Knowledge Elicitation Methods for Affect Modelling in Education

Kaśka Porayska-Pomsta, **Manolis Mavrikis**
(London Knowledge Lab, Institute of Education, University of London; K.Porayska-Pomsta@ioe.ac.uk, M.Mavrikis@ioe.ac.uk;
Sidney D'Mello
(Departments of Psychology and Computer Science, University of Notre Dame; sdmello@nd.edu);
Cristina Conati
(Department of Computer Science, University of British Columbia; conati@cs.ubc.ca);
Ryan S.J.d. Baker
(Teachers College, Columbia University; ryan@educationaldatamining.org)

Abstract. Research on the relationship between affect and cognition in Artificial Intelligence in Education (AIEd) brings an important dimension to our understanding of how learning occurs and how it can be facilitated. Emotions are crucial to learning, but their nature, the conditions under which they occur, and their exact impact on learning for different learners in diverse contexts still needs to be mapped out. The study of affect during learning can be challenging, because emotions are subjective, fleeting phenomena that are often difficult for learners to report accurately and for observers to perceive reliably. Context forms an integral part of learners' affect and the study thereof. This review provides a synthesis of the current knowledge elicitation methods that are used to aid the study of learners' affect and to inform the design of intelligent technologies for learning. Advantages and disadvantages of the specific methods are discussed along with their respective potential for enhancing research in this area, and issues related to the interpretation of data that emerges as the result of their use. References to related research are also provided together with illustrative examples of where the individual methods have been used in the past. Therefore, this review is intended as a resource for methodological decision making for those who want to study emotions and their antecedents in AIEd contexts, i.e. where the aim is to inform the design and implementation of an intelligent learning environment or to evaluate its use and educational efficacy.

Keywords: Research Methods, Affect, Learning, Knowledge Elicitation, Intelligent Learning Environments, Annotating Affect

1. Introduction

There is little doubt that affect and cognition are related, and that emotions can enhance or inhibit learning. As early as 1908, Yerkes and Dodson posited that an optimal level of emotional arousal, i.e. one that is neither too high nor too low, is necessary to enabling effective, long-term learning, a hypothesis seen in more recent theoretical accounts of cognition and emotion as well (Csikszentmihalyi, 1990; Shernoff et al., 2003; Barth and Funke, 2010; Clore and Huntsinger, 2007; Isen, 2008; Schwartz, *in press*). Emotions along with moods and related motivational states form a natural and arguably essential component of learning, but their nature, the conditions under which they occur, and their specific impact on learning for different learners still remain to be fully explored. The study of affect and cognition presents many challenges to those who attempt it, especially if the end goal is to inform the design of a new generation of intelligent technologies that are able to adapt their learning support to the individual learners' differences and needs in diverse situations in real-

time. This is because engineering such technologies requires access to relatively precise specification of affect-related and other relevant information of importance to both learning in general and to learning by individuals in specific domains and contexts. While many possible approaches emerge from psychology, cognitive science and affective computing, the field still lacks a cohesive account of the emotions that are relevant to learning and principled guidelines for how to identify and interpret learners' affect in context.

The question of what can be referred to as *affect* and what needs to be measured remain the objects of active research, to which the present review aims to contribute. Despite the lack of clear answers, much of the current affect modelling research seems to rest on the assumption that "true" affect can be both observed and defined objectively. This assumption is propelled by a combination of the engineering demand for precision in the definition of constructs on which various technological designs are based and of the growing availability and real-time feasibility of physiological (e.g., electrodermal response) and bodily (e.g., facial features, speech contours) sensors as means for detecting human emotions. Reflected in the growing number of extensive reviews of methods for affect detection that rely on such sensors (Calvo & D'Mello, 2010; Jaimes & Sebe, 2007; Pantic & Patras, 2006; Pantic & Rothkrantz, 2003; Valstar et al., in press; Zeng et al., 2009; D'Mello and Kory, 2012) is an ongoing debate about which approaches are best for capturing "true" affect, with instruments for measuring bodily and physiological sensors being seen increasingly as the appropriate tools for the job.

However, there exists a range of other, equally important methods for studying emotional and affective responses, which do not rely on bodily and physiological sensors, and which have not been the subject of a recent comprehensive review. The present review aims to bridge this gap by discussing knowledge elicitation methods that can be used with different stakeholders, including learners, teachers and external observers in educational interactions, to enable insight into observations and interpretations by those stakeholders of their own and of other people's emotional experiences that relate to learning. The value of such interpretations is not only in the way they can inform the study of what emotions might be important to learning, but also in their being essential to a formulation of a theory of how, when and why emotions occur and of their real or perceived impact on the way we communicate and learn (Wosnitza and Volet, 2005; Järvenoja and Järverlä, 2005). Therefore, the present review rests on the understanding that gaining access to such interpretations is fundamental to our being able to explain and model computationally learners' emotional reactions in a way that contributes to both pedagogical practices in general and to the design and use of intelligent learning environments in particular.

Despite their undisputable potential and growing usefulness, bodily and physiological sensors are not considered in the present review as the primary basis for evaluating learner's affect, because automated affect detectors derived from these sensors inherently depend on some measure of *ground truth* for their development – a view also expressed by other researchers, e.g. by Afzal and Robinson (2011). In turn, such ground truth is derived from the stakeholders' interpretations obtained through the knowledge elicitation methods discussed in this paper, which are treated here as primary to physiological sensors research. Furthermore, the present focus is on annotations of affect by humans and not through physiological sensors, the latter of which have been extensively reviewed elsewhere (Calvo & D'Mello, 2010; Jaimes & Sebe, 2007; Pantic & Patras, 2006; Pantic & Rothkrantz, 2003; Valstar et al., in press; Zeng et al., 2009; D'Mello and Kory, 2012).

One of the major attractions of using technology to study affect in relation to learning is the fact that it supports automatic logging and systematic evaluation of models. It also facilitates the construction of real-time dynamic models of affect in educational interactions (e.g. Baker et al., 2012; Conati & Zhou, 2002; D'Mello et al., 2008; Sabourin et al., 2011; Woolf et al., 2009). Data obtained through the use of the methods discussed in this review is fundamental to endowing a technology-enhanced learning (TEL) environment with an ability to interpret

learners' behaviours and to act on them in a contingent manner and in real-time. In turn, such an interpretative ability of a TEL environment (also known as user or learner modelling) constitutes a cardinal pre-requisite of intelligent learning technologies, in the AIEd sense, i.e. technologies that are motivated by and built based on Artificial Intelligence techniques and paradigms (see e.g. Russell and Norvig, 1995; Woolf, 2010). Such technologies can further adhere to other technology-enhanced learning approaches, such as Intelligent Tutoring Systems (these traditionally focus on supporting one-on-one learning in a specific, often well-defined subject domain), Intelligent Learning Environments (these typically embrace the idea that learners do not represent a homogenous population, that learning is life-long, and that the environment as a whole influences what, when and how is learned), or Educational Games (these leverage the affordances, e.g. intrinsic motivation associated with competition and winning, and structure, e.g. levels and rewards, of commercial computer games, for educational purposes). In this review, so long as a technology-enhanced learning approach is endowed with, or aspires to be endowed with diagnostic, inferential and /or adaptive capabilities, it is treated as an intelligent technology for learning in the AIEd sense.

The chief focus of the present review is on approaches and instruments for eliciting knowledge (objective and subjective) about which emotions are important to learning and to tailoring educational interactions, and on how to measure and label those emotions in real educational interactions reliably, based on situated input from educators, learners and external observers. The objective is to provide a resource and a map for methodological decision making for those who want to study emotions and their relationships to other theoretically important constructs, such as learning, engagement, attitudes, goals, etc., in AIEd and other technology-enhanced learning contexts.

1.1. Relevant reviews

As well as building on the authors' combined first-hand experience in using the specific methods, the present review extends two similar accounts available to date in the context of learning technologies use and/or design. Relevant details contained in these accounts are brought to the reader's attention in the paper as and when relevant, however a brief summary of their main contributions is given immediately, in order to outline their relationship to the present review, and to exhibit this review's novel contribution.

Many of the knowledge elicitation methods discussed in earlier reviews have been used to study affect before the advent of technology and affective computing (Coan & Allen, 2007), but technology makes it possible to use the established methods in new ways, e.g., through graphical user interfaces and often in combination with data synchronisation, data mining, machine learning and physiological and behavioural sensing.

Wosnitza and Volet (2005) review knowledge elicitation and interaction analysis methods that can be used to study learners' affect during social online learning. In particular, they propose that methods for accessing emotions of relevance to learning belong to one of three temporal categories: (i) snapshot measurements (immediately before and/or after learning), (ii) continuous measurements (during learning) and (iii) stimulated recall measurements (after learning), and they highlight how the different instruments (e.g. questionnaires, interviews, verbal protocols, etc.) when they are applied at different times can help in identifying the origin (the cause) and direction (the subject or object) of learners' emotions in general, and in online social learning contexts in particular. The specific instruments are discussed along with their relative strengths and weaknesses, which are illustrated by reference to several empirical studies. Therefore, Wosnitza and Volet's review examines a range of knowledge elicitation methods in terms of *when* they may be applied in order to elicit *why* learners experience certain emotions and *towards what* they direct those emotions during learning.

Afzal and Robinson (2011), extend Wosnitza's and Volet's proposal by considering different

qualitative and quantitative approaches linking them directly to their goal to enable automatic, real-time multimodal detection of emotions in learning contexts. They are careful to highlight the kind of interpretations that the knowledge elicitation methods considered facilitate. To Wosnitza and Volet's temporal categorisation of the methods, they add further dimensions of objectivity and subjectivity, thereby implicitly pointing towards the role of the informants and the resulting type of data (qualitative vs. quantitative and subjective vs. objective), in the knowledge elicitation process. Therefore, Afzal and Robinson's review maps out the knowledge elicitation instruments in terms of *when* they can be applied (as per Wosnitza and Volet's proposal) and in terms of the *type of data* that each yields.

While Wosnitza and Volet's review does not discuss the different measures in detail, Afzal and Robinson's review focuses chiefly on automated detection of affect. The present review both compliments and extends these two methodological accounts by examining the different measures along three dimensions:

- *what* instruments are available to elicit information about a learner's affective states e.g., tools to elicit *forced-choice vs. free responses*
- *who* generates the emotion reports, i.e. the learners, trained coders, or the tutors
- *when* is the elicitation of affective knowledge undertaken, i.e., *concurrent* to the interaction or *retrospectively*.

While these dimensions are in line with the two reviews, the 'what' and the 'who' dimensions in particular are treated here at a much finer grain level of detail than previously available, with specific case studies illustrating the particular applications of the individual methods at different times (the 'when' dimension). Most of the methods to study learner's affect described in the following sections rely on a combination of instruments. Advantages and disadvantages of the different methods and instruments are discussed.

1.2 Outline of the review

The structure of the review is as follows. Section 2 highlights three preliminary considerations that need to be taken into account to enable researchers to make informed choices of the different methods. Section 3 examines the different types of tools used to elicit information about learners' emotions. Section 4 focuses on the different types of informants, placing them on a spectrum of proximity to the emotional experiences studied. These include learners who have the most intimate insight into their own emotional experiences, though not always possess the means to verbalise those experiences, tutors who are internal to the tutoring situations in which the emotions are detected and labeled, and observers who are external to the tutoring situations. Section 5, considers different times at which the detection and labelling of learners' emotions may be undertaken and presents different variants on the protocols that have been followed to date. In all the sections advantages and disadvantages of the different techniques and protocols are considered. Section 6, summarises the review's main aims and focus, and offers the reader a glimpse beyond the knowledge elicitation methods discussed therein towards some initial challenges related to the analysis of data. In section 7 three tables are provided as a quick look-up reference to and a summary of the methods reviewed along with their advantages and disadvantages.

2. Preliminary considerations

Prior to embarking on a study of affect in learning, a number of important questions need to be considered when deciding on the appropriate methods and instruments. These questions relate to: (i) what is being measured, (ii) the predominant purpose of the data being sought and (iii) the expected fidelity of the resulting data given the different methods and instruments. These questions are pertinent to any domain of research relying on empirical

evidence, but they also reflect many debates which are specific to the research related to affect in learning, where the dependency between what is being measured (i.e. emotions) and immediate context (settings, actors, domain of interaction, etc.) in which the measurements are conducted may preclude an exact replication of the conditions. It is important to consider the various points at the studies' preparatory stages, including during piloting, to ensure coherence and rigour of the evidence generated.

2.1 What is being measured?

Before answering the question of *which instruments* can be used to measure affect, it is prudent first to consider *what is to be measured*. Affective phenomena range from persistent emotional predispositions (e.g., hostility), prolonged mood states (e.g., depression), intermediate-length mood states (e.g., being downhearted because of bad weather), and transient emotions (e.g., frustration from being stuck), to rapid reflexive responses (e.g., startled) (Barrett, Mesquita, Ochsner, & Gross, 2007; Ekman, 1992; Izard, 2010; Rosenberg, 1998; Russell, 2003). Clearly, the temporal granularity of the affective phenomena must play a role in the selection of the appropriate instrument to monitor affect.

One broad temporal organisation of affective phenomena involves distinguishing between affective traits, background moods, and emotions (Rosenberg, 1998). Affective traits are relatively stable, mostly unconscious predispositions towards particular emotional experiences. They operate by lowering the threshold for experiencing certain emotional states (i.e., hostile people have a lower threshold for experiencing anger but not necessarily other negative emotions). Moods also perform a threshold reduction function on emotional elicitation in that they make it easier for emotions that are congruent with the mood to be activated (e.g., feeling sad when one is in a general negative mood). However, moods are considered to be more transitory than affective traits and have a background influence on consciousness. In contrast to affective traits and moods, emotions are brief, intense states that occupy the forefront of consciousness, have significant physiological and behavioural manifestations, and rapidly prepare the bodily systems for action (Ekman, 1992; Izard, 2010). According to this framework, affective traits, moods, and emotions occupy varying positions along the dimensions of *duration*, *pervasiveness in consciousness*, and *distributive breadth* (i.e., the extent of the influence each have on other psychological and physiological processes). The present review focuses on how to measure and label the immediate emotions (henceforth referred to as *affective states* or *emotions*) that arise in a given learning situation instead of longer-term moods and stable affective traits. It should be noted that the present focus on immediate emotions raises some important issues when it comes to analysing the context surrounding an emotional expression. Ekman hypothesised that emotions are intense events that last for approximately 0.5 – 4 seconds (Ekman, 1984), but there is a notable paucity of research when it comes to understanding the temporal dynamics of the emotions. Therefore, one should analyse the learning context at a minimum of 5 seconds before the emotional episode, with several researchers focusing on larger windows that span 10-20 seconds (D'Mello et al., 2008).

Another important consideration in determining what to measure, relates to identifying the specific affective states to annotate. Pekrun and Stephens (2012) provide a taxonomy of the affective states that occur in educational contexts, which groups these so-called *academic emotions* into the four categories: (i) achievement emotions, (ii) topic emotions, (iii) social emotions, and (iv) epistemic emotions. *Achievement emotions* (e.g., contentment, anxiety, and frustration) are linked to learning activities (e.g., homework, taking a test) and outcomes (e.g., success/ failure). *Topic emotions* are aligned with the learning topic (e.g., empathy for a protagonist while reading classic literature). On the other hand, *social emotions* such as pride, shame, and jealousy are not directly related to the topic but reflect the fact that educational activities are socially situated. Finally *epistemic emotions* arise primarily from cognitive

information processing, such as surprise when novelty is encountered, or confusion when the student experiences an impasse. This taxonomy of academic emotions posits a large set of affective states that are presumably relevant in a diverse set of educational contexts, such as attending lectures, completing homework, taking tests. This taxonomy provides a useful initial list of states to annotate in a given learning context. An important goal for AIEd is to facilitate the identification of the correspondences between these different emotion categories and the specific contexts in which they tend to occur. For example, social emotions are less relevant in one-on-one student-computer interactions than in student-student interactions.

2.2 Reliance on labels

The methods discussed in this paper and the studies that rely on them typically use *labels* to characterise affect. Such labels can refer to binary categories denoting the presence of specific affective states (e.g., confusion = 1, frustration = 0), categories with intensities (confusion = 0.9, frustration = 0.3), dimensions with intensities (valence = 0.8, arousal = 0.4), or dimensions grouped into categories along with their values (positive deactivating states, where valence is high and arousal is low). The advantage of labels, either categorical or dimensional, is that they offer a language with which to characterise the affective phenomena under consideration. Such language is important for: (a) aligning observations with emotion theories, (b) interpreting behavioural signals (e.g., noting that the furrowed brow co-occurs with annotations of confusion), (c) human annotation of affect, because humans use language, and (d) deriving strategies to respond to affect (e.g., the learner is confused so a hint might be appropriate). However, labels have a liability because language is essentially imprecise and fluid, instead of being rigid and discrete, and the meaning of a label is constrained by context and linguistic repertoire of the labeller, ultimately residing in the mind of the interpreter (see also Ortony et al., 1988, for a discussion of related issues).

An alternative approach is to do away with labels entirely, and proceed with identifiable stable configurations in the data. For example, one could log interaction events, cluster these events, identify stable clusters, and correlate these clusters with performance. This approach can be applied with any data stream (physiological, behavioural, etc.). The advantage of this purely data-driven approach is that it avoids the problem of interpretation consistency introduced by labels. The obvious trade-off is that one loses the advantages provided by labels as discussed above. It should also be noted that the study of affect and learning is not only to service engineering goals of building more effective learning environments. There is the equally important scientific goal of studying affect to advance basic understanding of its relationship with learning. Labels facilitate this goal, because psychological and educational theories use labels to identify constructs. Nevertheless, the debate of the role of labels in science has been discussed for centuries, no consensus has been reached, it is unlikely that this issue will ever be fully settled, and it is beyond the scope of this paper to attempt to settle it. Therefore, the review proceeds with a description of methodologies and studies that use labels to measure affect while being mindful of the potential pitfalls introduced by their use.

2.3 Validity and Reliability of Measures

In any scientific domain, the use of instruments for measuring phenomena of interest raises the question of whether the data gathered through them and the associated conclusions can be trusted and replicated. This is of particular importance in a domain such as the study of human affect in learning, because, as discussed thus far, both affect and any description or interpretation thereof is prone to subjectivity and is context-dependent. Studying human affect in relation to learning imposes additional demands on the instruments used, because the conclusions drawn with their help may have an impact on educational practice and life-long outcomes for individuals.

Validity of the affect measurements refers to whether the data obtained reflect the *true* phenomena in *representative* contexts and thus, whether the data generalise beyond the specific context in which they have been generated. Assuming the equivalence of conditions under which the measurements are conducted through different means, **reliability** of affect measurements accounts for whether the same data can be generated through other means.

Validity in affect measurement is critical, because similar to most psychological variables, affect cannot be measured directly and one can only approximate its true value. This approximation raises critical questions in the measurement of human emotions (Rosenthal & Rosnow, 1984) that relate to:

- *Conclusion validity*, i.e. the ability to infer a relationship between any two variables of interest – e.g. are increased levels of happiness related to an increase in learning gains? (Shadish, Cook & Campbell, 2002),
- *Internal validity*, i.e. whether a relationship between two variables is causal – e.g. does happiness cause positive learning gains? (Campbell & Stanley, 1963),
- *Construct validity*, i.e. whether the operational definitions of a construct accurately reflect that construct – in other words: are we measuring what we are claiming to be measuring? (Campbell & Fiske, 1959),
- *External validity*, i.e. the extent that any relationship observed in the lab settings can generalise to other people, places, and times (Shadish, et al., 2002),
- *Ecological validity*, i.e. the extent to which the environment (including settings, learners and tutors) within which the observations are made is truly representative of the environment that we want to model and/or emulate. Note that some researchers elide external validity and ecological validity.

Ecological validity, which arguably is fundamental to the study of learners' affect, refers to whether affect judgments are made in an environment that is representative of the environment that we want to model theoretically or through technology. Here, *environment* refers to all: temporal settings, location and participants, as well as learning tools used. Achieving ecological validity in laboratory settings is challenging in most empirical research, but it is especially difficult in relation to affect. This is because emotions are shaped by the specific moment-by-moment contexts and settings, and this in turn always raises a possibility that the observations made in one context may not be valid in another context. Examples of attempts to achieve (or preserve) ecological validity given in this review include studies by Porayska-Pomsta et al. (2008) and Baker et al. (2004). Achieving ecological validity is arguably one of the most important issues in affect-related research, because despite presenting many logistical and practical challenges, it also carries the promise of the affect judgements being generalisable to other similar learning contexts, whether traditional or mediated through technology.

Establishing construct validity (*are we measuring what we are claiming to be measuring?*) requires the demonstration of reliability, convergent validity and discriminant validity (Campbell & Fiske, 1959). Reliability implies that the same or similar measurement device should produce measurements that are highly correlated. For example, affect judgments provided by two or more annotators observing a learner should be strongly correlated (inter-rater agreement, i.e. whether two or more judges agree about the definitions of labels and the phenomena that they denote in the specific contexts). 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 self-reports of affect (Measure 1) can be correlated with facial expressions annotations (Measure 2) (Bonnano and Keltner, 2004; Mauss, Levenson, McCarter, Wilhelm, & Gross, 2005). In turn, this implies the need to distinguish between

multiple measures (e.g. researchers' observations and learners self-reports) versus multiple instances of the same measure (i.e. judgments by two observers). In particular, 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 observers (see for example Graesser et al., 2006).

Employing multiple judges in the evaluation of learners' emotional reactions and correlating the resulting judgements to establish inter-rater agreement is standardly expected by the research community, because it informs the reliability of the measures used (e.g. coding schemes) and of the data generated (actual judgements). However, it is important to bear in mind that whilst the judgements by multiple annotators may be highly correlated, this does not guarantee their accuracy or the correctness of the specific coding scheme. To enhance further the fidelity of the data gathered, it is important to employ observers who are independent from the design and execution of the research for which the data is gathered to remove the possibility of overfamiliarity with the research goals and the bias that such familiarity may cause in the judgments and the use of a given schema (see also Section 4.3).

3. *WHAT?:* Types of instruments and tools for eliciting information about learners' emotions.

Gaining access to learners' affective states is necessary to inform the specific pedagogical and technological designs and it constitutes the focus of contemporary research in this area. The most common types of instruments used to obtain information about learners' affective states are self-reporting instruments. In general such instruments can be divided into *forced-choice* and *free-response* (also known as *open-ended*) instruments. Forced-choice instruments can be further subdivided into discrete and dimensional instruments. Each instrument is briefly described, followed by an outline of their respective advantages and disadvantages. The reader is referred to the original sources cited for a detailed description of each instrument, illustrations and downloads.

Forced-choice Discrete Affect Instruments provide the reporter with a pre-defined list of words describing the affective states of interest to the researcher (e.g. Cowie, 2005). The reporters[1] are asked to rate the learners' emotions along nominal, ordinal, or interval type of scales (Arroyo et al., 2009; Conati & Maclaren, 2009; Sabourin, Mott, & Lester, 2011; Strain & D'Mello, 2011). For example, a reporter can be asked to select one affect label from a set of labels (nominal or categorical response) or to report, via a Likert type scale, the degree to which an emotion is experienced. As an example of label-selection instruments, Strain and D'Mello (in review) asked learners to select *one* of six emotions at multiple points during a reading comprehension task. In this approach, the selection of the choices was facilitated by a simple drop-down list.

As an example of the alternative approach, Arroyo et al. (2009) asked learners to rate the extent to which they were experiencing *four* emotions on a 1-5 Likert scale during interactions with a maths ITS. The ends of the scales corresponded to the extreme negatives for the lower end (e.g. 1= 'I feel very anxious') and to the extreme positives for the higher end of the scale (e.g. 5= 'I feel very confident'), with the value of 3 corresponding to a neutral state. The students were prompted to enter the values for the four emotions every 5 minutes.

There are crucial methodological differences between single-choice responses and Likert

[1] Note that "reporters" can be the learners themselves, tutors or other observers as discussed in Section 4.

scale type responses, which lead to different types of data being generated and consequently shedding light on different research questions: the first leads to emotions being labelled as discrete occurrences and it often serves to ascertain which emotions, out of a set of possible emotions established *a priori*, are the ones that describe the particular learning domain the best. The second may be used to capture the extent to which different emotions are being experienced and provide a fine-grained insight into the quality of these emotions. Apart from shedding light on the complexity of the emotional experiences, as reported by the learners, such information can be of particular use in enhancing the adaptive power of communication modules in intelligent learning environments, for example, by providing the basis for modulating the feedback according to the degree to which the learner is thought to experience a particular emotion (e.g. Porayska-Pomsta and Mellish, 2013; Porayska-Pomsta et al., 2008).

There are also examples of combining forced-choice discrete instruments, for example, where reporters are asked to both select a label from a predefined list as well as to indicate a value from a Likert scale during interactions. In Porayska-Pomsta, Mavrikis and Pain (2008) drop-down lists were provided to the tutors who were asked at every point at which they have given feedback to the learner, to select several emotions that they thought the learner may have experienced at those points and to indicate the intensity of each emotion selected on five point, fuzzy linguistic scales (see also Figure 2, "Current Situation" window, in Section 5.3 for an illustration).

Forced-choice Dimensional Affect Instruments rely on structured descriptions of emotion within a dimensional space. There are several of these instruments and the most notable, albeit not the only, examples are implemented in tools such as Feeltrace (Cowie & Cornelius, 2003), NTX Feeltrace (Reidsma, Hofs, & Jovanovic, 2005), the Self Assessment Manikin (SAM) (Bradley & Lang, 1985), the Geneva Emotion Wheel (Scherer, 2005), and the Affect-Grid (Russell, Weiss, & Mendelsohn, 1989). When using such instruments, the reporters are expected to locate the emotional states within the space represented by one or more dimensions. Example dimensions include valence (pleasantness vs. unpleasantness) and arousal (active vs. inactive) (Cowie & Cornelius, 2003; Larsen, McGraw, Mellers, & Cacioppo, 2004). The Geneva Emotion Wheel arranges emotions in two-dimensional space and, by denoting the distance from the origin, it is able to represent the intensity of the associated feeling.

Open-ended (free-response) Affect Instruments allow reporters to freely discuss (or write) about the affective states. The reporter can enter a specific affect term (e.g., happy, sad, confused), a general valence term (slightly displeased), or an arousal term (e.g., very active at the moment). Alternatively, the researcher can simply ask the reporter how they feel at the present time or in response to a particular event or stimulus or, if they are acting as observers, how they interpret other people's emotions. The reporter can also be given the opportunity to enter their feelings or their observations/interpretations of other people's feelings in a text box. Verbal protocols represent another form of free-response and they can be used to capture and to record free-flowing reports of learners' emotions or interpretations thereof by observers during interactions as shown for example in D'Mello et al. (2006). Porayska-Pomsta et al. (2008) combined tutors' verbal protocols with their authoring of affect terms. This free-response capability was embedded within the same forced-choice discrete tool described earlier and was designed to capture the points of potential mismatch between the tutors' interpretations of the learners' affect and the pre-scribed labels if the participating tutors' felt the choices available were not adequate.

In order to organise the open-ended responses, the researcher needs to develop a coding scheme. Such scheme will allow them to analyse and to report the emotions detected in the context studied. However, to ensure validity and reliability, such a scheme is subject to tests of inter-rater agreements about the definitions of labels and the phenomena that they denote in the specific contexts. The schemes can be developed from the ground-up or adapted from the

existing schemas. If developed from the ground-up, typically this can be done based on a combination of two or more of the following sources: (i) exiting research describing or analysing emotions in the context or domain of interest, (ii) existing schemes such as one of the examples discussed in relation to forced choice response instruments, (iii) the researchers' hypotheses about the emotions of importance/relevance to a specific context and/or target population, (iv) tutors' input based on detailed analysis of relevant data, for example through post-task walkthroughs intended to disambiguate and explain the labels assigned to specific episodes in some interactions in which the same tutors were asked to provide free-choice responses. Whatever combination of sources is chosen either for practical or scientific reasons, a number of substantial cross-cultural resources for labelling and defining emotions are available in different European languages, including the HUMAINE handbook (Petta, Pelachaud and Cowie, 2011) and the Emotion Mark-up Language (Schröder et al., 2010). Such resources are of great importance in the study of affect, because they reflect many years of collaborative and principled research aimed to establish a consistent, replicable and sharable basis for emotion labelling that ultimately can be extended and built on by a wider community of researchers.

3.1 Advantages and disadvantages of the reporting instruments

All of the types of instruments presented thus far have both their strengths and weaknesses and none can be said to be absolutely better or worse than one another as this depends on the context of their use and the target population. The choice does not have to be an exclusive one either, as the instruments can be combined and used at different research stages as deemed necessary and as determined by the desired outcomes of any given study. However the choices need to be made based on a balanced understanding of the advantages and potential challenges associated with each of them. This subsection reviews the common strengths and weaknesses of the different types of instruments introduced, but for the details about each particular instrument, especially the forced-choice instruments, the reader is referred to the original research cited.

One advantage of forced-choice discrete response is to ensure homogenous data and to ease the researcher's task of analysing the data. However, both the definitions of the emotions and the instrument itself may influence the resulting reports. It is often difficult to gauge the extent to which the reporters' understanding and labelling of the affective states actually correspond to the predefined labels of possible emotions. Hence, a preparatory session is needed to align the reporters' labels with those of the researchers. Such preparation is standardly expected when the reporter is a trained expert, and inter-rater reliability checks are an expected part of the procedure. But such preparation can be, and often is, more informal when self-reporting is used, creating a risk that different participants interpret the definition of an affective state in subtly different ways. Reporters might also want to report an emotion that is missing from the assigned list. Forced-choice response tools do not facilitate this, though often "catch-all" or "other" categories are included to prevent labelling that is inconsistent with the forced-choices given. Furthermore, the use of forced-choice response does not ensure that the predetermined labels do not influence reporters to report an affective state that they would not report had they not been primed by the choices provided (Russell, 1994). Finally, it is difficult to compare results from different studies when different scales, sets of labels and definitions are used by different researchers (Scherer, 2005). This lack of systematicity is one important limitation of the forced-choice discrete measures, although there are growing efforts by researchers to ensure availability of common references such as the HUMAINE and EmotionML resources discussed earlier. Such references are created and continuously improved to help researchers in this field increase both the validity and reliability of the data generated by them.

Dimensional response approaches allow the reporter to report affect in a more systematic way. However, it can be difficult for reporters to relate affective states to the complex dimensions that are typically the focus of these measures. For example, *dominance* is commonly used in characterising affective states, but it is considerably less easy to interpret for non-experts than other dimensions such as valence (unpleasant vs. pleasant) and arousal (active vs. inactive). 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 to aid such differentiation (Fontaine, Scherer, Roesch, & Ellsworth, 2007). For example, one intriguing finding is that images of a "pizza" and an "erotic male" yield surprisingly similar ratings in a valence, arousal, and dominance space (Kaernbach, 2011), thereby indicating that an additional dimension – preferably one that relates to the context in which the specific affective states occur – is needed to further discriminate the affective response elicited by such different stimuli (unless the affective responses are in fact similar).

Open-ended instruments address several of the problems associated with forced-choice instruments but are laborious to categorise and code. In general, a composite approach consisting of affect labels, dimensions of emotions, and free responses might provide the most consistent and defensible way of obtaining information about learners' affect. For example, a reporter might be asked to annotate (a) a specific discrete emotion from a list of pre-specified emotions with an option of an "other" category where an open-ended response can be made if deemed necessary (as in Porayska-Pomsta et al., 2008) and (b) position the emotion in a pre-specified dimensional-space. A study by Sazzad and colleagues (2011) illustrates this point. In this study, 20 participants provided discrete emotional labels (i.e., they selected one emotion from a list of eight emotions) as well as reported levels of valence and arousal by clicking on an appropriate cell in a valence-arousal grid. In addition to obtaining three affect measures (the emotion label, valence, and arousal), the researchers were able to project the valence and arousal ratings corresponding to the different discrete emotion labels in a valence-arousal space to ensure that the labels were being consistently interpreted (e.g., boredom should appear in the quadrant representing low valence and low arousal).

4. WHO?: Sources of information about learners' emotions

The reporting instruments can be used in a variety of different contexts and as part of different methods and methodologies. They can also be used to elicit affective information from different types of reporters. Possible reporters can be categorised according to their proximity to the emotions experienced and the context of those experiences. These include: (i) the learners themselves reporting on their emotions experienced first-hand and the resulting reports are termed *self-reports*, (ii) external annotators who are participants in a learning situation, including peer learners and tutors, and who have direct access to the emotional reactions to be judged in the context in which such reactions occur, and (iii) external annotators who are observers not involved in a learning situation. In the next three sub-sections, we discuss *pros* and *cons* of each category, highlighting the different factors that need to be taken into account when selecting the reporters (e.g. experience, age, metacognitive skills, multitasking abilities, type of affective states to be labelled). All of these factors, individually or in combination, may impact the nature and the quality of the resulting data. Note that an additional influencing factor of whether the annotations are conducted concurrently with the learning task or retrospectively, will be discussed in Section 5.

4.1 Learners as reporters

Learners themselves provide a common and frequently used source of information about their emotions. It is important to consider that all learners' reports result in data that is subjective in nature and that may reflect their folk theories about what their affective states are or should be. Crucially, the quality of the data obtained from learners depends upon both the type and subtlety of the emotions to be reported, as well as on the learners' ability to report their affective experiences. This ability, in turn, is impacted by learners' individual differences such as their meta-affective skills, personality, culture and age. For example, studies with children suggest that it is difficult to elicit reliable, coherent self-reports from younger learners (e.g. Conati & Mclaren, 2009). This is also conditioned by the fact that children's understanding of emotions is very crude until approximately the age of 8 and correspondingly, there are substantial differences between the specific age groups' abilities to recognise, categorise and label their own and others' emotions into fine-grained affective categories (Safyan & Lagattuta, 2008). Thus, any instrument used for eliciting self-reports from young children must be adapted appropriately to their cognitive, metacognitive, and affective capabilities. This typically means adjusting the questions asked and providing the children with additional tools, such as pictorial representations of emotions (Read, MacFarlane, & Casey, 2002; 2006; Frauenberger et al., 2012), for communicating their feelings and perceptions. Thus, different populations of learners will have to be assessed in terms of their needs and abilities to generate the reports, and the data gathered will need to be qualified and interpreted in relation to such assessments. Beyond learners' ability to accurately report their affect, there is the challenge of the demand [that is being imposed on the learner] and the self-presentation effects (Tourangeau & Yan, 2007). Learners may be uncomfortable making affect annotations that indicate a lack of capability (e.g. frustration), or negative attitudes about the technology (e.g. boredom), particularly if there is a perception that the researcher collecting the data may be angered, upset, or disappointed by the data (cf. Nielsen, 1991). Finally, self-reports are limited to situations where the emotional episode is sufficiently pronounced to enter learners' consciousness so that it can be subjectively accessed.

4.2 Tutors as reporters

Understanding the relationships between affective states and specific behaviours 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 actively involved in generating the data (i.e. who are involved in the tutorial sessions) are also asked to annotate the data.

In contrast to data collected from independent annotators, collecting reports from tutors has the important advantage of generating data that helps not only in diagnosing learners' affective states but also in modelling pedagogy and in designing appropriate responses to learners' affect. In particular, it allows the researchers to link the tutor diagnoses of learners' 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 data collected through tutor reports include information about what affective states the tutors think the learners experience during specific interactions, the learners' behaviours that lead to tutors' specific judgements, and concrete tutoring actions committed as a consequence of those judgements. For example, in Porayska-Pomsta et al. (2008) tutors were asked to report on learners' affect, using a partially specified list of common states (derived from previous studies), while engaging in tutorial dialogue with the learner *via* a chat interface. While the data thus collected provided a direct mapping between tutors' feedback and their diagnosis of the learners' emotions in context (as judged by the tutors *in the moment*), post-task walkthroughs with the same tutors allowed the researchers to further establish what

learner actions led each tutor to make their judgements (see also Subsections 5.3 and 5.4). Typically, such sources of evidence 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. Sometimes tutors are able also to point to their own specific actions as potentially impacting learners' affective states. For example, in the same Porayska-Pomsta et al. study, tutors sometimes anticipated "dips" in learners' confidence when they set a question or problem difficulty level as 'high'.

Tutors' reports can be compared with and complemented by results derived from studies using learners' self-reports. In this way, the two different perspectives, that of the tutor and the learner, can be reconciled to derive a more accurate model. Ideally, such rich data should be further combined with other data such as pre- and post- tests of learners' knowledge to yield information about what affective states lead to increased learning and what specific tutoring actions are most effective.

Expertise in the subject domain taught and, crucially, tutoring experience may also have a significant impact on annotators' ability to provide 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.). The reports by tutors, who are not used to considering the pragmatics of the interaction explicitly, often focus solely on the content, thereby yielding little information about their judgements and interpretations of the learners' behaviours and their underlying emotional causes. The more experience the tutor has, 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).

The disadvantage of relying on tutors' reports *in situ* is that they may impose additional complexity due to multitasking required, which has been associated with both decreased reaction times and increased error rates (for a detailed review of results from task shifting experiments see Waszak et al. (2003)). For example, in the most extreme cases, tutors may be asked to tutor in real-time, communicate with the student, and report their observations of the student's affect (see section 5.3). Switching between so many tasks may impact the quality of the tutoring and/or of the resulting data and, consequently, it may increase the effort needed to explain the data *post-hoc*. However, Ericsson and Simon (1999, pp. 91-101) point out that although increased cognitive load caused by multitasking may indeed interfere with the verbalisation in favour of the task-oriented processes, this is mostly the case for perceptual-motor and visual encoding processes and not necessarily during tasks where participants provide rationale for their task-related decisions (i.e. tasks relevant to tutors' reporting). Related research (Fidler, 1983) also provides empirical evidence that concurrent verbalisation does not seem to affect the reliability of decision outcomes. In the absence of conclusive evidence, however, it seems prudent to reduce the effort required on behalf of the tutor both with respect to the tutoring task and reporting. Therefore, if technology is used to aid the knowledge elicitation effort, it is worth investing in an interface design that is as seamless as possible. One way of lessening the effort involved is to prepare a set of questions or problems that the tutor might give to the student and allow the tutor simply to click on each problem to reduce the amount of typing needed. This is particularly useful in the domains such as mathematics where the typing of formulae can be cumbersome and time consuming (e.g. 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 increase reaction time, with appropriate prior training, the tutors tend to become more efficient at systematic reporting and find that it eases their task in the long run (Porayska-Pomsta et al., 2008).

4.3 External annotators as reporters

Another method for eliciting affective information is for an independent annotator to provide a report based on observations of learners' participating in the target learning experience. Such reports are often and interchangeably referred to as *annotations*, *codes* or *observations* and the reporting task is referred to as the *reporting* or *coding* task. An annotator can either be a peer student, a researcher, a teacher or any individual with appropriate experience or training.

The choice of annotators requires careful consideration depending on the goals and context of the study. Although there are no hard and fast rules, the age of the annotators, their familiarity with the context under investigation, and cultural proximity to the target population tend to be the primary factors to consider. For example, in the context of a simple multiple-choice environment for language learning, de Vicente (2003) reported that postgraduate students found it relatively easy to annotate other students' interactions and the inter-rater agreements between their annotations were also high. This supports the hypothesis that annotators who are themselves learners, even if they are not directly involved in a given learning situation, may be better qualified to interpret the affective experiences of their peers than annotators who are no longer in formal education (Mavrikis, 2008). However, there is also some evidence that not all learners make good judges of other learners' affect. Specifically, learners who are less experienced, e.g. undergraduate students, may lack the ability to accurately judge emotions experienced by their peers (Graesser, et al., 2006). Some researchers also posit that the ability to detect emotions accurately requires considerable teaching or tutoring experience (Goleman, 1995; Lepper & Woolverton, 2002) and/or prior training in assessing emotions (Sayette, Cohn, Wertz, Perrott, & Parrott, 2001).

The complexity of the interaction and the annotators' familiarity with a given learning environment may further impact the quality of the resulting annotations, especially if these are done post-hoc. For example, Mavrikis et al. (2007) note that annotators found it difficult to report on learner affect when annotating screen recordings of learners' online interactions 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 (i.e., 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 were neither accustomed nor specifically trained to interpret. However, similar difficulties may not occur in field observations where an annotator has access to facial expressions, postures and other *in situ* behaviours of learners.

Cultural differences between observers and the learners might also influence the accuracy of the observers' judgments of learners' affective states. While Ekman et al. (1987) propose that expressions of basic emotions (e.g., anger, fear, sadness) are universal and are 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 (Elfenbein & Ambady, 2002; Russell, 1994). Additionally, practical experience of the fifth author suggests that when conducting live observation methods to study affect in classrooms, observers from socio-cultural backgrounds other than the study population sometimes may achieve poor inter-rater agreement, when compared to observers from the same background as the study population. This pattern has been noted both in the USA (with observers from Taiwan and Brazil), and in the Philippines, where two annotators – one Philippine and one Cambodian – were employed.

One of the main advantages of the data collected from external annotators, as opposed to the data collected from self-reports, is that so long as inter-rater agreement is validated, there can be fairly high confidence that all reports of an affective state involve the same constructs.

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. By contrast, if multiple students report on their affect, it is difficult to be certain that they are all referring to the same construct, because of the idiosyncrasies in the criteria they may use (in different situations) for categorising and naming the emotions experienced.

On the flip side, external annotations may lack the "internal perspective" of learners' self-reports or the long-term information about a specific learner's responses that a tutor or a peer familiar with the learner and the learning situation 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 themselves. If 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 overlook vital instances of learner affect. In our experience, this is seldom a problem with young learners, but can produce significant challenges when studying adults who are more likely to mask their emotions. In addition, the more fine-grained the target set of emotions, the more difficult it is to tell emotions apart. For instance, in a study by Conati, et al., (2003), two observers were unable to consistently recognise instances of pride/shame towards oneself and admiration/reproach towards a virtual 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.

5. *WHEN* to elicit information about learners' emotions?

Information on the learners' affective states can be collected as the learners experience them during a learning activity (*concurrent* reports) or after a learning activity is completed (*retrospective* reports). In the rest of this section each approach is discussed both in the case of the learners reporting their emotions as well as when learners' emotions are annotated by external annotators or coders.

5.1 Concurrent learners' reports

Eliciting concurrent learners' reports involves employing any of the instruments described in Section 2 to allow learners (as reporters) to report their emotions while they are experienced during a learning activity. *Concurrent free-response* techniques can be seen as a variation of the *think-aloud* protocols that have been extensively used to help learners verbalise their mental processes 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 a variant of this method as *emote-aloud*. In this implementation of the method, students were asked to report their emotions verbally as they perform a learning task, whereby the task involved interacting with an Intelligent Tutoring System (ITS), with verbal reports having been recorded for offline analyses. Students were given a list of emotions (anger, boredom, confusion, contempt, curious, disgust, eureka, and frustration) along with definitions. They were instructed to verbalise any of these emotions (e.g., "I'm so confused right now" or "This frustrates me") as they were subjectively experienced during the ITS interaction. They were also encouraged to express any affective state not included in the list provided, as well as instances in which they experienced multiple emotions. Affect reports in this approach were always voluntary in that neither the system nor the experimenter ever prompted the students to make an emote-aloud.

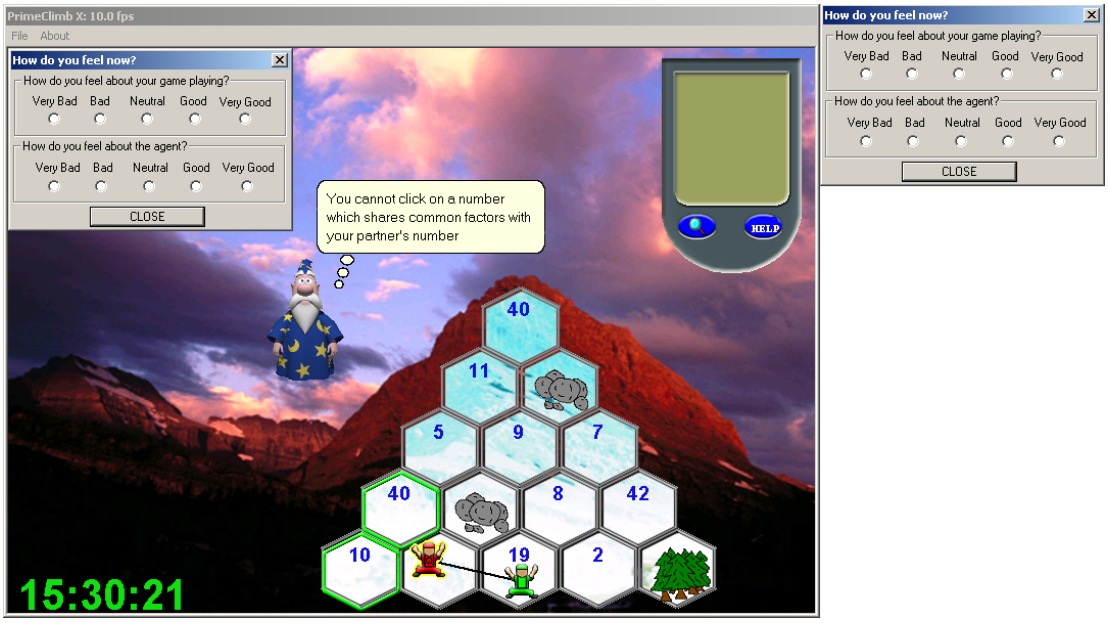

Figure 1: Pop-Up selection boxes for emotion-self reporting in the Prime Climb educational game (Conati 2004; Conati & Maclaren, 2009). One selection box (on the right) is always present, whereas the one inside the games screen pops-up as needed.

Concurrent forced-response techniques usually involve interface mechanisms that are integrated with the learning environment and that are designed to facilitate the expression of pre-defined emotional states in real-time while interfering as little as possible with the interaction flow. Examples of such mechanisms include carefully timed pop-up selection boxes (e.g. Conati 2004, 2009), combo-boxes in-between pages of a reading task (Strain & D'Mello, in review), sliders and drop-down lists (e.g., de Vicente & Pain, 2002), or more recently, tools that allow students to report their emotions as status updates in social networks (cf. Sabourin et al., 2011).

An example of this technique is seen in the pop-up selection boxes described in Conati (2004) and Conati & Maclaren (2009), where they were used to elicit two pairs of non-mutually exclusive emotions during interaction with an educational computer game: *joy vs. regret* towards states of the game and *admiration vs. reproach* towards a virtual agent which provided didactic help during the task (see Figure 1). One version of the selection box is permanently present on the side of main game window, for students to volunteer self–reports on their emotional states (right hand side of figure 1). However, the selection box also pops up whenever either: (i) the student has not used the permanent dialog box for longer than a set threshold or (ii) an underlying model designed to automatically detect changes in the student affective states estimates that one such change has happened. The thresholds that influence the appearance of the pop-up box were adjusted through pilot studies to maximize the amount of data that it allows researchers to collect while minimizing the level of interference that it generates during game playing (Conati, 2004).

The main advantage of concurrent learners' reports is that they can provide *in-the-moment* insights into the emotions experienced by the learners, especially when free response techniques are used (see Section 3). Their main disadvantage is that it is non-trivial to seamlessly incorporate concurrent emotion reporting in the learning experience. First, the sheer act of learners' reporting on their affective states may actually influence and change them. Second, asking learners to engage in a meaningful learning experience while self-analysing their emotions may have *cognitive load* implications (see also Section 4.2) and may interfere both with learning and with the actual reporting task. This is particularly the case

for the younger learners (c.f. Branch, 2000). This raises further concerns about the validity of concurrent reporting that have been thoroughly discussed in the literature, especially with respect to reactivity i.e. the extent to which the verbalisation influences the way in which a specific type of task is performed (see for examples Russo et al. 1989; Ericsson and Simon 1980, 1993). Unfortunately, the extent to which concurrently self-reporting emotions may impact the learners' affect is not well understood, with limited research having been conducted on this issue to date.

Exceptions are the work by Conati (2004), and D'Mello and Mills (in review). Conati (2004) discusses an exploratory study in which 20 students interacted with an educational game outfitted with emotion-selection boxes (as introduced above), and completed a post-questionnaire on interface acceptance. The results were quite positive in that the students' average ratings (on a Likert scale, where 1 = strongly disagree, and 5 = strongly agree) for the statement "The popup dialog box interfered with my game playing" was 2.8 (st. dev. 1.4), while the average ratings for "It bothered me having to tell the system how I feel" was 2.1 (st. dev. 1.1). The same study found that the negative emotions reported were only a small fraction of the self-reports generated by the students who reported annoyance with the dialog box. These results suggest that, even when subjects expressed annoyance with the dialog box, this annoyance did not translate into annoyance with the game or the agent.

A recent study by D'Mello and Mills (in review) obtained similar results. In this study, the researchers interrupted participants with a brief 4-item affect-rating questionnaire every 2 minutes during a 12-minute essay-writing task. After writing the essay, participants were asked if they found the affect rating questionnaire frustrating. To answer they had to select from one of the following three options: very frustrating, somewhat frustrating, and not frustrating. Only eight of the 166 participants (4.8%) reported that they found the affect-rating panel very frustrating, 50 (30.1%) reported some frustration, while the majority – 108 (65.1%) – reported no frustration. Importantly, there was no difference in performance (essay quality as coded by standard rubrics) for those reporting some frustration vs. no frustration. Taken together, the findings by Conati (2004) and D'Mello and Mills (in review) are encouraging for researchers interested in evaluating affective models, because they indicate that subjects can tolerate an extent of interference caused by the artefacts designed to measure their emotions. Nevertheless, depending on the exact context in which the reporting takes place, e.g. how complex is the learning task, presence of peers or other potential observers, etc., additional checks are advised as a way of validating any concurrent self-reports data.

One approach to address the potential unreliability of concurrent self-reports is to further clarify the responses collected during the task through *post-hoc* discussions with the learner. This method is often interchangeably referred to as *retrospective*, *post-task*, or *post-hoc walkthroughs* and requires access to a record of the learning episodes (e.g. Porayska-Pomsta et al., 2008). Such a record may consist of video- and/or audio-recordings of the learner engaging in a learning task, or if a tutoring system 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 engaged in a learning task. However, this approach requires the learners to be available to participate in research over longer a period and thus may often be unfeasible due to practical reasons. The drawbacks of learners' concurrent self-reports are reduced, to some extent, by making learners report their emotions retrospectively, as described next.

5.2 Retrospective learners' reports

With retrospective reports, learners engage with a learning task first and report on their emotions during that task later. Reporting is usually elicited by engaging participants in an audio- or video-stimulated recall interview, and can involve either free responses based on appropriately designed open-ended questionnaires, or forced responses based on dimensional

or discrete emotion response tools. For example, Mills and D'Mello (2012) used a retrospective affect judgment procedure to monitor affect during an argumentative writing task. Participants were given 15 minutes to write essays on two topics. They typed their essays on a computer interface and the text was saved for offline analyses. Videos of participants' faces and computer screens were recorded during the writing session. Participants provided self-judgments of their affective states immediately after the writing session. The procedure began by playing a video of the face along with the screen capture video on a widescreen monitor. The screen capture included the writing prompt and dynamically presented the text as it was written, thereby providing the context of the writing session. Participants were instructed to make judgments on what affective states were present at any moment during the writing session by manually pausing the videos. They were also instructed to make judgments at each 15-second interval where the videos automatically stopped. Participants made their ratings via a computer interface that allowed them to select one out of 15 affective states from a drop-down list. Hence, judgments were made on the basis of the participants' facial expressions, contextual cues via the screen capture, the definitions of the states (presented on a sheet), and their memories of the recently completed writing session.

Applying this method requires significant time commitment on the part of researchers and the reporters and therefore may not be feasible on a large scale, for example if the reporters are students who are about to complete a course. Another noteworthy limitation of retrospective learners reporting of affect is the temporal duration between the time the learners are engaged in a task and the time of the report. In addition to differences between the affective states that are retrospectively reported compared to the ones experienced during the task (Masthoff & Gatt, 2006), the learner may also not remember exactly what they felt at specific points during the task. Furthermore, learners' memories may be biased by post-hoc rationalisations of the affective experiences and whether or not the learning task was completed successfully.

However, one advantage of retrospective reporting is that it eliminates the problem of increased cognitive load and interference with the learning activity, while offering an opportunity to cross-validate and qualify data gathered by other means, including concurrent self-reports and interaction data logging. Furthermore, this approach 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 to the research questions. By the same token, this method can offer to the participants an opportunity for in-depth reflection about their learning and a possibility to verbalise and discuss those reflections with a researcher or peer, which arguably, may contribute to the learners' developing better metacognitive skills of essence to learning. Systematic research is needed to validate this last claim.

5.3 Concurrent annotation of affect by tutors

As with the previous methods, tutor reports of learners' affect can be done concurrently or retrospectively and the methods are often combined to yield more informative data. If the concurrent method is used, the tutor is asked to simultaneously tutor the learners and to code or to comment on learners' emotions observed *in-the-moment*. This has the added advantage of resulting in a data that consolidate both the tutor judgements of the learner's affective states and their tutoring actions and aims to achieve external and/or ecological validity (see section 2.3 for definitions). For example, in Forbes-Riley et al. (2008) tutors and learners interacted through an adaptive *Wizard-of-Oz* tutoring system, in which learner's uncertainty was manually annotated by the tutor in real-time.

Similarly, in Porayska-Pomsta et al. (2008) five tutors were asked to tutor the learners remotely, through a specially designed chat interface (see fig 2), while also engaging in (i) talking-aloud about any and every possible aspect of the interaction as they engaged in tutoring and (ii) selecting values for a set of possible affective states of the learner and other

relevant situational factors such as 'difficulty of material' or 'correctness of student answer'. The idea behind combining the different concurrent methods was threefold: (1) to collect *in-the-moment* data about the affective states in the domain studied (differential calculus), (2) to access the tutors' *in-the-moment* inferences about learners' affect including the information about the specific behaviours of learners that led to the given inferences and (3) to map between the specific behaviours identified, the affective states diagnosed and the tutorial feedback provided. The ultimate goal was to elicit knowledge that would enable the implementation of an intelligent system for teaching mathematics to a wide range of learners. The achievement of this goal was additionally facilitated by the fact that the chat interface used constrained the bandwidth of information available to the tutors to that which would be available to the system itself.

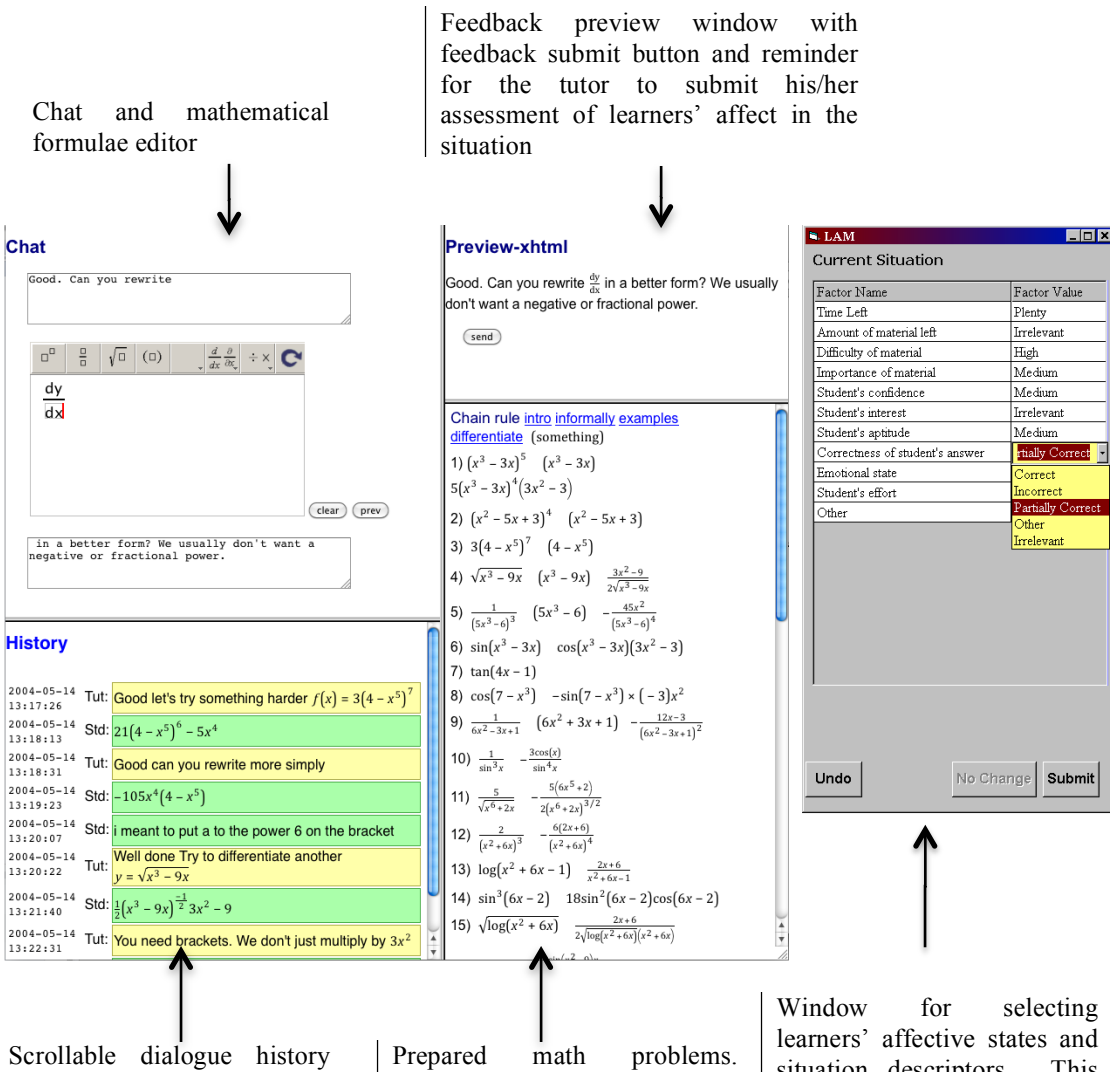

Figure 2: Dialogue and data collection interface used in Porayska-Pomsta, Mavrikis and Pain (2008) studies.

The advantages of concurrent annotations by tutors are similar to those found when the reporters are learners. However, tutors perspective is different to that of the learner in that, arguably, tutors will always view learners' emotions from the point of view of how such emotions impact learning and how they can be modulated to stimulate learning. Therefore, in addition to concurrent annotations by tutors providing a unique access to their *in-the moment* interpretations of learners' emotions, if combined with other methods such as dialogue and contextual information logging, they allow to record the specific sources of evidence used to make those interpretations, i.e. the specific behaviours and task-related actions of the learner. Examining such annotations and the recorded contextual information together with the corresponding tutoring feedback provides fine-grained insight not only into what learner emotions are tutors able to perceive, but also why they may think that they perceive them in the specific contexts and about the pedagogical strategies that they deem appropriate for modulating those emotions to stimulate learning. Such information would be difficult to capture retrospectively, because the temporal distance from the individual interactions may cause the tutors to forget the relevant detail. Even if some of the tutor interpretations of learners' affective states are incorrect at times, recording those interpretations in the context of an entire interaction as it happens, allows the tutor to identify the sources of ambiguity and to observe the specific repair strategies employed by them, when they realise their mistake. Think-aloud protocols that are collected concurrently with the annotations and the interactions provide further qualitative data that can be used to verify them and they can be used as the basis for any follow-up questions of special interest.

However, as discussed in detail in Section 4.2, adopting this method with tutors may be distracting to them and consequently it may alter the quality of their tutoring. When combined with other methods, as described above, the need to simultaneously teach and explain their teaching in the context of individual learners' affect is not a natural task: typically, tutors are not trained in explicitly identifying affective states of learners, or to providing running commentaries on how and why they identify them. Instead they often focus primarily on the content of learners' answers. Therefore, in order to apply this method successfully it is necessary to train the tutors in verbalising their observations about the learner. If a graphical interface is used to collect data, it is important to consider its design carefully to ensure that it supports the multi-tasking demanded of tutors. More training in the use of the specific tools and procedures is likely to be required, if the participating tutors are less experienced in tutoring (also see Section 4.2).

5.4 Retrospective annotation of affect by tutors

When used with tutors, who have been involved in given tutoring interactions, the retrospective annotation method has a clear advantage of not interfering with the tutoring task, while also giving the tutors the opportunity to pause and reflect on the detail in learners' behaviours that may be indicative of the specific emotions detected. The way in which this method can be used with tutors – the procedure, the tools needed and its further advantages and disadvantages are very much the same as when the method is used with external annotators and therefore the reader is referred to section 5.6, where these are discussed in detail. This subsection briefly considers the use of this method as a way to retrospectively validate and elaborate on annotations that were previously made by the same tutors concurrently. When used for this purpose, this method allows the researcher to discuss the reasoning behind the concurrent annotations made by the tutor and to resolve any inconsistencies therein. In Porayska-Pomsta et al. (2008), the tutors were invited to participate in retrospective annotations (also referred to as *post-task walkthroughs*), following each completed interaction. Given that the concurrent annotations employed in these studies imposed high cognitive demands on the tutors, it was very likely that some important episodes may have been mis-annotated. For example, although the tutors were forced to submit their annotations of learners' affect, they also had an option to press a 'no change' [observed] button, which allowed them to move on with the interaction without necessarily

making a note of the actual observation. Therefore, post-task walkthroughs were also necessary to allow the tutors to double-check their concurrent annotations and to correct them accordingly. Note that concurrent annotations logs were preserved and separate logs were created for the retrospective annotations to enable future comparison.

The tools needed for retrospective annotations by tutors typically include replays of videos of the interactions in which they participated. These can be paused at specific places of interest. In Porayska-Pomsta et al., replays of videos of students' screens were used and these were synchronised with replays of the corresponding learner affect annotations and think-aloud protocols by tutors. The tutors observed the replays while listening to their own narrative of what was happening in the moment. These tools provided memory props (or cues) for the tutors and as such facilitated informed elaboration of the annotations. Both tutors and researchers were allowed to pause at any point in the replays to elaborate or to ask for elaboration. Using such memory props is all the more important if the temporal distance between the tutorial interactions and retrospective (re-) annotations is long – in Porayska-Pomsta et al. (*ibid*), such temporal distances were up to 2 months owing to tutors' busy schedules.

Another procedural variant on the retrospective annotations that rely on prior concurrent annotations involves two tutors: one who has done the tutoring and another who is external to it. In this setup the researcher acts as a facilitator of the discussion between the two tutors and the scribe. Here, the idea is to encourage the annotations and their explanation between two professionals who share in-depth understanding of the tutoring goals, the common problems that the learners' encounter in a given domain and a qualitative insight into the differences of perception between the individual tutors. This method is particularly valuable in determining learners' affect in ill-defined domains such as learning of social interaction skills, as will be reported in Porayska-Pomsta and Bernardini (in preparation). In these studies, the two practitioners discussed social and emotional cues of the learners during replays of video-recorded training sessions, with the tutor internal to the training session having the final say about what annotations to enter. The discussion between the tutors, involved a negotiation of what cues where actually displayed and resulted in many disagreements, which were accompanied by both practitioners providing explicit reasons for the differing interpretations. These annotation sessions were video recorded, with the researcher present and recording the discussions through notes made directly in the linguistic and video annotation tool called Élan (Sloetjes and Wittenburg, 2008) at crucial points in the interaction, as indicated by the internal tutor.

An important aspect of retrospective annotations that are made on top of the concurrent ones is that it gives the tutors an opportunity to consolidate their perception (or training) about what constitutes good practice: while professional tutors and teachers are taught to pay attention to learners' moods and predispositions in general, as observed earlier, they rarely have experience of thinking about learners emotions on a moment-by-moment basis. They are also rarely made to verbalise their reasoning about the individual learners or explain why they are tutoring in a specific way. As reported in the Porayska-Pomsta et al. (2008) and Porayska-Pomsta and Bernardini (in preparation), accessing past interactions and examining them in minute detail can be revelatory not only to the researchers who aim to use the knowledge thus elicited as the basis for designing intelligent technologies for learning, but also to tutors, who have identified this process as an excellent teacher-training method.

5.5 Concurrent annotation of affect by external annotators

In this method, one or more annotators observe students engaging in a specific learning task and code for affective categories that are usually predefined, in real-time. The learning task may involve traditional (i.e. human-human) or technology-enhanced interaction. These

concurrent methods are especially suitable for obtaining affect information in 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., 2007, 2008; Baker et al., 2012). These concurrent affect annotation methods build on prior methods for concurrent behavioural observation by researchers in classrooms (e.g. 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.

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 exactly when an observation is occurring. The purpose of this is to reduce observer effects as much as possible. 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. Despite these methods, there is always some risk that the observers' presence may inhibit or suppress the learners' affective expressions. This issue is considerably less critical in laboratory settings where it is possible to hide the observers behind a one-way mirror. However, live observations of affect are less common in laboratory settings (though examples do exist in the literature – e.g. Craig, Graesser, Sullins, & Gholson (2004)), because in these settings it is usually somewhat easier to collect video data and thus rely on the higher flexibility afforded by retrospective annotations, described in Section 5.6. The main disadvantage of concurrent annotation of affect is that it is 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 assessments of inter-rater agreement. This can present challenges in terms of physical positioning, if multiple observers annotate the same student at the same time (for this reason, typically observers annotate separate parts of the classroom, except when establishing inter-rater agreement). This method also depends on scheduling observation times at schools or other real-world settings, a logistics challenge not present to the same degree for laboratory studies. The issues of logistics and procedure in this type of field observation are discussed in further detail in (Ocumpaugh et al., 2012).

5.6 Retrospective annotation of affect by external annotators

In retrospective coding by external annotators, the annotators work on a video recording of the learner engaging in a 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, video of the learner's computer screen can be recorded so that the context of the learning session can be considered in the retrospective judgments of affect. However, 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. Despite the reduced complexity of the task compared to concurrent report, the challenge with retrospective annotation is to strike the right balance between providing enough information on all the relevant aspects of the interaction to be coded, while avoiding cognitive overload and divided attention effects e.g. due to the multiple sources of information available. 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 when they cannot reconcile the information that they can observe on the screen, e.g. learners' mouse movements, learners typing and deleting half-constructed responses, correctness of and working-outs in learners' answers, etc. Therefore, several factors should be taken into account when deciding the informational channels made available to coders (the *bandwidth* of information), including coders' expertise with the coding process, their familiarity with the task to be coded and the granularity of the affective states to be captured. Reducing the amount of information sought in any given study is often better than trying to achieve too much at once. The coding process ought to be validated via a pilot study.

The retrospective judgment by external annotator protocol has three main advantages. The first is that it is easier to include multiple judges in the affect judgment process. For example, peer learners can first judge the students' affective states, followed by trained judges, and expert teachers (D'Mello, Taylor, Davidson, & Graesser, 2008; Graesser et al., 2006). Agreement among these different types of judges can then be assessed in order to obtain a measure of *convergent validity*, i.e. validity obtained through use of multiple measures (see also Section 2.3). A second advantage of the retrospective coding protocol 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 one theoretical question. At a later time, the videos can be recoded for affect at another set of points to answer different 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 answer different questions as the research progresses. The third advantage is that retrospective allows the annotators to be more careful, ultimately increase precision in the annotations obtained.

An overarching limitation of retrospective annotations is that they often represent information about the events presented *post factum* rather than being representative of diagnoses *in-the-the-moment*, i.e. such annotations may represent the annotators' theories about what emotions they observed rather than the actual observations. Another disadvantage is that retrospective annotations can require significantly more overall research time to annotate the data, with Baker, Corbett, & Wagner (2006) estimating that retrospective annotations of lab study data takes approximately four times as long as concurrent classroom annotation[3]. One reason for the increased time is that annotators can and often do re-watch a video clip repeatedly, potentially increasing precision at the cost of time and effort. Ultimately, the amount of time needed for retrospective annotation will depend on the exact data to be annotated, the complexity and clarity of the coding scheme as well as the experience and/or training of the coders. All of these factors need to be taken into account prior to embarking on such annotations to ensure sufficient time is available for the task.

5.7 Variants of Concurrent and Retrospective Protocols

Both concurrent and retrospective methods can facilitate detection of any occurrence of a visible 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 et al., 2008) while de Vicente and Pain (2002) employed semi-structured questionnaires and asked postgraduate tutors to elaborate on their reports 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 (e.g. D'Mello et al., 2008; 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

[3] More recent estimations by the third author suggest that retrospective annotation time can take approximately 1.5 times the length of the learning session with appropriate annotation tools. Hence, it would take roughly 90 minutes to annotate a 60-minute learning session.

selected intervals and in a pre-determined order (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. Observations are typically either twenty seconds long or thirty seconds long (cf. Baker et al., 2004; Karweit & Slavin, 1982). While justification for the time window size is not often given in detail, it seems reasonable to suppose that 20-30 seconds is long enough to be able to make valid judgments, without frequently seeing multiple affective states in one observation. These time intervals are also convenient to work with. In some cases, the first affective state observed is the only affective state coded; 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 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.

6. Discussion and Conclusions

This paper presented a review of knowledge elicitation methods to aid detection, labeling and studying learners' affect in and for intelligent technologies for learning in the AIEd sense. The focus of the review is on approaches and instruments used for eliciting affect-related knowledge from different stakeholders including learners, tutors, researchers and external annotators. Many existing reviews surveyed affect-elicitation methods that were used to study affect before the advent of technology and affective computing, with the exception of Wosnitza and Volet's (2005) and Afzal and Robinson's (2011) reviews, which are complimentary to the present review.

The present review demonstrates how the traditional knowledge elicitation methods remain relevant to modelling learners' affect and how they have been enhanced for use in educational technology design and affective computing research, especially with respect to establishing measures of *ground truth* needed to validate the various models and theories of affect in learning. The goals of the review were three-fold: (1) to critically examine the different methods in terms of *what instruments* they involve, as well as *who* generates emotion information and *when*; (2) to highlight both advantages and disadvantages of each method to provide the reader with a basis for creating balanced empirical design decisions and to be able to align those empirical designs with those of other researchers in the field; (3) to generate a methodological resource for other Artificial Intelligence in Education researchers interested in learners' emotions and their antecedents in AIEd contexts. A summary of the different methods that were reviewed, along with their advantages, disadvantages and illustrative references are given in tables 1-3, at the end of this section. The issues raised and the guidelines provided are the result of several years of discussions and knowledge exchanges among the authors, who all have many years of experience in applying, testing and refining these methods in the contexts of their own research. These issues and guidelines are by no means exhaustive, but we believe that they are representative of the considerations that need to be examined in the context of affect modelling, whether for theoretical or computational purposes. The reader is referred to the many sources cited throughout the paper for details on any specific approach or instrument cited.

Most of the methods for studying learner's affect rely on a variety of ways to measure and monitor learners' affect during learning. As highlighted throughout the paper, the selection of specific methods and instruments and the related challenges depend on the specific affect-related questions asked, the level of detail sought, the target learner population, the resources available and the desired generalisability of the resulting conclusions. To aid the assessment of the fit of different methods to a desired purpose, the methods available were categorised

according to three main considerations: (i) *what* instruments are available to elicit information about learners' affective states, (ii) *who* can be the prospective informants, and (iii) *when* the elicitation of the affective knowledge may be undertaken. It is important to bear in mind that the choices of *what*, *who* and *when* are often mutually dependent and may have to be made in tandem with one another with precisely defined research questions and goals for guidance, as well as careful consideration of how the validity and reliability of the desired data can be at least enhanced if not ensured.

While the review focuses specifically on the knowledge elicitation methods, as with any data collection methods, questions arise as to how the data generated can be combined and analysed. Specifically, in the field of affect modelling there are sometimes situations when multiple instruments, reporters, and affective phenomena are used in conjunction. This can be made manageable when the affect measurement methodology focuses on one aspect of the measurement. For example if a retrospective affect judgment procedure collects ratings from the learners' and from tutors at the same time intervals and using the same instruments, then it is possible to discriminate instances where both annotators agree from cases when they both disagree. A more in-depth analysis of patterns of disagreement can also be performed and this can be particularly informative because it might suggest that the different annotators are sensitive to different cues. The situation is more complex when multiple aspects of the measurement methodology are simultaneously varied. For example, one could use a dimensional instrument, such as the Affect Grid, to collect self-reports of valence and activation every five minutes along with online observations of discrete emotions by trained researchers every 20 seconds. There is the important question of how these multiple measures (valence-arousal vs. discrete emotions) that were recorded at different time scales (5 minutes vs. 20 seconds) and by different annotators (self-reports vs. external observers) can be reconciled. This makes it more difficult to paint a coherent picture of the learner's affective states, so it might be beneficial to construct separate models (a learner model and an observer model) and to focus on high-level similarities and differences across models.

A further important consideration relates to the Kappa score (Cohen, 1960), which is a measure of agreement that is used consistently in affect-related research. Statisticians have claimed that kappa scores ranging from *.40 – .60* are typically considered fair, *.60 – .75* are good, and scores greater than *.75* are excellent (Robson, 1993). Many psychology journal reviewers expect scores over *.75* for all constructs. Kappas associated with annotation of naturalistic affect experiences are seldom this high, although values vary somewhat depending on the type of information available. For coding voice data, values around *.40* are often seen. 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 six emotions. For video coding of affect, similar values are seen; for example, Graesser et al. (2006) obtained a kappa of *.31* for video coding conducted on 20-second sequences of behaviour. For field observation, higher values have been seen in multiple studies. Baker, D'Mello, Rodrigo, & Graesser (2010) report kappa scores of *.63* in two studies comparing six affective states in the Philippines. Baker et al. (2012) reports a kappa of *.72* for the same coding scheme in the United States, conducted on high school students using mathematics tutoring software for Algebra; Pardos et al. (in press) reports a kappa of *.72* for the same procedure with middle school students using a different mathematics tutor. One potential explanation for the higher values of kappa seen in field observations is that it is easier for a field observer to change to an ideal viewing position and to see the observed student's context and posture than with other methods. Another possibility is that the natural setting of observation, combined with the use of methods to obfuscate who is being observed, may make student behaviour more natural, and therefore more demonstrative. Still, the kappa scores obtained even in these field observation studies would be considered to be below the standards adopted by most psychology journals. However, such

claims address the reliability of multiple judges when the phenomenon is salient and when the researcher can assert that the decisions are clear-cut and decidable. A kappa score above *.80* can be expected when judges code some simple human behaviours, such as facial action units, basic gestures, and other observable behaviour, but it is unlikely that perfect agreement will ever be achieved in affect measurement because there is no objective gold standard. In general, kappas are lower when emotions are not intentionally elicited, contextual factors play an important role, and the unit of analysis is on individual emotion events (Aviezer et al., 2008; Matsumoto, Olide, Schug, Willingham, & Callan, 2009; Naab & Russell, 2007; Stemmler, Heldmann, Pauls, & Scherer, 2001).

Relatively low kappa scores also raise questions as to the difference between reliability and validity of the judgements made. Reliability is not the same as validity and clear-cut decisions, as expected by statisticians, are rarely possible in relation to affective judgments, which are fuzzy, ill-defined, and possibly indeterminate. The argument here is that every phenomenon we study has inherent characteristics that we have to live with and sometimes reliability is modest for individual observations. But that should not prevent us from studying the phenomenon – it just makes our task harder. Other methods, such as self-reports, have their own limitations, e.g. demand and self-presentation effects. As such, there is no "magic bullet" to assessing affect; a combination of methods is needed, and each has its limitations. It is worth noting, however, that difficulties in measuring a construct do not imply that it should not be measured. Given the many studies showing an important role for affect in key learning processes, it is incumbent on us to do our best to assess this challenging construct, continually improving our methods towards obtaining steadily more reliable results. The knowledge elicitation methods presented in this review can be enhanced with more objective physiological measurements to provide further grounding to the resulting observations. We believe that intelligent technologies as introduced at the beginning of this review, whose goal is to capture the knowledge generated through knowledge elicitation methods such as discussed, have a fundamental role to play in enhancing the reliability of the data by providing base models which can be inspected, manipulated and changed and which increasingly can learn from interactions in real-time.

In conclusion, emotions form a natural and arguably essential part of learning, but the study of affect and learning presents many challenges to those who attempt it, especially if the end goal is to inform the design of intelligent technologies for learning that are capable of recognising and managing learners' emotions in real-time. While many possible approaches emerge from diverse disciplines, the field still lacks a cohesive account of which emotions are relevant to learning and principled guidelines for how to measure learners' affect in context. This review was motivated by the need for such an account and it is intended as a resource for any researcher interested in understanding the role of emotions during learning with technology. None of the methods discussed in the present review, together or individually, provide a definitive tool for accessing all emotions in all contexts with all types of learners. However, a clear understanding of their individual advantages and disadvantages and a careful reflection as to the end results to which they are likely to lead in terms of data validity and reliability will increase our confidence in the observations that we make. As much as guiding novice researchers in their endeavour, this review is intended to provide a basis for critically examining the different methods available and consequently for improving them, as more data becomes available, new contexts of learning are explored and new technologies become available.

7. Summary Tables

Table 1: Summary of methods and instruments involving concurrent and retrospective reports of affect by *learners*

<table>
<tr><th colspan="2">WHO</th><th colspan="2">WHEN</th></tr>
<tr><td colspan="2">Learners as Reporters</td><td>Concurrent</td><td>Retrospective</td></tr>
<tr><td rowspan="2">WHAT</td><td rowspan="2">Instruments and tools</td><td colspan="2">Video- audio- recording
Dimensional response
Discrete emotion response</td></tr>
<tr><td colspan="2">Free response
(think-, talk-, emote-aloud) (interviews)</td></tr>
<tr><td colspan="2">Involves</td><td>Students reporting emotions during a learning task</td><td>Students reporting emotions after a learning task</td></tr>
<tr><td colspan="2">Advantages</td><td>• Provides heat-of-the-moment reports
• Allows stream of consciousness reports</td><td>• Allows elaboration and focus on details
• Reduces cognitive load[4]
• Learners do not need to know that their emotions are the focus of the study during the learning task
• Easy to prepare, administer and elicit</td></tr>
<tr><td colspan="2">Disadvantages</td><td>•Imposes high cognitive load
•Interferes with the primary task
•Learners know that they are being monitored
•May influence the emotions experienced
•Requires participant's ability to coordinate engagement in task and self analysis;
•Generates subjective data</td><td>•Requires more time per subject (time to engage in the task + time needed to obtain offline affect annotations)
•Increased distance between learning task and affect-reports
•Requires the learners to have significant meta-cognitive skills
•Generates subjective data</td></tr>
<tr><td colspan="2">For adults</td><td>Yes</td><td>Yes</td></tr>
<tr><td colspan="2">For children</td><td>Possibly for older children, with appropriately designed tools</td><td>Possibly for older children, with appropriately designed tools</td></tr>
<tr><td colspan="2">Illustrative Research</td><td>deVincente & Pain (1999);
Conati & McLaren (2009)
D' Mello et. (2006)</td><td>D'Mello, Lehman, & Person (2010)
Mavrikis et al. (2007)
Masthoff & Gatt (2006)</td></tr>
</table>

[4] Relative to concurrent self-reports, in which cognitive load is typically increased.

Table 2: Summary of methods and instruments involving concurrent and retrospective annotation of affect by *tutors*

WHO		WHEN	
Tutor participant annotation		**Concurrent**	**Retrospective**
WHAT	**Instruments and tools**	Tool for computer-mediated interaction (e.g. for wizard-of-Oz type of setting) Free response (talk-aloud) - Dimensional response - Discrete emotion response	video-; audio-recordings Free response (interviews) Dimensional response Discrete emotion response
Involves		Tutor reporting on student's affect *during* a computer-mediated tutor-student interaction	• Tutor annotating affect (for the first time) after the learning episode • Tutor revisiting concurrent annotations after the learning episode
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 student reports - Repeated use can result in tutors' improved reporting skills for a given learner	
		Provides *in-the-moment*, stream of consciousness reports	• Allows to elaborate and focus on details
Disadvantages		• requires an investment of time by researcher 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	• Relies on the bandwidth of available information in the computer-mediated interaction • 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.
Illustrative Research		Forbes-Riley et al. (2008) Porayska-Pomsta et al. (2008)	Porayska-Pomsta et al. (2008)

Table 3: Summary of methods and instruments involving concurrent and retrospective annotation of affect by *external observers*

WHO		WHEN	
External annotation		**Concurrent**	**Retrospective**
WHAT	**Instruments and tools**	Dimensional response Discrete emotion response	video- audio-recordings Free response Dimensional response Discrete emotion response
Involves		Observer annotating students' interaction with emotional judgements *during* a learning task	Observers coding students' affective states *after* a learning task
Advantages		After validating inter-rater reliability, one can be fairly confident that, compared to student self-reports, data are referring to the same psychological construct.	
		• Provide *heat-of-the-moment*, stream of consciousness reports • Rapid to conduct • Facilitate comprehensive view of interaction at a specific moment	• Allows to elaborate and focus on details; • Reduces cognitive load[5] • Easier for multiple annotators to make affect judgments • Permits multiple rounds of annotations
Disadvantages		- Lacks the "internal perspective" that a student's self-reports can give - Relies on the degree to which the learner displays his/her emotional expressions - Requires observers to be able/trained to judge emotional experiences	
		• Logistically challenging e.g. multiple observers are required	• Reports often represent theories about the events presented *post-factum* rather than being representative of diagnoses *in-the-heat-of-the-moment*. • Potentially reduces the bandwidth of available information to the reporter
Illustrative Research		Baker et al (2004) Rodrigo et al (2008) Craig, Graesser, Sullins, & Gholson (2004)	Graesser et al (2006) D'Mello et al. (2008)

[5] Relative to concurrent self-reports

Acknowledgements

Sidney D'Mello was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF. The authors would like to thank Rose Luckin and Benedict du Boulay who provided the motivation for the authors to write this review and comments on early drafts.

References

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Paper presented at the *International Conference on Spoken Language Processing*, Denver, CO.

Arroyo, I., Woolf, B., Cooper, D., Burleson, W., Muldner, K., & Christopherson, R. (2009). Emotion sensors go to school. In V. Dimitrova, R. Mizoguchi, B. Du Boulay & A. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence In Education* (pp. 17-24). Amsterdam: IOS Press.

Aviezer, H., Hassin, R., Ryan, J., Grady, C., Susskind, J., Anderson, A., Moscovitch, M., & Bentin, S. (2008). Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological Science, 19*(7), 724-732.

Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z.,Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J., Rossi, L. (2012). Sensor-free automated detection of affect in a Cognitive Tutor forAlgebra. Proceedings of the 5th International Conference onEducational Data Mining, 126-133.

Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-task behaviour 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., 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.

Barrett, L., Mesquita, B., Ochsner, K., & Gross, J. (2007). The experience of emotion. [Review]. *Annual Review of Psychology, 58*, 373-403. doi: 10.1146/annurev.psych.58.110405.085709

Barth, C. M., & Funke, J. (2010). Negative affective environments improve complex solving performance. *Cognition and Emotion, 24*(7), 1259-1268. doi: 10.1080/02699930903223766

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.

Branch, J. L. (2000). Investigating the information-seeking processes of adolescents: The value of using think-alouds and think afters. *Library and Information Science Research, 22*(4), 371–392.

Calvo, R. A., & D'Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18-37. doi: 10.1109/T-AFFC.2010.1

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 Company.

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences, 11*(9), 393-399. doi: 10.1016/j.tics.2007.08.005

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

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement, 20*(1), 37-46.

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.

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

Conati, 2004. How to Evaluate a Model of User Affect? In proceedings of ADS 2004, Tutorial and Research Workshop on Affective Dialogue Systems. Kloster Irsee, Germany, 288-300.

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 Communicatio*n, 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. *Journal of Educational Media, 29*, 241-250. doi: 10.1080/1358165042000283101

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

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., Taylor, R., Davidson, K., & Graesser, A. (2008). Self versus teacher judgments of learner emotions during a tutoring session with AutoTutor. In B. Woolf, E. Aimeur, R. Nkambou & S. Lajoie (Eds.), *Proceedings of the 9th international conference on Intelligent Tutoring Systems*. Berlin, Heidelberg: Springer-Verlag.

D'Mello, S., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence In Education, 20*(4), 361-389.

D'Mello, S., & Kory, J. (2012). Consistent but Modest: Comparing multimodal and unimodal affect detection accuracies from 30 studies. In L.-P. Morency, D. Bohus, H. Aghajan, A. Nijholt, J. Cassell & J. Epps (Eds.), *Proceedings of the 14th ACM International Conference on Multimodal Interaction* (pp. 31-38). New York: ACM

D'Mello, S., & Mills, C. (in review). Emotions during emotional and non-emotional writing.

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.

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

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion, 6*(3-4), 169-200.

Elfenbein, H., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*(2), 203-235. doi: 10.1037//0033-2909.128.2.203

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions of a different cultural group. *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.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data, Psychological Review, 87:215-51.

Fidler, Eduard J. (1983), "The Reliability and Validity of Concurrent, Retrospective and Interpretive Verbal Reports," in Analyzing and Aiding Decision Processes, P. Humphreys, Ola Svenson, and A. Vari, eds. Amsterdam: NorthHolland Publishing Company, 429-40.

Fontaine, J., Scherer, K., Roesch, E., & Ellsworth, P. (2007). The world of emotions is not two-dimensional. *Psychological Science, 18*(12).

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.

Frauenberger, C., Good, J., Alcorn, A., Pain, H. (2012). Supporting the Design Contributions of Children with Autism Spectrum Conditions. Proceedings of the 12th International Conference on Interaction Design and Children.

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

Graesser, A., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 285-290). Austin, TX: Cognitive Science Society.

Grimm, M., E. Mower, K. Kroschel, and S. Narayan. 2006. Combining Categorical and Primitives-Based Emotion Recognition. In 14th European Signal Processing Conference (EUSIPCO), Florence, Italy.

Isen, A. (2008). Some ways in which positive affect influences decision making and problem solving. In M. Lewis, J. Haviland-Jones & L. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 548-573). New York, NY: Guilford.

Izard, C. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review, 2*(4), 363-370. doi: 10.1177/1754073910374661

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding, 108*(1-2), 116-134. doi: 10.1016/j.cviu.2006.10.019

Järvenoja, H. and Järverlä, S. (2005). How students describe the source of their emotional and motivational experiences during the learning process: A qualitative approach, Learning and Instruction, vol. 15, pp.465-480.

Kaernbach, C. (2011). *On dimensions in emotion psychology.* Paper presented at the Proceesings of the 1st International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE) held in conjunction with the IEEE Automatic Face & Gesture Recognition Conference, Santa Barbara, CA.

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

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.

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., 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.

Litman, D. J., and K. Forbes-Riley. 2004. Predicting Student Emotions In Computer-Human Tutoring Dialogues. In Proceedings Of The 42nd Annual Meeting Of The Association For Computational Linguistics, East Stroudsburg, PA: 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

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.

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.

Matsumoto, D., Olide, A., Schug, J., Willingham, B., & Callan, M. (2009). Cross-cultural judgments of spontaneous facial expressions of emotion. *Journal of Nonverbal Behavior, 33*(4), 213-238. doi: 10.1007/s10919-009-0071-4

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

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.

Mills, C., & D'Mello, S. K. (2012). Emotions during writing on topics that align or misalign with personal beliefs. In S. Cerri, W. Clancey, G. Papadourakis & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems* (pp. 638-639). Berlin Heidelberg: Springer-Verlag.

Naab, P. J., & Russell, J. A. (2007). Judgments of emotion from spontaneous facial expressions of new guineans. [Article]. *Emotion, 7*(4), 736-744. doi: 10.1037/1528-3542.7a.736

Nielsen, P. A. (1991). Approaches to appreciate information systems methodologies: a soft system survey. Scandinavian Journal of Information Systems , Volume 2, University of Aalborg.

Ocumpaugh, J., Baker, R.S.J.d., Rodrigo, M.M.T. (2012) Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report. New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B., 36*(2), 433-449. doi: 10.1109/tsmcb.2005.859075

Pantic, M., & Rothkrantz, L. (2003). Toward an affect-sensitive multimodal human-computer interaction. [Review]. *Proceedings of the IEEE, 91*(9), 1370-1390. doi: 10.1109/jproc.2003.817122

Petta, P., Pelachaud, C., Cowie, R. (eds.): Emotion-Oriented Systems: The Humaine Handbook. Cognitive Systems Series, Springer Heidelberg / Dordrecht, 2011.

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.

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. and Mellish C.S (2013). Modelling human tutor's feedback to inform natural language interfaces for learning. *International Journal of Human - Computer Studies 71, pp. 703-724. Elsevier.*

Porayska-Pomsta, K., and Bernardini, S. (in preparation). Modelling Affect in TARDIS Social Skills Training Environment.

Read, J., McFarlane, S., and Cassey, C. (2002). Endurability, engagement and expectations: Measuring children's fun. In *Proceedings of International Conference for Interaction Design and Children*.

Read J. C. and MacFarlane, S.(2006). Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 conference on Interaction design and children* (IDC '06). ACM, New York, NY, USA, 81-88.

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.

Robson C.1993. *Real word research: A resource for social scientist and practitioner researchers*. Oxford: Blackwell.

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.

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.

Rosenberg, E. (1998). Levels of analysis and the organization of affect. *Review of General Psychology, 2*(3), 247-270. doi: 10.1037//1089-2680.2.3.247

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

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.

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. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145-172.

Russell, S. and Norvig, P. (1995). Artificial Intelligence: A Modern Approach. Prentice Hall.

Russo, J.E., Johnson, E.J. and Stephens, M.L. (1989) 'The validity of verbal protocols', Memory and Cognition, 17:759-69

Sabourin, J., Mott, B., & Lester, J. (2011). Modeling learner affect with theoretically grounded dynamic bayesian networks In S. D'Mello, A. Graesser, B. Schuller & J. Martin (Eds.), *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction* (pp. 286-295). Berlin Heidelberg: Springer-Verlag.

Safyan, L. and Lagattuta, K. H. (2008). Grown ups 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.

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.

Sazzad, M. S., AlZoubi, O., Calvo, R. A., & D'Mello, S. K. (2011). Affect detection from multichannel physiology during learning. In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 131-138). New York / Heidelberg: Springer.

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

Schröder, M (ed.), 2010: http://www.w3.org/TR/emotionml/

Schwarz, N. (in press). Feelings-as-Information Theory. In P. Van Lange, A. Kruglanski & T. Higgins (Eds.), *Handbook of theories of social psychology*: Sage.

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)*.

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18*(2), 158.

Sloetjes, H. and Wittenburg, P (2008). Annotation by category: Elan and iso dcr. In *LREC*. European Language Resources Association.

Stemmler, G., Heldmann, M., Pauls, C., & Scherer, T. (2001). Constraints for emotion specificity in fear and anger: The context counts. *Psychophysiology, 38*(2), 275-291.

Strain, A., & D'Mello, S. (in review). Cognitive Reappraisal to Alleviate Boredom and Disengagement during Learning.

Strain, A., & D'Mello, S. (2011). Emotion regulation during learning. In S. Bull & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 566-568). New York / Heidelberg: Springer.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859. doi: 10.1037/0033-2909.133.5.859

Valstar, M. F., Mehu, M., Jiang, B., Pantic, M., Scherer, K., Jiang, B., Valstar, M., Pantic, M., Valstar, M., & Jiang, B. (in press). Meta-Analysis of the First Facial Expression Recognition Challenge. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics*.

de Vicente, A., Pain, H. (2002) Informing the detection of the students' motivational state: an empirical study**.** In S. A. Cerri, G. Gouarderes, F. Paraguacu, editors, *Proceedings of the* Sixth International Conference on Intelligent Tutoring Systems, volume 2363 of Lecture Notes in Computer Science, pages 933-943, Berlin. Heidelberg. Springer.

de Vicente, A. (2003) Towards tutoring systems that detect students' motivation: an investigation. Ph.D. thesis, School of Informatics, University of Edinburgh, UK.

Florian Waszak, Bernhard Hommel, and Alan Allport, "Task switching and Long-term Priming: Role of Episodic Stimulus- task Bindings in Task-shift Costs," Cognitive Psychology 46, no. 4 (June 2003): 361–413.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: Recognizing and responding to student affect. *International Journal of Learning Technology, 4*(3/4), 129-163.

Woolf, B. (2008). Building Interactive Intelligent Tutors. Morgan Kaufman.

Wosnitza, M. and Volet, S. (2005). Origin, direction and impact of emotions in social online learning. Learning and Instruction, vol.15, pp. 449-464.

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