cognitive Appraisals. *Cognition and Emotion, 7*(3/4), 271-293.

Robertson, J., Good, J., & Pain, H. (1998). BetterBlether: The Design and Evaluation of a Discussion Tool for education. *International Journal of Artificial Intelligence and Education, 9*(3-4), 219-236.

Robson, C. (1993). *Real World Research: A resource for social scientists and practitioner-researchers*. Oxford: Blackwell.

Rodrigo, M. M. T., Baker, R. S. J. d., D'Mello, S., Gonzalez, M. C. T., Lagud, M. C. V., Lim, S. A. L., et al. (2008). Comparing Learners' Affect While Using an Intelligent Tutoring Systems and a Simulation Problem Solving Game *Intelligent Tutoring Systems, 9th International Conference, ITS 2008, Montreal, Canada, Proceedings* (Vol. Lecture Notes in Computer Science 5091, pp. 40-49): Springer.

Rodrigo, M. M. T., Baker, R. S. J. d., Lagud, M. C. V., Lim, S. A. L., Macapanpan, A. F., Pascua, S. A. M. S., et al. (2007). Affect and Usage Choices in Simulation Problem-Solving Environments. In R. Luckin, K. R. Koedinger & J. Greer (Eds.), *Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work* (Vol. Frontiers in Artificial Intelligence and Applications 158). Amsterdam: IOS Press.

Rogers, C. (1961). *On Becoming a Person*. London: Constable.

Rosas, R., Nussbaum, M., Cumsille, P., Marianov, V., Correa, M., Flores, P., et al. (2003). Beyond Nintendo: design and assessment of educational video games for first and second grade students. *Computers & Education, 40*(1), 71-94.

Roschelle, J., & Teasley, S. D. (1995). The Construction of Shared Knowledge in Collaborative Problem Solving. In C. O'Malley (Ed.), *Computer-Supported Collaborative Learning* (pp. 69-97). Berlin: Springer.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of Behavioral Research: Methods and Data Analysis*. New York: McGraw-Hill.

Rosiek, J. (2003). Emotional Scaffolding: An Exploration of the Teacher Knowledge at the Intersection of Student Emotion and the Subject Matter. *Journal of Teacher Education, 54*(4), 399-412.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569-600.

Russell, J. A. (1994). Is There Universal Recognition of Emotion From Facial expression? A Review of the Cross-Cultural Studies. *Psychological Bulletin, 115*(1), 102-141.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review, 110*(1), 145-172.

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-Item Scale of Pleasure and Arousal. *Journal of Personality and Social Psychology, 57*(3), 493-502.

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, Massachusetts: MIT press.

Sayette, M. A., Cohn, J. F., Wertz, J. M., Perrott, M. A., & Parrott, D. J. (2001). A psychometric evaluation of facial action coding system for assessing spontaneous expression *Journal of Nonverbal Behavior, 25*(3), 167-185.

Sayfan, L., & Lagattuta, K. H. (2008). Grownups Are Not Afraid of Scary Stuff, but Kids Are: Young Children's and Adults' Reasoning about Children's, Infants', and Adults' Fears. *Child Development, 79*(4), 821-835.

Scherer, K. R. (2005). What are emotions? And how can they be measured? . *Social Science Information, 44*(4), 695-729.

Schwartz, D. (1998). The productive agency that drives collaborative learning. In P. Dillenbourg (Ed.), *Collaborative learning: Cognitive and computational approaches* (pp. 197-218). New York: Elsevier Science/Pergamon.

Scott, S., Mandryk, R., & Inkpen. (2003). Understanding Children's Collaborative Interactions in Shared Environments. *Journal of Computer Assisted Learning*(19), 220-228.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental & Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.

Shafran, I., Riley, M., & Mohri, M. (2003). *Voice Signatures*. Paper presented at the Proceedings IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

Sharp, H., Rogers, Y., & Preece, J. (2007). *Interaction design: Beyond Human-Computer Interaction*. West Sussex: John Wiley & Sons.

Smith, C. A., & Ellsworth, P. (1985). Patterns of cognitive appraisal in emotions. *Journal of Personal and Social Psychology, 84*(4), 813-838.

Stein, N. L., & Levine, L. J. (1991). Making sense out of emotion. In W. Kessen, A. Ortony & F. Craik (Eds.), *Memories, Thoughts, and Emotions: Essays in Honor of George Mandler* (pp. 295-322). Hillsdale, NJ: Lawrence Erlbaum Associates.

Storey, J. K., Kopp, K. J., Wiemer, K., Chipman, P., & Graesser, A. C. (in press). Critical thinking tutor: Using AutoTutor to teach scientific critical thinking skills. *Behavioral Research Methods*.

Susaeta, H., Jimenez, F., Nussbaum, M., Gajardo, I., Andreu, J., & Villalta, M. (in press). From MMORPG to a Classroom Multiplayer Presential Role Playing Game. *Educational Technology & Society*.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of educational psychology, 82*(2), 306-314.

Tobias, S. (1995). *Overcoming math anxiety (revised and expanded)* New York: W.W. Norton & Company.

Tong, E. M. W., Bishop, G. D., Enkelmann, H. C., Why, P. Y., & Diong, M. S. (2007). Emotion and appraisal: A study using ecological momentary assessment. *Cognition and Emotion, 21*(7), 1361 - 1381.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When Are Tutorial Dialogues More Effective Than Reading? . *Cognitive Science, 31*(1), 2-62.

Wages, R., Grünvogel, S. M., & Grützmacher, B. (2004). How Realistic is Realism? Considerations on the Aesthetics of Computer Games *Entertainment Computing – ICEC 2004* (pp. 83-92).

Walonoski, J. A., & Heffernan, N. T. (2006). Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems. In M. Ikeda, K. D. Ashley & T.-W. Chan (Eds.), *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, Proceedings* (Vol. Lecture Notes in Computer Science 4053, pp. 382-391): Springer.

Wegerif, R., Mercer, N., & Lyn, D. (1999). From social interaction to individual reasoning: an empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction, 9*(6), 493-516.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36-45.

Wentzel, K. R. (1997). Student Motivation in Middle School: The Role of Perceived Pedagogical Caring. *Journal of educational psychology, 89*(3), 411-419.

Whitelock, D., & Scanlon, E. (1996). *Motivation, media and motion Reviewing a computer supported collaborative learning experience*. Paper presented at the European Conference on Artificial Intelligence in Education.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*, 459-482.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(1), 39-58.