The final publication is available at Springer via
http://dx.doi.org/10.1007/s40593-015-0059-7

Adapting Progress Feedback and Emotional Support to Learner Personality

Matt Dennis, Judith Masthoff and Chris Mellish *University of Aberdeen, UK*
m.dennis, j.masthoff,c.mellish@abdn.ac.uk

Abstract. As feedback is an important part of learning and motivation, we investigate how to adapt the feedback of a conversational agent to learner personality (as well as to learner performance, as we expect an interaction effect between personality and performance on feedback). We investigate two aspects of feedback. Firstly, we investigate whether the conversational agent should employ a slant (or bias) in its feedback on particular test scores to motivate a learner with a particular personality trait more effectively (for example, using "you are slightly below expectations" versus "you are substantially below expectations" depending on learner conscientiousness). Secondly, we investigate which emotional support messages the conversational agent should use (for example: using praise, emotional reflection, reassurance or advice) given learner personality and performance.

We investigate the adaptation of this feedback to a learner personality, in particular the traits in the Five Factor Model. Five experiments were run where participants gave progress feedback and emotional support to students with different personalities and test scores. The type of emotional support given varied between different personalities (e.g. neurotic individuals with poor grades received more emotional reflection). Two algorithms were created using different methods to describe the adaptations and evaluated on how well they described the experimental data using DICE scores. A refined algorithm was created based on the results. Finally, we ran a qualitative study with teachers to investigate the algorithm's effectiveness and further refine the algorithm.

Keywords. Feedback, Personality, Emotional Support, Motivation, Adaptation

1 INTRODUCTION

Keeping learner motivation high is a key challenge in digital educational systems, with the lack of personalized approaches traditionally delivered by human tutors increasing drop out rates. As also reported in this journal, there has been considerable research in developing Intelligent Tutoring Systems and Adaptive Learning Environments which intelligently adapt the learning environment to learner characteristics (e.g. Wenger, 1987; Cerri et al., 2012; Lane et al., 2013b). Common learner characteristics for this adaptation include affective state (Nkambou, 2006; Woolf et al., 2009), motivational state (McQuiggan et al., 2008), learning styles (El-Bishouty et al., 2014), learner skills (Desmarais and Baker, 2012) and performance on a task (Varnosfadrani and Basturkmen, 2009). This paper investigates adaptation to an under-explored learner characteristic, namely personality, in particular the traits in the Five Factor Model (also known as the Big 5).

Table 1
Four examples of learners

Matt procrastinates and wastes his time. He finds it difficult to get down to work. He does just enough work to get by and often doesn't see things through, leaving them unfinished. He shirks his duties and messes things up. He doesn't put his mind on the task at hand and needs a push to get started. Matt tends to enjoy talking with people.	Peter is always prepared. He gets tasks done right away, paying attention to detail. He makes plans and sticks to them and carries them out. He completes tasks successfully, doing things according to a plan. He is exacting in his work; he finishes what he starts. Peter is quite a nice person, tends to enjoy talking with people, and quite likes exploring new ideas.
David often feels sad, and dislikes the way he is. He is often down in the dumps and suffers from frequent mood swings. He is often filled with doubts about things and is easily threatened. He gets stressed out easily, fearing the worst. He panics easily and worries about things. David is quite a nice person who tends to enjoy talking with people and tends to do his work.	Andrew seldom feels sad and is comfortable with himself. He rarely gets irritated, is not easily bothered by things and he is relaxed most of the time. He is not easily frustrated and seldom gets angry with himself. He remains calm under pressure and rarely loses his composure.

As feedback is an important part of learning and motivation (Deci and Ryan, 1980), we investigate how a conversational agent could adapt feedback to learner personality and performance.

Consider the learners in Table 1 – suppose they have all achieved the same score on a test. Would you praise them? Obviously, this would depend on the score they achieved. Suppose the pass mark is 50%. Would you provide praise if the learner scored 90%? Most likely, yes. However, what if the learner only achieved 55%? We believe that the answer to this question depends on the circumstances surrounding the learner. It may depend on how much effort the learner has put in[1], how you believe they will be feeling about the score, and how they may react to feedback. We postulate that teachers use the learner's personality as a guide for both the effort they put in (e.g. conscientiousness; compare example learners 'Matt' and 'Peter' in Table 1) and how they would respond to their progress (e.g. neuroticism; compare 'David' and 'Andrew'). For example, if a student is highly conscientious (such as 'Peter') then the teacher may assume that have put in considerable effort and praise them even if they achieved a bare pass. If the student is highly neurotic, then they may be easily upset about a failing score and may need more encouragement than an emotionally stable learner (such as 'Andrew').

In this paper, we investigate two aspects of feedback. Firstly, we investigate whether the conversational agent should employ a slant (or bias) in its feedback on particular test scores to motivate a learner with a particular personality trait more effectively (e.g., using "you are slightly below expectations" versus "you are substantially below expectations" depending on learner conscientiousness). Secondly, we investigate which emotional support messages the conversational agent should use (e.g., using praise, emotional reflection, reassurance or advice) given learner personality and performance. Table 2 shows examples of feedback. The empirical research will lead to an algorithm for providing tailored feedback.

The paper is organized as follows. Section 2 describes the background to the research and the related work. Section 3 introduces the methodology used in the studies. This includes a summary of

[1]It could also depend on the learner's past performance or learner goal. This is beyond the scope of this paper, but it will be discussed under future work.

Table 2
Examples of different types of feedback on performance

Feedback Example	**Emotional Support used**	**Slant**
You are meeting my expectations on topic A. I am proud of you, just keep practising.	praise & advice	neutral if score 55%
You are meeting my expectations on topic A. Just keep practising – you will get the hang of it eventually.	advice & reassurance	neutral if score 55%
You are below my expectations on topic A. I understand that you may be upset, you will get the hang of it eventually, just keep practising.	emotional reflection, reassurance & advice	neutral if score 30%
You are substantially below my expectations on topic A. You will get the hang of it eventually, just keep practising.	reassurance, advice	negative if score 30%

how personality stories were constructed and validated, how slant was defined and validated, and how emotional support categories and statements were produced and validated. Sections 4-8 present five studies, one for each factor of the Five Factor Model. This results in empirical insights on what feedback (slant and emotional support) people use when providing feedback to a learner with a certain personality and a certain performance. Section 9 describes the creation of an algorithm to produce the adaptations in feedback discovered to through the studies and an initial evaluation of the algorithm in terms of DICE scores. Section 10 presents a qualitative study with teachers and trainee teachers to evaluate whether the adaptations produced by the algorithm are appropriate and to further refine the algorithm. Section 11 concludes the paper, discusses its limitations and presents future work.

2 BACKGROUND AND RELATED WORK

The overarching goal of our research is to tailor feedback to learner personality and performance in order to motivate learners to study. Motivation is a complicated concept, so we first wanted to establish through a literature review how it relates to personality, and how it can be influenced through feedback. The Intelligent Tutoring Systems community have started investigating the relationships between feedback types, momentary confidence and learning (Boyer et al., 2008; Calvo and Ellis, 2010). However, they do not provide a fine grained model of the relevant factors. This section first discusses the concepts of motivation, personality, affective state, empathic support and performance feedback, and the relationship between them (as investigated through a literature survey in (Dennis et al., 2012c) and summarized in Figure 1). Next, it discusses related work on adaptive learning systems.

2.1 Motivation

Motivation is defined as "the process that energizes and/or maintains a behaviour"[2]. Many theories of motivation exist (see Graham and Weiner (1996) for an overview). Table 3 shows relevant modern theories of motivation. There are also several theories specifically about learner motivation. For example, the ARCS model identifies four factors that facilitate motivation, namely the learner's attention, the

[2]AllPsych Dictionary: http://allpsych.com/dictionary/m.html

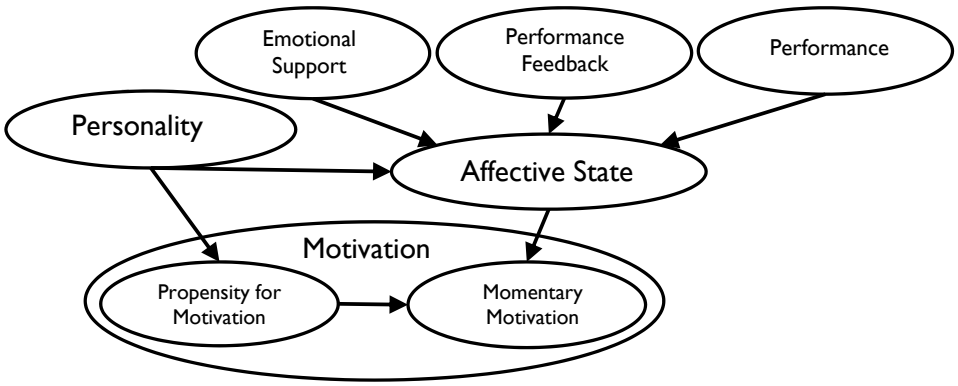

Fig.1. Relationships between concepts

relevance of instruction to the learner (goals, past experience, learning styles), the learner's confidence (including self-efficacy and attribution theory) and satisfaction (establishing a positive feeling towards the learning experience) (Keller and Suzuki, 2004). A distinction is often made between intrinsic (coming from oneself; doing things for personal enjoyment), and extrinsic (coming from external sources, such as the need to pass) motivation (Ryan and Deci, 2000). It has been argued that people have an evolved propensity for intrinsic motivation (Ryan et al., 1997). We will distinguish between the propensity for motivation (motivation in general) and momentary motivation (motivation at a particular moment).

Motivation is an important part of the learning process- as without the drive to achieve, students may fail to complete tasks and lose the will to learn new skills. If a learner experiences high levels of motivation, they are likely to perform better at academic tasks, and also experience positive affective states while learning (Zakharov et al., 2008). Blanchard and Frasson (2004) found that motivation was important for successful learning, and that a lack of motivation would have negative emotional impacts.

2.2 Personality

Chambers Dictionary defines Personality as "a person's nature or disposition; the qualities that give one's character individuality"[3]. Personality is a complex area, which has led to the development of theories of personality to aid understanding of oneself and others including Psychoanalytic, Neopsychoanalytic, Trait, Life-span, Humanistic, Cognitive, Behavioural, Social Learning and Limited-Domain theories (Nunes, 2008). In this paper, we focus on Trait theories.[4].

Personality traits can be defined as "relatively enduring patterns of thoughts, feelings and behaviours that represent a readiness to respond in particular ways to specific environmental cues" (Fayard et al.,

[3]http://www.chambers.co.uk

[4]We also considered the Social Learning models of Locus of Control (Rotter, 1966) and Self-Efficacy (Bandura, 1994). However, Locus of Control seemed most useful when adapting feedback that compares the learner's performance to that of others, and Self-Efficacy has been shown to be a subfacet of Emotional Stability

Table 3
Contemporary Motivational Theories and Constructs

Theory / Construct	Description
Expectancy / Value	E/V theories advocate increasing the expectancy for success, which will raise the value of the task, increasing motivation. For example, task value theory aims to explain why individuals do different tasks. This can be defined as four motivational components of task value: Attainment Value: The personal importance of doing well on a task; Intrinsic Value: the enjoyment one gets from doing the activity; Utility Value: How well the task relates to current and future goals; Cost: How doing one task affects other possible tasks (see (Wigfield et al., 2008)). These self-concepts and expectancies have been shown to reliably predict performance in mathematics, English and sport activities (Wigfield et al., 2008).
Attribution	Attribution theory gets the learner to focus on effort and controllable causes. Weiner's Attribution Theory (Weiner, 1985) states that an individual's causal attributions of achievement affect subsequent behaviours and motivation. A primary assumption is that people interpret their environment in such a way as to maintain a positive self-image .
Goal Orientation	Goal theories argue that setting reasonable goals and encouraging mastery increases motivation.
Intrinsic Motivation	Intrinsic Motivation to complete a task comes from oneself; such as a person who voluntarily masters a skill. Extrinsic motivation is motivation coming from an external source, such as pressure from parents to pass an exam. (Deci and Ryan, 1985) found that rewards for a task were usually perceived as controlling, and lead to a decrease in intrinsic motivation. They argue that people have a built in tendency to become competent at tasks.
Self-Determination (SDT)	Building on Intrinsic Motivation theory, SDT describes behaviour that originates from the individual themselves (Wigfield et al., 2008). Deci and Ryan (2002) state that there are three basic needs: The need for competence, for autonomy, for relatedness. They developed a taxonomy to describe the different stages of motivation experienced when transitioning from being externally to internally motivated. They investigate how external rewards may undermine intrinsic motivation, and title this cognitive evaluation theory. There is debate about when this occurs, but convincing evidence that this does occur in the real world (Wigfield et al., 2008).
Flow	Flow is defined as the immediate subjective experience that occurs when engaged in an activity. Characteristics of being *in flow* are: Feeling of being immersed in an activity; Merging of action and awareness; Focus of attention on a limited stimulus field; Lack of Self-Consciousness; Feeling in control of one's actions and the environment (Csikszentmihalyi, 1988).
Self-Efficacy	Self-Efficacy is a person's belief in how well they can complete a task. This has been argued to be one of the main determinants of motivation and performance (Bandura, 2012).
Control	Rotter (1966) distinguishes between internal or external locus of control, where individuals with an internal locus of control believe they can control events and consequently feel responsible; whereas those with an external locus of control believe they cannot control much and may proportion blame onto others. Later, control beliefs were categorised into a model (e.g. Wigfield et al. (2008)): Strategy Beliefs: particular causes produce a certain outcome; Control Beliefs: expectations that individuals have about how they can produce desired events and avoid undesired ones; and Capacity Beliefs: the belief that an individual has about whether they have access to the materials they need to produce outcomes.
Social Cognitive	In social cognitive theory, the self (thoughts, feelings and actions) plays a key role in human behaviour. It has been applied to learning, borrowing constructs such as Self-Efficacy, Locus of Control and Self-Worth (Schunk and Ertmer, 1999).

2012). Over the past century, trait theorists have sought to identify and categorise these traits in various ways, resulting in many trait models. The number of traits identified has varied over the years, with several competing theories. Some of the most well known are Eysenck's three factors (Eysenck, 2013), Cattell's 16PF (Cattell, 1957), and the Five-Factor Model (Goldberg, 1993). However, in recent years there has been a general trend towards Five main traits (or dimensions) (e.g. Digman, 1990).

We have adopted the Five Factor Model, as it is considered robust by most psychologists (Magai and McFadden, 1995), and a 45 year study found that the levels of the traits in individuals remained relatively stable (Soldz and Vaillant, 1999). See Digman (1990) on how the most popular trait models can be mapped to the Five-Factor Model. Psychologists may generally agree that there are five traits, however they do not agree on their names. In this paper, we adopt the nomenclature and definitions from John and Srivastava (1999) and refer to them as follows:

- Extraversion (I): How talkative, assertive and energetic a person is.
- Agreeableness (II): How good natured, cooperative and trustful a person is.
- Conscientiousness (III): How orderly, responsible and dependable a person is.
- Emotional Stability *(vs neuroticism)* (IV): How calm, non-neurotic and imperturable a person is.
- Openness to Experience (V): How intellectual, imaginative and independent-minded a person is.

The Five-Factor model has several advantages for use in our research. In addition to having broad consensus amongst psychologists, it provides a relatively simple way of identifying the personality of an individual as there are several validated questionnaires which measure it, with varying numbers of items from as few as 10 to 300 (e.g. Gosling et al., 2003; Goldberg et al., 2006).

The link between personality and propensity for motivation is well documented. Colquitt and Simmering (1998) examined the relationship between conscientiousness and motivation to learn, and found that conscientiousness was linked to Self-Efficacy and learner motivation. In further research, Major et al. (2006) found that Extraversion, Openness to Experience, Conscientiousness and Proactivity are good predictors of motivation to learn.

According to (Fayard et al., 2012), emotions can constitute part of personality traits. As such, there has been considerable research on the link between personality and emotions, with the link between Extraversion and positive affect, and Neuroticism and negative affect being well established (Rusting and Larsen, 1997). Conscientiousness is linked to both positive and negative affect: it is related to the positive affect facet attentiveness (Watson et al., 1998) and the negative facet guilt (with conscientious individuals likely to feel guilt when failing to meet their aspirations) (Fayard et al., 2012). Conscientious individuals also become more stressed when they receive negative feedback on their performance as they are more ambitious (Cianci et al., 2010). Personality can be used as a predictor of affective state, when coupled with performance related to a goal (Robison et al., 2010).

2.3 Affective State and Momentary Motivation

Although personality provides an indication of how likely somebody is to become motivated, whether they are feeling motivated at any given time (so, momentary motivation) also depends on the individual's affective state (which as discussed above, is influenced by personality). Earlier research treated

affective state as an output after motivation (e.g. (Weiner, 1985)), and affective state does not form part of many motivational constructs (Meyer and Turner, 2002). However, more recent research has found that affective state is important as an input to motivation; it contributes to the establishment of goals and self-efficacy (Turner et al., 1998). Positive affective states, such as interest, also increase motivation in learners (Meyer and Turner, 2002) and negative affective states such as anger and anxiety can reduce motivation (Assor et al., 2005). Effective motivational strategies need to consider affect (Meyer and Turner, 2002). The relationship between affective state and momentary motivation is complex. In general, positive affective states are seen as conducive to motivation, however certain negative emotions, in particular guilt, can increase motivation (Fayard et al., 2012). Propensity for motivation however can outweigh this relationship (Fayard et al., 2012).

2.4 Emotional Support

The Psychology Dictionary defines emotional support as: "the reassurance, encouragement and understanding we give . . . to a person"[5]. Emotional support has been described as messages or actions assuring an individual they are loved, cared for, esteemed and valued (Cobb, 1976). Emotional support is important in learning (Meyer and Turner, 2002) as it can encourage and reassure a learner and support their well-being when faced with negative affect. Table 4 shows examples of emotional support types used by other researchers. Statements can provide reassurance ("don't worry"), encouragement ("you can do it") and praise ("good job"). They can also empathise with emotional state ("I know you may be feeling anxious") and provide perspective ("yes, this topic is difficult").

Several researchers have built systems which provide emotional support to improve learning (e.g. Robison et al. (2009a)), and reduce negative affect (e.g. Prendinger and Ishizuka (2005); Nguyen and Masthoff (2009); Klein et al. (2002)). Studies have shown that providing emotional support leads to increased user satisfaction and system likeability (e.g Brave et al. (2005); Paiva et al. (2004); Nguyen and Masthoff (2009); Prendinger and Ishizuka (2005); Klein et al. (2002)) and can indeed impact emotions.

2.5 Performance Feedback

Performance is how well the learner has performed a learning related task. There are three common types of feedback in Intelligent Tutoring Systems literature: Empathic Feedback, Task-Based Feedback, and Progress-Based feedback (Robison et al., 2009b; Jackson and Graesser, 2007). Empathic feedback is a type of Emotional Support and has been discussed above. Task-based feedback gives practical and domain specific advice on how to complete an activity (e.g. hints on a maths quiz) or how to avoid mistakes made next time. Progress feedback is an assessment provided to a learner on their advancement (Jackson and Graesser, 2007), and reflects on their performance compared to teacher expectations.

Feedback can be slanted by emphasising good or bad aspects of the learner's performance. This can be achieved by omitting items (e.g., "you are below expectations on A and above expectations on B" versus "you are below expectations on A"), and by changing the phrasing (e.g., "you are substantially below expectations" versus "you are slightly below expectations"). There have been some studies on

[5] http://psychologydictionary.org/emotional-support/

Table 4
Examples of types of Emotional Support

Author	Empathy	Praise	Advice	Reassurance	Encouragement	Other
Barbee et al. (1993)	solace		solve			
Brave et al. (2005)	empathy					
Burleson and Picard (2007)	mirroring, affect					
Cutrona and Russell (1990)	concern	love				interest, care
Cutrona (1996)	nuturant	esteem	problem solving			
Fogg and Nass (1997)		praise				
Gilliland (2011)		praise	explaining	reassurance	encouragement	
Hone (2006)	affect support					
Johnson et al. (2004)		praise				
Klein et al. (2002)	affect support					
Lee (2008)		flattery				
Masthoff (1997)		praise				
Nguyen and Masthoff (2009)	sympathy, empathy					perspective
Paiva et al. (2004)	empathy					
Picard and Klein (2002)	sympathy, empathy					
Prendinger et al. (2004)	empathy					
Prendinger and Ishizuka (2005)	empathy	congratulate			encouragement	
Robison et al. (2010)	parallel empathy		reactive empathy			
Rook and Underwood (2000)		appreciate, respect		reassurance	encouragement	

the language used in feedback. Subtle variations in language can result in differences in self-reported affective state (van der Sluis and Mellish, 2010), with positive phrasing of feedback after an IQ test having a beneficial effect on study participants' affective state. Wang et al. (2008) describe a study where learners received feedback that was "polite" appealing to the desire for approval (positive face), or direct. The learners that received the "polite" feedback reported higher learning gains than those with the direct feedback, and the authors highlight the need to focus on the social-intelligence of agents, rather than just their appearance. Porayska-Pomsta and Mellish (2013) describe the modelling of feedback by human tutors to inform a Natural Language Generation system which can provide politeness in feedback as well as feedback tailored to the learner's immediate situation. Feedback can be given during or after a task. In this paper, we focus on Emotional Support and Progress feedback given after a task.

2.6 Adaptation of Learning Systems to Motivation and Affective State

There has been considerable research on adapting learning systems to learner motivation and in particular affective state. For example, Autotutor (D'Mello et al., 2008) is a sophisticated intelligent tutoring

system which can recognise moods such as boredom, confusion, frustration and engagement. Autotutor interacts with learners in two ways – via natural language dialogues and an embodied pedagogical agent which expresses affective responses to the learner. Autotutor aims to change negative affective states into positive ones which promote more engagement in learning using feedback tailored to emotions. As feedback, the embodied pedagogical agent can produce surprise, delight, disappointment, compassion and skepticism. Autotutor uses many physiological measurements to establish learner affect, and has shown considerable success in using learning tactics from expert tutors to improve learning gains.

Robison et al. (2009b) describe a methodology to decide whether to use task based (hints on how to proceed with the task) or affect based feedback (feedback depending on the affective state of the learner). They found that affect-based feedback was often rated as helpful, and that learners often benefit more from affect-based feedback than hints and reminders.

VanLehn et al. (2014) describe the development of an affective intervention using a learning companion. The learner's behavioural categories after a task were modelled, and an affective message given afterwards. For example, if a learner was suspected of gaming the system, the companion would respond with messages such as "It seems that you need to put quality time into your tasks. Maybe trial-and-error is not always the best strategy". The system's performance shows promising accuracy at predicting learner behaviour. However, the effects of the intervention on learners are yet to be investigated.

Lane et al. (2013a) describe 'Mike' an animated conversational agent which is placed in a science museum and teaches children to program. They describe an experiment where Mike's feedback on learner progress was designed to increase the learner's self efficacy at programming a robot. Mike had two conditions: enthusiastic and cold. In the enthusiastic condition, Mike offered praise in feedback such as "I am so impressed" when the learner performed a correct action, and offered more human like hints ("think about what you do when you turn around" vs "the robot needs to turn left and right"). A small improvement in self-efficacy was found in the condition where Mike provided enthusiastic feedback.

Martin et al. (2011) describe the development of the SIgBLE framework which aims to provide adaptable feedback for teachers, learner and learning environments (actors), with the goal of improving learning outcomes. The main objectives are to detect failure and success from the actions of the actors and provide reliable, adaptable feedback for each actor. Presently, their work has focussed on deducing learner characteristics (in the SIgBLE prototype), which teachers can use when deciding on feedback.

Arroyo et al. (2014) describes a system that identified and targeted affective states, additionally using measures of progress and motivation to improve student cognition, engagement and affect.

Outside of learning, Prendinger and Ishizuka (2005) developed an empathic companion to help with the preparation of a job interview. The agent was able to interpret the affective state of users by measuring physiological data. It then provided empathic support with the goal of mitigating negative affective states. Whilst an overall effect was not found, the empathic support did succeed in reducing stress when the user was being asked interviewer questions.

2.7 Adaptation of Learning Systems to Personality

There has also been some work on adapting to learner personality. Firstly, there is research on adapting motivational tactics. Del Soldato and Du Boulay (1995) describe a motivational planner which is able

to react and use motivational tactics. They describe how to model confidence and effort in learners, and how to select different tactics based on these levels. For example, if a learner has completed a task, then the planner will select a tactic to increase confidence. However, if the learner has given up, then the planner will pick a tactic to increase effort.

Secondly, there is research on adapting tutoring materials to learner personality and learning styles. For example, Beal and Lee (2005) describe the creation of a pedagogical model that adapts instruction (problem selection, problem difficulty, topic area, choice of activity, choice of help type, and availability of help) to learner motivation (including Self-Efficacy and domain specific criteria) and mood. El-Bishouty et al. (2014) developed a smart e-course recommender tool which analyses online courses and recommends learning objects which improves the support level for those with different learning styles.

Finally, there has been research into the role that personality plays when different types of feedback are given to learners. Robison et al. (2010) investigated whether personality could be used as a predictor of the effect of different types of feedback on learner affect. In their experiment, learners played an interactive game to solve a mystery about the cause of an outbreak of disease on a fictional island. When they were asked a question, a virtual agent appeared and asked them how they were feeling. Learners indicated this by completing a short self-report questionnaire, which measured the valence of nine emotions (*anger*, *anxiety*, *boredom*, *curiosity*, *confusion*, *delight*, *excitement*, *flow* and *frustration*). The agent then gave a randomized type of feedback to the learner, which was selected from *task-based* (e.g. "You might consider reading a book on pathogens, you can find a good book in the lab"), *parallel-empathic*, which gave feedback reflecting the learner's current affective state (e.g."I know it's frustrating not knowing what is causing the problem") or *reactive-empathic*, which aimed to place the learner in a more positive affective state (e.g. "I know this is a tough problem, but if you keep working at it, I'm sure you'll get an answer soon"). Learners then rated how effective they thought the feedback was using a scale, before taking a second self-report affect questionnaire. Robison et al. were thus able to measure the *transition* of affective state that the feedback had caused in the learner. As they also required the learners to take a personality test prior to the study, they were able to use this to correlate the transitions that occurred for each feedback type with their scores on the Five-Factor model. Personality had a significant effect on the transitions overall (for example, Conscientious learners were less bored after feedback), however the effects for each feedback type were not investigated in this paper.

3 STUDY DESIGN FOR INVESTIGATING HOW TO ADAPT EMOTIONAL SUPPORT AND SLANT IN FEEDBACK TO PERSONALITY AND PERFORMANCE

This paper investigates if those taking the role of a teacher adapt emotional support strategies in addition to slant based on a learner's personality and performance. As the Five-Factor model presents a complete model of personality, we performed five experiments with the same design, one for each of the traits, using *personality trait stories* to convey learner personality. This series of experiments utilize the *User-as-Wizard* approach (Masthoff, 2006), where participants take the role of a system giving feedback to a learner. This section describes the common design of the experiments, sections 4-8 present the results, section 9 provides an over-arching discussion of the results.

Table 5

Stories used for each of the traits, high and low.

Trait	**Level**	**Story**
Extraversion	low	Jack has little to say to others, preferring to stay in the background. He would describe his life experiences as somewhat dull. He doesn't like drawing attention to himself, and doesn't talk a lot. He avoids contact with others and is hard to get to know. He retreats from others, finding it difficult to approach them. He keeps people at a distance. Jack is quite a nice person.
	high	Jack feels comfortable around people and makes friends easily. He is skilled in handling social situations, and is the life and soul of the party. He knows how to start conversations and easily captivates his audience. He warms up quickly to others, and likes talking to a lot of different people at parties. He doesn't mind being the centre of attention and cheers people up. Jack can sometimes be insensitive.
Agreeableness	low	Charlie has a sharp tongue and cuts others to pieces. He suspects hidden motives in people. He holds grudges and gets back at others. He insults and contradicts people, believing he is better than them. He makes demands on others, and is out for his own personal gain. Charlie tends to be calm and quite likes exploring new ideas.
	high	Charlie has a good word for everyone, believing that they have good intentions. He respects others and accepts people as they are. He makes people feel at ease. He is concerned about others, and trusts what they say. He sympathizes with others' feelings, and treats everyone equally. He is easy to satisfy. Charlie tends to be quite anxious.
Conscientiousness	low	Josh procrastinates and wastes his time. He finds it difficult to get down to work. He does just enough work to get by and often doesn't see things through, leaving them unfinished. He shirks his duties and messes things up. He doesn't put his mind on the task at hand and needs a push to get started. Josh tends to enjoy talking with people.
	high	Josh is always prepared. He gets tasks done right away, paying attention to detail. He makes plans and sticks to them and carries them out. He completes tasks successfully, doing things according to a plan. He is exacting in his work; he finishes what he starts. Josh is quite a nice person, tends to enjoy talking with people, and quite likes exploring new ideas.
Emotional Stability	low	James often feels sad, and dislikes the way he is. He is often down in the dumps and suffers from frequent mood swings. He is often filled with doubts about things and is easily threatened. He gets stressed out easily, fearing the worst. He panics easily and worries about things. James is quite a nice person who tends to enjoy talking with people and tends to do his work.
	high	James seldom feels sad and is comfortable with himself. He rarely gets irritated, is not easily bothered by things and he is relaxed most of the time. He is not easily frustrated and seldom gets angry with himself. He remains calm under pressure and rarely loses his composure.
Openness	low	Oliver is not interested in abstract ideas, as he has difficulty understanding them. He does not like art, and dislikes going to art galleries. He avoids philosophical discussions. He tends to vote for conservative political candidates. He does not like poetry and rarely looks for a deeper meaning in things. He believes that too much tax money goes to supporting artists. He is not interested in theoretical discussions. Oliver is quite a nice person, and tends to enjoy talking with people.
	high	Oliver believes in the importance of art and has a vivid imagination. He tends to vote for liberal political candidates. He enjoys hearing new ideas and thinking about things. He enjoys wild flights of fantasy, getting excited by new ideas.

Table 6
Selected emotional support statements to be used in the series of studies

Cat.	Statements used
P	That was hard but you did it; I am proud of you; Well done
ER	I know what you're feeling; You must be really happy; I understand that you may be upset
R	Everyone is wrong sometimes; Everyone finds this hard; You will get the hang of it eventually
A	Just keep practising; Just take a bit longer next time; Just read the questions more carefully

3.1 Materials

Personality Stories. To vary learner personality, participants were given one of two stories about a fictional learner, (see Table 5), which described a Big 5 factor at a high or low level. By taking this approach, rather than a more simplistic one (such as "John is an extrovert"), we hoped to provide enough information for participants to identify and empathise with them. We adapted personality *Self-Report Questionnaires* (the IPIP-NEO scales from Goldberg et al. (2006)), to create short stories which describe one personality trait at a polarized level. To ensure that the stories expressed the trait that they were designed for, and did not inadvertently express another trait at the same time, we designed a validation experiment where participants saw the story about a learner, and rated the story using a validated personality questionnaire (the Mini-Markers scale, (Saucier, 1994)). If needed, the stories were adjusted and re-validated; Dennis et al. (2012b) describes the development and validation of the stories in detail.

Progress Feedback Options. In (Dennis et al., 2011), we developed a systematic way for participants to provide progress feedback. Participants can use a set of descriptions (above, below, meeting) and modifiers (slightly, substantially) to describe the learner's performance compared to their expectations, see Table 7.

Emotional Support Options. We provided participants with a range of different types of emotional support that they could use to support the learner (see Table 6). To establish these, we ran a brainstorming exercise which asked three teachers to generate as many examples of emotional support as they could. A card sorting exercise was performed to group the emotional support statements that were related together. These became a set of categories, which were labelled. A validation experiment was implemented which showed participants each emotional support statement in turn and asked them to place it into one of the categories. Statements which were not reliably categorised were discarded. This resulted in a set of categories of emotional support, each having a set of validated statements that were examples of them. See (Dennis et al., 2013) for details.

Score. The performance of the learner was conveyed by a percentage score that they had achieved on a mock test on Aromathy (a fake topic). There were six possible scores: a poor fail (10%), a fail (30%), a marginal fail (45%), a marginal pass (55%), a good pass (70%), and a strong pass (90%).

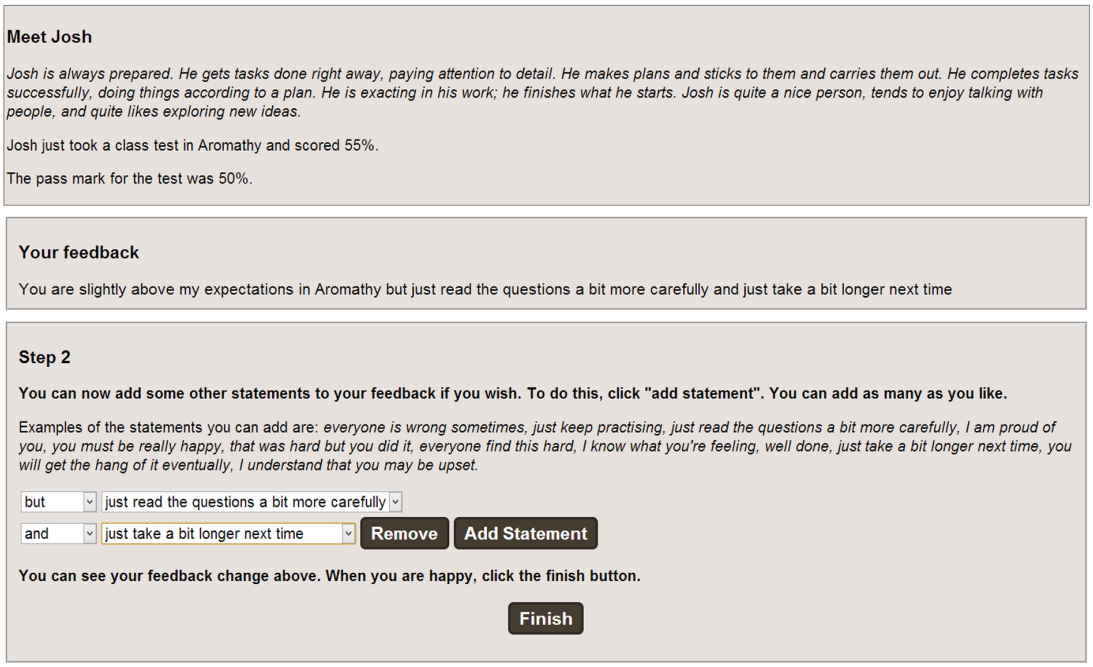

Fig.2. The interface for the main study. In this example, the participant has already given performance feedback in step one, and is now adding emotional support.

3.2 Experimental Procedure

We used a 6x2 between-subjects design; participants were shown a story about a learner conveying the personality trait (e.g. Conscientiousness) at high or low levels, and a score they had achieved on a test. Thus there were 12 variants of the study per trait, with participants only giving feedback on a single score. Participants were asked to give the learner progress feedback: whether he was above, meeting or below expectations. They could also use the two modifiers (substantially or slightly) if they wished.

Participants could opt to add emotional support statements (shown in Table 6). They could add as many of the statements as they wished, but they could only add each statement once. A conjunction (one of "and", "but", "however", a full stop, a semi-colon or a comma) could also be added between the statements to make the feedback statement flow more naturally. Participants were then shown their feedback paragraph, and given the opportunity to give any comments. Figure 2 shows the interface used.

3.3 Variables

There were two independent variables: The learner's personality *Trait Level*: high or low, and the *Score* the learner has achieved. We used six measures for the dependent variables: The number of Emotional Support statements used for each category (*Reassurance*, *Praise*, *Emotional Reflection* or *Advice*), the total number of Emotional Support statements (*NS*), and the *Slant* exhibited.

Table 7
The slant expressed by each description/modifier combination for each score.

Passing				Failing			
Score	Description	Modifier	Slant	Score	Description	Modifier	Slant
90%	above	substantially, none	neutral	45%	above	all	positive
		slightly	negative		meeting	n/a	positive
	meeting	n/a	negative		below	slightly, none	neutral
	below	all	negative			substantially	negative
70%	above	substantially	positive	30%	above	all	positive
		slightly, none	neutral		meeting	n/a	positive
	meeting	n/a	negative		below	slightly, none	neutral
	below	all	negative			substantially	negative
55%	above	substantially	positive	10%	above	all	positive
		slightly, none	neutral		meeting	n/a	positive
	meeting	n/a	neutral		below	slightly	positive
	below	all	negative			none, substantially	neutral

The slant is the result of progress feedback, and can be positive, negative or neutral. For example, if a score of 90% is described as 'slightly above' expectations then this constitutes a negative slant. To establish and validate which modifier and description combinations exhibited which type of slant, we conducted a focus group with three judges (who were teachers) and asked them to indicate which combinations exhibited a positive, negative or neutral slant (Dennis, 2014). This resulted in a set of rules for slant, for each modifier/description combination and score, shown in Table 7.

3.4 Participants

The experiments were administered as an online questionnaire on Amazon's Mechanical Turk (MT, 2012). Mechanical Turk allows the creators of tasks (*requesters*) to approve or reject completed work before payment. Mechanical Turk holds many statistics on each participant (*worker*), including their location and *acceptance rate*. The acceptance rate is a global statistic available to all *requesters* on Mechanical Turk. Thus if a worker consistently submits poor or incomplete work, their acceptance rate will drop. As requesters usually set a high acceptance rate as as requirement for their tasks, this causes workers to take their acceptance rate very seriously, and to complete the tasks to the best of their ability.

In our experiments, we included a Cloze Test for English Fluency that served as an attention check to ensure that workers were reading the instructions carefully, and possessed enough literacy skills to understand the language based nature of the task. Participants had to have an acceptance rate of 90%, be based in the United States and pass the fluency test. Each of the experiments described in this paper used unique participants, and Table 8 shows their demographic information.

Table 8
Participant demographics for each trait experiment (C = conscientiousness, ES = Emotional Stability, EXT = Extraversion, AGR = Agreeableness, OE = Openness to Experience).

trait	n	age				gender			occupation			avg time
		16-25	26-40	41-65	over 65	M	F	unknown	student	teacher	other	
C	242	34%	41%	24%	1%	46%	53%	1%	33%	7%	60%	2.5m
ES	240	45%	39%	15%	1%	63%	36%	1%	38%	5%	58%	2.5m
EXT	241	28%	47%	23%	2%	57%	42%	1%	23%	7%	70%	2.5m
AGR	240	21%	45%	31%	3%	52%	47%	1%	22%	7%	71%	2.5m
OE	240	40%	42%	17%	1%	59%	41%	0%	31%	5%	64%	2.4m

3.5 Analysis of Results

To analyse the results for each of the experiments, a 2-way MANOVA was used to test the effect of the two independendent variables: trait-level (high or low) and Score on each of the dependent variables[6]. Where a significant effect was found, we performed a post-hoc pairwise comparison to produce homogeneous subsets of score, and the interaction of trait-level $\times$ score on the dependent variables. All comparisons are Bonferroni corrected. For Emotional Support type, if the subset mean for a particular range of scores was ≥ 0.5, we recommended that type of emotional support to be used. For Slant, if the subset mean was ≤ -0.5, we recommended a negative slant be used. If the subset mean was ≥ 0.5, we recommended a positive slant be used. Otherwise, a neutral slant was recommended. To decide on the number of emotional support statements to recommend, we rounded the subset mean.

4 EXPERIMENT 1: CONSCIENTIOUSNESS

In this study, we examine the effect of learner Conscientiousness (III). Conscientiousness describes how efficient, organized, reliable and responsible an individual is (McCrae and John, 1992).

We would expect tutors to use Emotional Support when giving feedback to conscientious learners, when performance has been poor, to attempt to mitigate negative affective states such as stress and guilt.

4.1 Hypotheses

Based on the results of a previous experiment on progress feedback and conscientiousness described in (Dennis et al., 2012a), we expected the use of positive slant for a mark which was almost passing (around 45%) and a learner with high conscientiousness. However, the design of this study differs as participants only see one grade as opposed to a range, and can also provide emotional support.

We expect Praise to be used only for passing scores, as good work is usually rewarded, and praising failing scores may seem inappropriate, patronizing or sarcastic. We expect Emotional Reflection to be used when strong emotions are being experienced, so it seems likely that it will be used for the highest and lowest scores. We do not expect Advice to be used for the two highest scores, as the learner performed

[6]The dependent variables distribution were checked for normality, and the homogeneity of variance was checked to ensure that the assumptions for the MANOVA were met.

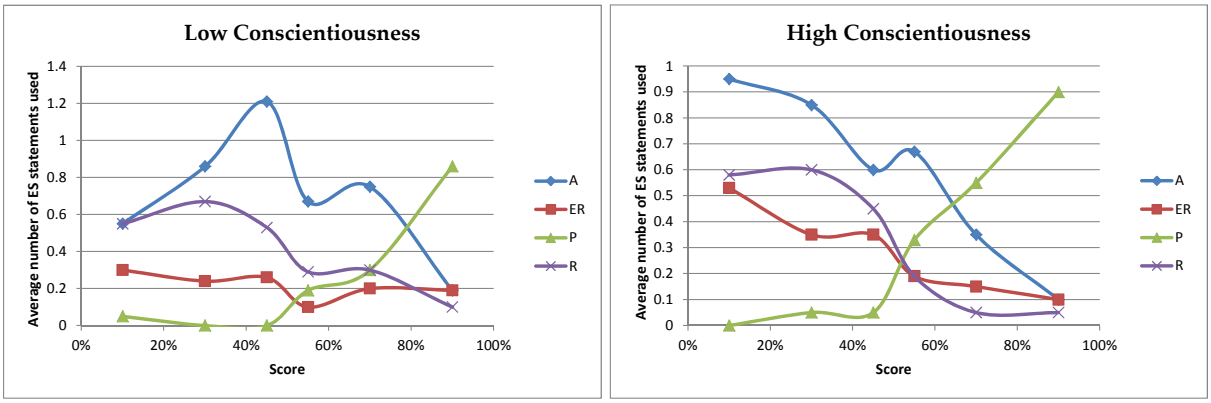

Fig.3. Average number of Emotional Support statements used, for high and low conscientiousness, as score increases.

well and further advice may not be required. We expect Reassurance to be used for the middle scores (30%, 45%, 55% and 70%), as reassuring a learner who scored 10% that they can do better is perhaps not true, and a score of 90% does not require it. Therefore the hypotheses were:

H1: The type of emotional support given will depend on the score the learner achieved

- **H1a:** Praise will only be used when the score is passing (>50%).
- **H1b:** Emotional Reflection will mainly be used for 10% or 90%.
- **H1c:** Advice will not be given for the very high scores (70% or 90%).
- **H1d:** Reassurance will be given for scores between 30% and 70%.

H2: The quantity of emotional support utilized will be higher for high learner conscientiousness

H3: The slant utilized will differ depending on the level of conscientiousness

- **H3a:** Slant will be positive for a 45% score and high conscientiousness

In addition to testing specific hypotheses, we investigated more in general which types of emotional support and slant participants gave for learners with high or low conscientiousness for the different scores, in order to provide recommendations for what a system should do.

4.2 Results

Figure 3 shows the different types of Emotional Support (A, ER, P, R) used per trait-level. Table 9 shows the significant effects of the 2-way MANOVA of trait-level $\times$ score on the use of *A*, *ER*, *P*, *R*, *Slant*, *NS*.

Effects of Score

The score had a significant effect on all of the different types of emotional support used and slant employed. Hypothesis H1 (the type of emotional support given will depend on the score the learner achieved) is thus confirmed.

Table 9

Significance values of 2-way MANOVA for Conscientiousness. Independent variables are the columns, and dependent variables are the rows. – indicates no significance. P = number of Praise statements, A = number of Advice statements, ER = number of Emotional Reflection statements, R = Number of Reassurance Statements, NS = Total number of Emotional Support Statements.

	score	*trait-level*	*trait-level* × *score*
A	$F(5, 230) = 7.43, p < .01$	–	$F(5, 230) = 2.90, p < .05$
ER	$F(5, 230) = 2.32, p < .05$	–	–
P	$F(5, 230) = 24.17, p < .01$	–	–
R	$F(5, 230) = 6.30, p < .01$	–	–
NS	–	–	–
Slant	$F(5, 230) = 7.48, p < .01$	$F(1, 230) = 18.72, p < .01$	$F(5, 230) = 4.73, p < .01$

Table 10

Conscientiousness: homogeneous subsets arising from a post-hoc pairwise comparison of score, and the interaction effect of trait-level × score, on the average number of each type of emotional support statement given (A, ER, P, R), slant, and mean total amount of emotional support given (NS).

	Effect of Score		Effect of trait-level x score			
			High		Low	
Variable	Scores (%) in subset	mean	Scores (%) in subset	mean	Scores (%) in subset	mean
A	10, 30, 45, 55, 70	0.74	10, 30, 45, 55	0.77	30, 45	1.04
	90	0.15	70, 90	0.23	10, 30, 55, 70	0.71
					10, 90	0.19
ER	10, 30, 45, 55, 70, 90	0.25	no effect			
P	90	0.88	no effect			
	55, 70	0.35				
	10, 30, 45, 55	0.09				
R	10, 30, 45	0.56	no effect			
	10, 45, 55	0.43				
	45, 55, 70	0.30				
	55, 70, 90	0.16				
Slant	10, 45, 90	-0.08	10, 30, 45, 55, 90	-0.08	10, 45	-0.06
	45, 55, 90	-0.17	30, 45, 55, 70, 90	-0.13	45, 50, 90	-0.23
	30,55,70,90	-0.31			30, 70	-0.66
NS	no effect		no effect			

The left hand columns in Table 10 show the results of the pairwise comparisons of the effect of score on the average number of each type of emotional support statement given, the slant used, and the average number of emotional support statements given. For Advice, a pairwise comparison indicates that there are two homogeneous subsets. These indicate that advice should be used for all scores except 90%. This partially supports hypothesis H1c, in that Advice is not used for 90%, however it is used for 70%. For Emotional Reflection, the significance was not powerful enough to generate more than one subset. None of the means are above 0.5, which indicates that ER was seldom used by participants. Thus, hypothesis H1b is not supported. For Praise, a pairwise comparison yields three subsets. Praise is

only used for passing scores, which supports H1a. However, whilst Praise is used for 90%, it is seldom used for the other passing scores (55% and 70%). For Reassurance, a pairwise comparison reveals four homogeneous subsets, with reassurance being given for failing scores. This partially supports H1d in that it was given for 30% and 45%, however reassurance was only used for the failing scores rather than all the middle scores and was also given for 10%, contrary to what we expected. For slant, a pairwise comparison shows 3 subsets. However, the means still indicate a neutral slant for all scores.

Effects of Trait-Level

Hypothesis H2 does not hold, the total number of emotional support statements used does not significantly differ between levels of conscientiousness. However, there is a significant effect of trait level on slant, supporting H3. The mean for high trait level is -0.11 and low trait level mean is -0.33. This indicates that there is more negative slant used for the low trait level, which is contrary to H3a which predicted a positive slant for learners with high conscientiousness. Unlike our previous findings in (Dennis et al., 2012a), there is no evidence for a positive slant being used.

Effects of Trait-Level $\times$ Score

The right-hand columns in Table 10 show the results of the pairwise comparisons of the effect of score x traitlevel on the average number of each type of emotional support statement given, the slant used, and the average number of emotional support statements given.

Interestingly, the MANOVA did show an interaction effect between the level of conscientiousness and the score achieved on the usage of advice and slant. For high conscientiousness, there are 2 homogeneous subsets for advice. This indicates that for scores between 10% and 55% inclusive, advice is used. For low conscientiousness, the pairwise comparison reveals three homogeneous subsets for Advice. This indicates that when conscientiousness is low, advice is used for 10% to 70% inclusive, which contrasts with high conscientiousness, where Advice is not used for 70%. This might be because participants thought that this learner had worked hard, and that adding advice was therefore less appropriate for these scores. Contrasting with low conscientiousness, they do offer advice at 70%, which may mean that participants thought that if the learner had achieved 70% with little assumed effort, they could gain a higher score if they worked harder. For high conscientiousness, 2 homogeneous subsets were found for slant, however a neutral slant was used for both groups. For low conscientiousness, there were 3 homogeneous subsets, with a negative slant being used for scores of 30% and 70%, and a neutral slant for all other scores. It seems likely that participants used negative slant to incentivise the learner to work harder.

5 EXPERIMENT 2: EMOTIONAL STABILITY

Emotional Stability (also frequently described as Neuroticism - see Section 2, Section 2.2) describes such personality features as moodiness, nervousness and temperamentality (Goldberg, 1993).

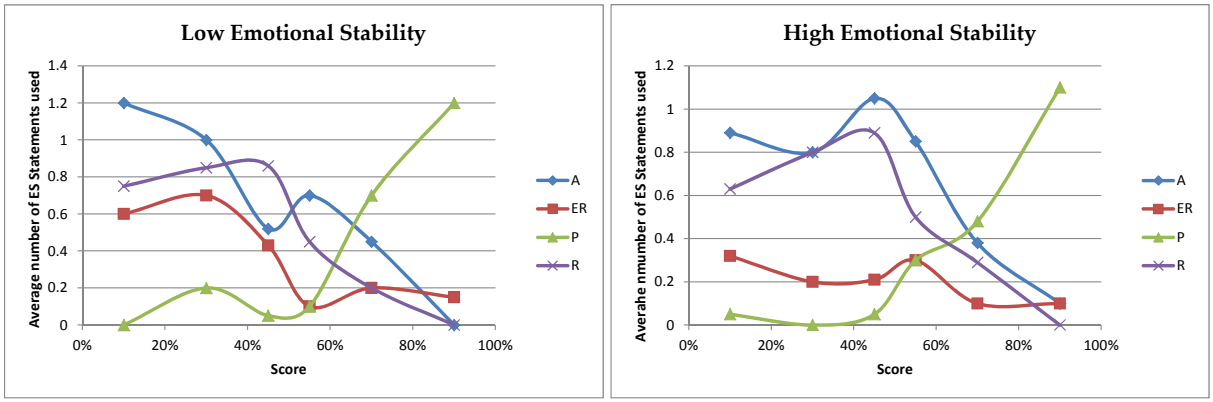

Fig.4. Average number of Emotional Support statements used, for high and low Emotional Stability, as score increases.

5.1 Hypotheses

As learners with low Emotional Stability (therefore being very neutrotic) are more likely to be nervous about their performance, we expect to see more Emotional Reflection being used, particularly when the scores are low. There may also be a tendency to praise neurotic learners when they pass, as opposed to it being reserved for the very high scores (like we saw in the Conscientiousness study). Overall, we expect this to lead to more Emotional Support being used for low Emotional Stability (H2). Conversely, a more emotionally stable learner would be able to accept criticism more readily. As such, there may be negative slanting used on poorer marks, and less Emotional Reflection used. Based on the results for the usage of Emotional Support on score from the previous study on Conscientiousness, we modified H1, as this concerns the use of Emotional Support on score only. H3a is based on the results of a previous study for Generalized Self-Efficacy (Dennis et al., 2011), as it is correlated with Neuroticism (Judge et al., 2002).

H1: The type of emotional support given will depend on the score the learner achieved

- **H1a:** Praise will only be used when the score is a good pass ($\geq$ 70%)
- **H1b:** Emotional Reflection will rarely be used
- **H1c:** Advice will not be given for the highest score (90%)
- **H1d:** Reassurance will be given for the failing scores (<50%)

H2: The quantity of emotional support utilized will be higher for low Emotional Stability

H3: The slant utilized will differ depending on the level of Emotional Stability

- **H3a** Slant will be positive for a 10% score and low Emotional Stability

5.2 Results

Figure 4 shows the different types of Emotional Support used per trait-level. Table 11 shows the significant effects of the 2-way MANOVA of trait-level $\times$ score on the use of *A*, *ER*, *P*, *R*, *Slant*, *NS*.

Table 11
Significance values of 2-way MANOVA for Emotional Stability. Independent variables are the columns, and dependent variables are the rows. – indicates no significance. P = number of Praise statements, A = number of Advice statements, ER = number of Emotional Reflection statements, R = number of Reassurance Statements, NS = Total number of Emotional Support Statements.

	score	*trait-level*	*trait-level* × *score*
Slant	$F(5, 228) = 3.72, p < .01$	$F(1, 228) = 4.28, p < .05$	–
A	$F(5, 228) = 14.68, p < .01$	–	$F(5, 228) = 2.42, p < .05$
ER	$F(5, 228) = 3.57, p < .01$	$F(1, 228) = 6.21, p < .02$	$F(5, 228) = 2.27, p < .05$
P	$F(5, 228) = 35.22, p < .01$	–	–
R	$F(5, 228) = 12.18, p < .01$	–	–
NS	$F(5, 228) = 4.61, p < .01$	–	$F(5, 228) = 2.30, p < .05$

Table 12
Emotional Stability: homogeneous subsets arising from a post-hoc pairwise comparison of score, and the interaction effect of trait-level × score, on the average number of each type of emotional support statement given (A, ER, P, R), slant, and mean total amount of emotional support given (NS).

	Effect of Score		Effect of trait-level x score			
			High		Low	
Variable	Scores (%) in subset	mean	Scores (%) in subset	mean	Scores (%) in subset	mean
A	10, 30, 45, 55	0.88	10, 30, 45, 55	0.90	10, 30	1.10
	70	0.41	70, 90	0.24	45, 55, 70	0.56
	90	0.05			90	0.00
ER	10, 30, 45, 55, 70	0.32	10, 30, 45, 55, 70, 90	0.21	10, 30, 45	0.58
	45, 55, 70, 90	0.20			45, 55, 70, 90	0.22
P	90	1.15	no effect			
	70	0.59				
	10, 30, 45, 55	0.10				
R	10, 30, 45	0.79	no effect			
	10, 30, 55	0.66				
	55, 70	0.36				
	70, 90	0.12				
Slant	10, 30, 45, 70, 90	-0.04	no effect			
	30, 45, 55, 70, 90	-0.11				
NS	10, 30, 45, 55	2.05	10, 30, 45, 55	1,97	10, 30	2.65
	45, 55, 70, 90	1.60	10, 30, 55, 70, 90	1.64	45, 55, 70, 90	1.53

Effects of Score

H1 is confirmed, as there is a significant effect of score on the use of Advice, Reassurance, Praise and Emotional Reflection. For Advice, a pairwise comparison (see Table 12) indicates that there are three homogeneous subsets. These indicate that advice should be used for all scores except 70% and 90%. This supports hypothesis H1c, in that Advice is not used for 90%, however, in contrast to what we found in the Conscientiousness study, Advice is now not used for a score of 70%, which would have been

in agreement with the stronger hypotheses we posed in the previous study. For Emotional Reflection, although there are two distinct homogeneous subsets, the subset mean is never above 0.5. This supports Hypothesis H1b in that Emotional Reflection seems to be rarely used. For Praise, a pairwise comparison yields three subsets, with Praise being used for 70% and 90%. This supports hypothesis H1a. For Reassurance, we hypothesized in H1d that it would be used for failing scores. The pairwise comparison supports this, but also indicates that Reassurance can be used for 55% additionally (though this is marginal). There was an unexpected effect of Score on the total number of Emotional Support statements given. The pairwise comparison indicates that the passing scores receive slightly less Emotional Support than the failing ones, however the subset average still recommends 2 statements overall for both passing and failing scores. There was also an effect of score on slant, the pairwise comparison reveals two homogeneous subsets, however the slant appears to be neutral for both.

Effects of Trait-Level

For slant, H3 expected that slant would differ based on the level of Emotional Stability. There was a significant effect for slant (shown in Table 11). The mean slant for high Emotional Stability was -0.13, and -0.02 for low Emotional Stability. Although High Emotional Stability has more negative slant than low, the means indicate that we should offer neutral slant for both, as neither is ≤ -0.5. Hypothesis H2 does not hold, the total number of emotional support statements used does not significantly differ between levels of Emotional Stability. There was also an effect for trait level on Emotional Reflection. The mean for high was 0.20 and low was 0.36 which indicates that leaners with low emotional stability receive slightly more ER overall. This is explored further in the next section, as there was also an interaction between trait level and score for the use of ER.

Effects of Trait-Level $\times$ Score

For high Emotional stability, the pairwise comparison suggests only one subset for Emotional Reflection, and does not recommend the use of this strategy. However, for low Emotional Stability, the pairwise comparison shown in indicates two subsets, with the mean for ER being ≥ 0.5 for failing scores. This may mean that participants felt it helpful to reflect on the stronger negative emotions that they expected learners with low Emotional Stability to be experiencing. There was also a significant effect of trait-level $\times$ score for Advice. For high Emotional stability, the pairwise comparison suggests two subsets for Advice, and Advice seems to be used for all but the top two scores (70% and 90%). However, when Emotional Stability is low, the pairwise comparison indicates three subsets, with Advice being used for all scores but 90% (though this is marginal for 70%). There was also an effect of trait-level $\times$ score on the total number of emotional support statements given. For high Emotional Stability, the pairwise comparison reveals two homogeneous subsets, however both recommend the use of two statements for all scores. For low Emotional Stability, the pairwise comparison indicates 2 homogeneous subsets, with scores 30% and 10% getting three statements and the remainder receiving two. Therefore participants may have wanted to give additional emotional support to learners whom they expected to be very upset.

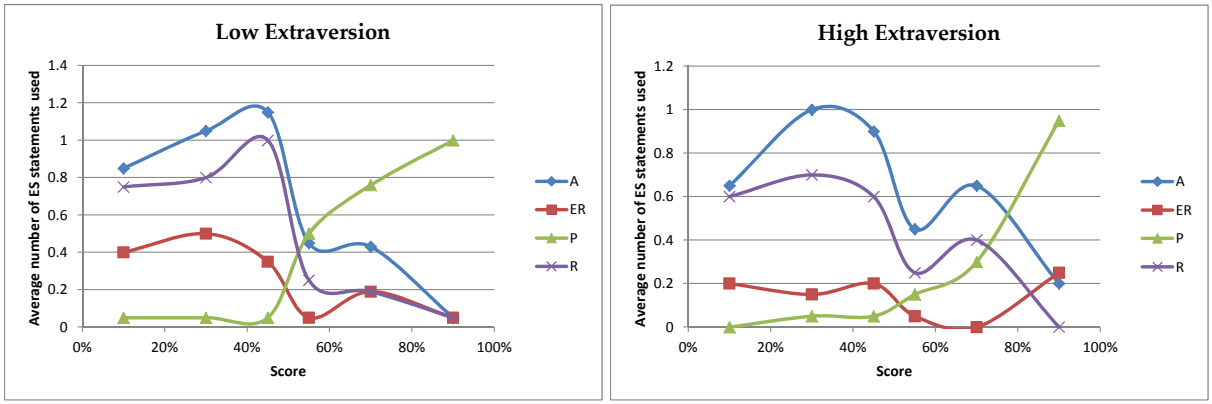

Fig.5. Average number of Emotional Support statements used, for high and low extraversion, as score increases.

6 EXPERIMENT 3: EXTRAVERSION

Extraversion describes how outgoing someone is and how comfortable they are in social situations (Costa and McCrae, 1992).

6.1 Hypotheses

We do not expect there to be great variance in the use of emotional support between the two levels of extraversion, perhaps Emotional Reflection would be used differently in this case.

H1: The type of emotional support given will depend on the score the learner achieved

- **H1a:** Praise will only be used when the score is a good pass ($\geq$ 70%)
- **H1b:** Emotional Reflection will rarely be used
- **H1c:** Advice will not be given for the highest score (90%)
- **H1d:** Reassurance will be given for the failing scores (<50%)

Although we do not expect there to be a difference based on the level of Extraversion, we left the hypothesis as is, as we can test this with the statistics:

H2: The quantity of emotional support utilized will differ depending on the level of extraversion.
H3: The slant utilized will differ depending on the level of Extraversion

6.2 Results

Figure 5 shows the use of the different types of Emotional Support (A, ER, P, R) per trait level. Table 13 shows the significant effects of the 2-way MANOVA of trait-level $\times$ score on the use of *A*, *ER*, *P*, *R*, *Slant* and *NS*. Table 14 show the results of the pairwise comparisons.

Table 13

Significance values of 2-way MANOVA for Extraversion. Independent variables are the columns, and dependent variables are the rows. – indicates no significance. P = number of Praise statements, A = number of Advice statements, ER = number of Emotional Reflection statements, R = number of Reassurance Statements, NS = Total number of Emotional Support Statements.

	score	*trait-level*	*trait-level* × *score*
Slant	$F(5, 229) = 2.72, p < .03$	–	–
A	$F(5, 229) = 10.90, p < .01$	–	–
ER	$F(5, 229) = 3.16, p < .01$	$F(1, 229) = 4.67, p < .04$	–
P	$F(5, 229) = 25.27, p < .01$	$F(1, 229) = 6.22, p < .02$	–
R	$F(5, 229) = 11.72, p < .01$	–	–
NS	$F(5, 229) = 5.59, p < .01$	$F(1, 229) = 5.67, p < .02$	–

Table 14

Extraversion: homogeneous subsets arising from a post-hoc pairwise comparison of score, and the interaction effect of trait-level × score, on the average number of each type of emotional support statement given (A, ER, P, R), slant, and mean total amount of emotional support given (NS).

	Effect of Score		Effect of trait-level x score			
			High		Low	
Variable	Scores (%) in subset	mean	Scores (%) in subset	mean	Scores (%) in subset	mean
A	10, 30, 45	0.94	no effect			
	10, 55, 70	0.58				
	55, 90	0.29				
ER	10, 30, 45, 70, 90	0.23	no effect			
	10, 45, 55, 70, 90	0.18				
P	90	0.98	no effect			
	55, 70	0.44				
	30, 45, 55	0.12				
	10, 30, 45	0.04				
R	10, 30, 45	0.74	no effect			
	55, 70, 90	0.19				
Slant	10, 30, 45, 70, 90	-0.10	no effect			
	30, 45, 55, 70, 90	-0.16				
NS	10, 30, 45, 70	1.88	no effect			
	10, 55, 70, 90	1.39				

Effects of Score

H1 is confirmed as there is a significant effect of Score on the type on the use of Advice, Praise, Reassurance and Emotional Reflection. For Advice, the pairwise comparison reveals three homogeneous subsets, and confirms H1c in that Advice should not be given for 90%, however it also indicates that Advice should not be given for 55% either. However, this is marginal (as 55% is also in the second subset, which recommends the use of Advice). The pairwise comparison for Emotional Reflection reveals two homogeneous subsets, however the use of Emotional Reflection is not recommended for either, confirming H1b. For Praise, a pairwise comparison reveals four homogeneous subsets. H1a is mostly confirmed,

with praise only recommended for 90%. For Reassurance, the pairwise comparison reveals two homogeneous subsets, which confirm H1d in that Reassurance should only be used for failing scores. We also found an effect of score on the slant employed. The pairwise comparison shows two homogeneous subsets, however these both result in neutral slant. There was an additional effect of Score on the total number of emotional support statements. There are two homogeneous subsets, which indicate that two statements should be used for 45% and 30%, and one statement should be used for 90% and 55%. 10% and 70% appear in both subsets, so could be offered one or two statements. Based on their means, we recommend using 2 for 10% and 1 for 70%.

Effects of Trait-Level

For Extraversion, there is an unexpected effect of the trait level overall on Praise, Emotional Reflection and Total number of statements. For Praise, the mean for high extraversion is 0.25 and the mean for low extraversion is 0.40, meaning learners receive slightly more praise if they are less extrovert. For Emotional Reflection, the mean for high is 0.14 and low is 0.26, showing a similar small effect. For number of statements, there is a similar effect again, the mean number of statements for high extraversion is 1.46 and the mean for low extraversion is 1.83, showing that less extroverted learners receive more emotional support overall. This confirms H2. There was no effect for Extraversion on Slant employed, so H3 is rejected, as expected.

7 EXPERIMENT 4: AGREEABLENESS

Agreeableness describes general affability of a person (Costa and McCrae, 1992).

7.1 Hypotheses

Although learners with high agreeableness may be more pleasant to work with, we do not expect that feedback given to learners should vary greatly. Futhermore, giving learners who are less agreeable negative slant on feedback, or less Emotional Support would not be ethical.

H1: The type of emotional support given will depend on the score the learner achieved

- **H1a:** Praise will only be used when the score is a good pass ($\geq$ 70%)
- **H1b:** Emotional Reflection will rarely be used
- **H1c:** Advice will not be given for the highest score (90%)
- **H1d:** Reassurance will be given for the failing scores (<50%)

H2: The quantity of emotional support utilized will differ between high and low Agreeableness[7]
H3: The slant utilized will differ between high and low Agreeableness[6]

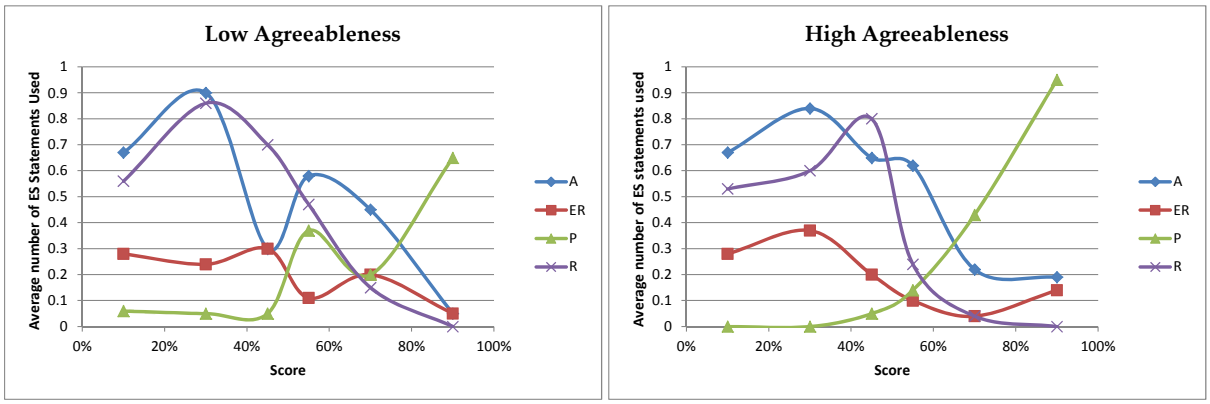

Fig.6. Average number of Emotional Support statements used, for high and low agreeableness, as score increases.

Table 15
Significance values of 2-way MANOVA for Agreeableness. Independent variables are the columns, and dependent variables are the rows. – indicates no significance. P = number of Praise statements, A = number of Advice statements, ER = number of Emotional Reflection statements, R = number of Reassurance Statements, NS = Total number of Emotional Support Statements.

	score	*trait level*	*trait level* × *score*
Slant	$F(5, 240) = 7.60, p < .01$	–	–
A	$F(5, 240) = 7.66, p < .01$	–	–
ER	–	–	–
P	$F(5, 240) = 24.85, p < .01$	–	$F(5, 240) = 2.74, p < .02$
R	$F(5, 240) = 13.01, p < .01$	–	–
NS	$F(5, 240) = 4.96, p < .01$	–	–

7.2 Results

Figure 6 shows the use of the different types of Emotional Support (A, ER, P, R) per trait level. Table 15 shows the significant effects of the 2-way MANOVA of trait-level × score on the use of *A*, *ER*, *P*, *R*, *Slant* and *NS*. Table 16 show the results of the pairwise comparisons.

Effects of Score

H1 is confirmed as there was a significant effect of score on the number of Advice, Praise and Reassurance statements given. H1b is confirmed as there is no significant effect for ER, and the average for each score (see Figure 6) is below 0.5. For Advice, a pairwise comparison indicates that there are three homogeneous subsets. These indicate that advice should not be used for 90%, confirming H1c. 45% and 70% appear in subsets 2 (which recommend Advice) and 3 (which does not), however the means are both below 0.5 so advice should probably not be given for these scores. For Praise, a pairwise comparison yields three subsets– this supports H1a in that Praise is only given for 90% as predicted, though not 70%.

[7] Although we do not expect there to be a difference, we left the hypothesis as is, as we can test this with the statistics.

Table 16
Agreeableness: homogeneous subsets arising from a post-hoc pairwise comparison of score, and the interaction effect of trait-level × score, on the average number of each type of emotional support statement given (A, ER, P, R), slant, and mean total amount of emotional support given (NS).

	Effect of Score		Effect of trait-level x score			
			High		Low	
Variable	Scores (%) in subset	mean	Scores (%) in subset	mean	Scores (%) in subset	mean
A	10, 30, 55	0.72				
	10, 45, 55, 70	0.52	no effect			
	45, 70, 90	0.31				
ER	no effect		no effect			
P	90	0.80	90	0.95	90	0.65
	55, 70	0.29	70	0.43	55, 70	0.28
	10, 30, 45, 55	0.09	10, 30, 45, 55	0.05	10, 30, 45, 70	0.09
R	10, 30, 45	0.66				
	10, 55	0.44	no effect			
	55, 70	0.22				
	70, 90	0.05				
Slant	10, 45, 90	-0.02				
	30, 45, 70, 90	-0.14	no effect			
	55, 70	-0.34				
NS	10, 30, 45, 55	1.56				
	10, 45, 55, 70, 90	1.24	no effect			

For Reassurance, the pairwise comparison reveals four homogeneous subsets. This supports H1d– that Reassurance should be given for failing scores. There was an unexpected effect of score on slant, the pairwise comparison reveals three homogeneous subsets, but each recommend a neutral slant. We also found an effect of score on NS. There are two homogenous subsets. For the failing scores and 55%, two statements are indicated. For the remaining scores, one statement is indicated, however 45%, 10% and 55% appear in both subsets. From their means, we recommend 2 for 45% and 10%, and one for 55%.

Effects of Trait-Level × Score

As expected, we found no evidence to support Hypotheses H2 and H3. However, we did find an interaction effect of trait-level × score on Praise. However, both pairwise comparisons for high and low Agreeableness recommends Praise to be used for 90% only, meaning there is no difference in strategy recommended between trait levels.

8 EXPERIMENT 5: OPENNESS TO EXPERIENCE

Openness to Experience describes qualities such as active imagination, aesthetic sensitivity, attentiveness to inner feelings, preference for variety, and intellectual curiosity (Costa and McCrae, 1992).

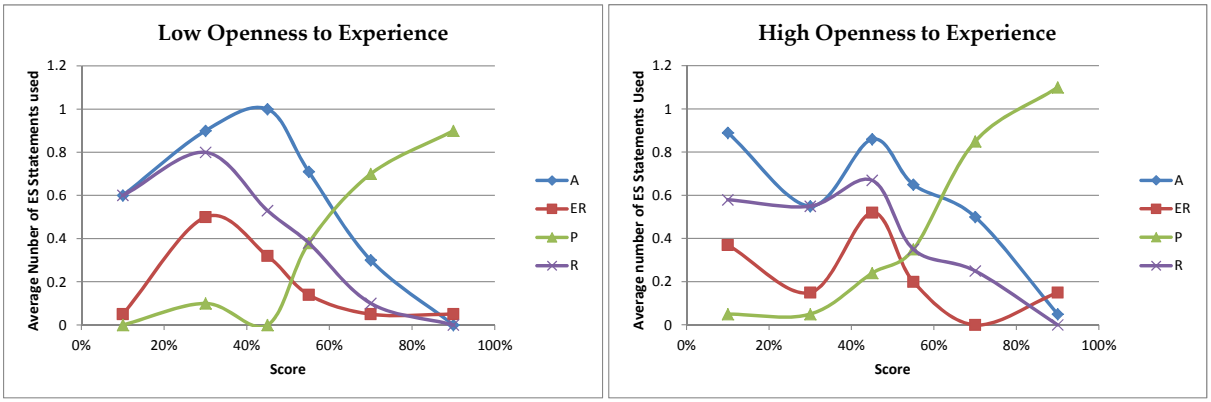

Fig.7. Average number of Emotional Support statements used, for high and low openness to experience, as score increases.

8.1 Hypotheses

Openness to Experience is a difficult trait to describe precisely, and as such, differences in feedback are hard to predict. We thus take the same approach to Extraversion and Agreeableness when forming our hypotheses.

H1: The type of emotional support given will depend on the score the learner achieved

- **H1a:** Praise will only be used when the score is a good pass ($\geq$ 70%)
- **H1b:** Emotional Reflection will rarely be used
- **H1c:** Advice will not be given for the highest score (90%)
- **H1d:** Reassurance will be given for the failing scores (<50%)

H2: The quantity of emotional support utilized will differ depending on the level of Openness to Experience.

H3: The slant utilized will differ depending on the level of Openness to Experience

8.2 Results

Figure 7 shows the use of the different types of Emotional Support (A, ER, P, R) per score and trait-level. Table 17 shows the significant effects of the 2-way MANOVA of trait-level $\times$ score on the use of *A*, *ER*, *P*, *R*, *Slant* and *NS*. Table 18 show the results of the pairwise comparisons.

Effects of Score

H1 is confirmed as there is a significant effect of score on the amount of Advice, Emotional Reflection, Praise and Reassurance given. For Advice, a pairwise comparison shows three homogeneous subsets, with Advice being recommended for all scores except 90%. This supports H1c. For Emotional Reflection, a pairwise comparison shows three subsets, however no subset mean is greater than 0.5, meaning

Table 17
Significance values of 2-way MANOVA for Openness to Experience. – indicates no significance. P = number of Praise statements, A = number of Advice statements, ER = number of Emotional Reflection statements, R = number of Reassurance Statements, NS = Total number of Emotional Support Statements.

	score	*trait-level*	*trait-level* × *score*
Slant	$F(5, 228) = 6.66, p < .03$	–	–
A	$F(5, 228) = 9.80, p < .01$	–	–
ER	$F(5, 228) = 4.12, p < .01$	–	$F(5, 228) = 2.63, p < .03$
P	$F(5, 228) = 26.14, p < .01$	–	–
R	$F(5, 228) = 8.23, p < .01$	–	–
NS	–	–	–

Table 18
Openness to Experience: homogeneous subsets arising from a post-hoc pairwise comparison of score, and the interaction effect of trait-level × score, on the average number of each type of emotional support statement given (A, ER, P, R), slant, and mean total amount of emotional support given (NS).

	Effect of Score		Effect of trait-level x score			
			High		Low	
Variable	Scores (%) in subset	mean	Scores (%) in subset	mean	Scores (%) in subset	mean
A	10, 30, 45, 55	0.77				
	10, 30, 55, 70	0.64		no effect		
	90	0.02				
ER	10, 30, 45, 55	0.28	10, 45	0.45	30, 45	0.41
	10, 30, 55, 90	0.20	30, 55, 70, 90	0.13	10, 45, 55, 70, 90	0.12
	10, 55, 70, 90	0.13				
P	70, 90	0.89				
	30, 45, 55	0.19		no effect		
	10, 30, 45	0.08				
R	10, 30, 45, 55	0.56				
	55, 70	0.28		no effect		
	70, 90	0.09				
Slant	10, 45, 90	0.01				
	45, 70, 90	-0.08		no effect		
	30, 55, 70	-0.28				
NS	no effect			no effect		

this strategy is not recommended for any score, which supports H1b. For Praise, a pairwise comparison shows three homogeneous subsets. Praise is recommended for 90% and 70%, supporting H1a. For Reassurance, hypothesis H1d is partially confirmed by the pairwise comparison, as reassurance is recommended for failing grades, but also 55%. However, this is marginal, as 55% also appears in subset 2, and has a mean less than 0.5, so reassurance should perhaps not be given for this score. We also found an effect of score on slant. A pairwise comparison reveals that although there are three homogeneous subsets, the slant should remain neutral for all scores.

Table 19
Summary of significant effects of the Emotional Support experiments. Bold indicates where the significant effect made a difference to the recommendations. A = amount of advice given, ER = amount of emotional reflection given, P = amount of praise given, R = amount of reassurance given, NS = total amount of emotional support given overall.

Trait	score	trait level	trait level × score
Extraversion	Slant, **A**, ER, **P**, **R**, **NS**	ER, P, **NS**	–
Agreeableness	Slant, **A**, **P**, **R**, **NS**	–	P
Conscientiousness	Slant, **A**, ER, **P**, **R**	Slant	**A**, **Slant**
Emotional Stability	Slant, **A**, ER, **P**, **R**, NS	Slant, ER	**A**, **ER**, **NS**
Openness to Experience	Slant, **A**, ER, **P**, **R**	–	ER

Effects of Trait-Level × Score

There was no evidence for H2 and H3, as we found no significant effect for the interaction between trait-level × score for slant or NS. However, we found an unexpected interaction between trait-level × score for Emotional Reflection. The pairwise comparison between high and low Openness to Experience on ER reveals differences in the homogeneous subsets, however there is no overall effect on the recommendation that Emotional Reflection should not be used for either level of Openness to Experience.

9 DISCUSSION OF RESULTS

Above, we have presented a series of experiments which examined how the provision of feedback (emotional support and slant) differs between trait levels for all five traits of the five-factor model. We found a number of significant effects (summarized in Table 19) and conducted a post-hoc pairwise analysis on each to make recommendations for the type of emotional support to give, how much to give and the slant to use, based on the significant effects. Table 20 summarizes the recommendations from the analysis.

To decide on the types of emotional support strategy to recommend for a particular trait level and score, we first considered the cases where there was a significant interaction of trait-level × score for each of the types of emotional support (see summary in Table 19), and used the recommendations from the post-hoc analysis of trait-level × score, by rounding the average of each of the homogeneous subsets to decide whether to recommend a particular emotional support strategy (if ≥ 0.5). These are the blue items in Table 20. For example, for conscientiousness, there was a significant effect of trait-level × score for Advice. Based on the homogeneous subsets (see Table 10), Advice should be used for 10%-55% for high conscientiousness, and 10%-70% for low conscientiousness, resulting in the blue 'A's in the Conscientiousness rows in Table 20.

Thereafter, if the score has a significant effect for the strategy type, we use the recommendations from the post-hoc analysis on score in a similar way to decide which strategies to give (excluding those strategies already used in the previous step). These are the red items in Table 20. For example, for extraversion, there was a significant effect of score on Praise. Based on the homogeneous subset table 14, Praise should only be given for a score of 90%. This resulted in the red P in the extraversion row.

Sometimes, a score could appear in a homogeneous subset for a particular strategy with an overall mean greater than 0.5, but the individual mean for that score in particular was less than 0.5, meaning

Table 20

Summary of findings for each trait and score. Strategy (the emotional support strategy), Slant and NS (number of statements) were recommended by mean (from pairwise analysis). – indicates no recommendation. Blue = recommendation from trait level × score. Red = recommendation from Score only. * indicates that the recommendation is made where the mean was lower than 0.5

Trait	Level		Score (%) 10	30	45	55	70	90
Extraversion	High & Low	Strategy	A, R	A, R	A, R	A*	A	P
		NS	2	2	2	1	1	1
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
Agreeableness	High & Low	Strategy	A, R	A, R	A*, R	A	A*	P
		NS	2	2	2	1	1	1
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
Conscientiousness	High	Strategy	A, R	A, R	A, R*	A	–	P
		NS	2	2	2	2	2	2
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Low	Strategy	A, R	A, R	A, R*	A	A	P
		NS	2	2	2	2	2	2
		Slant	Neutral	**Negative**	Neutral	Neutral	**Negative**	Neutral
Emotional Stability	High	Strategy	A, R	A, R	A, R	A, R*	P	P
		NS	2	2	2	2	2	2
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
	Low	Strategy	A, ER, R	A, ER, R	A, ER*, R	A, R*	A*, P	P
		NS	3	3	2	2	2	2
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral
Openness to Experience	High & Low	Strategy	A, R	A, R	A, R	A, R*	A*, P	P
		NS	2	2	2	2	2	2
		Slant	Neutral	Neutral	Neutral	Neutral	Neutral	Neutral

the recommendation is weak. These are the items marked with a * in Table 19. If a score appeared in two subsets, one for which a particular strategy was recommended and the other not, the strategy was recommended regardless.

There was one exceptional case. For high conscientiousness at 70%, we cannot make a recommendation: the subset for Advice and high conscientiousness does not recommend its use at this score, however disregarding the trait level, the subsets of score on Advice recommend it. Examining the subsets for score only, no other types of emotional support are recommended for this score.

To decide on the number of statements to give, we followed a similar procedure as for the type of emotional support strategies, by first considering the significant interaction of trait level × score to decide how many statements to give, then the significant effects of score based on the pairwise analysis. A score can appear in more than one subset. Where this occurred, we rounded the average for the individual score to decide. If there were no significant effects for the number of statements to give, we used the average across all scores as our recommendation.

To decide on the slant to use, we used the same process. The only trait where slant had a subset mean

greater than 0.5 or less than -0.5 was Conscientiousness, with a negative slant being used for learners with low conscientiousness at 30% and 70%.

There were a few cases where there was a significant effect for trait-level, but not the interaction of trait level $\times$ score. For Extraversion, these were for the amount of Emotional Reflection to give, the amount of Praise to give, and the number of statements overall. For Emotional Reflection and Praise, the mean for each trait level was below 0.5, so these can be discounted. For the total number of statements, we know that there is a difference, but as there is no interaction effect, we do not know at which scores these differences can be used (given there is also a significant effect on score). For Emotional Stability, there was a difference for slant, but the difference was too small to matter, and we found little evidence to recommend anything other than neutral slant. Thus these differences were discounted when we constructed our recommendations.

Finally, where the recommendations from the analysis were the same for both levels of a trait, these were collapsed into one row (Extraversion, Agreeableness and Openness to Experience).

10 CREATION OF AN ALGORITHM TO ADAPT EMOTIONAL SUPPORT AND SLANT TO PERSONALITY & PERFORMANCE

We now have a large dataset of the use of Emotional Support for different learner personalities and scores. To enable a Conversational Agent to use these adaptations, we need to create an algorithm which describes the adaptations present in the data. Above, we performed a statistical analysis on the dataset, and created a set of recommendations for the type(s) of Emotional Support and slant to use for each personality trait and score. However, this is not yet an algorithm as there are cases where we have a mismatch between the number of statements to give and the number of strategies recommended, and for high conscientiousness at 70%, we have no strategy recommendation at all. We also have not considered the order of the Emotional Support types. For example, we know that Advice and Reassurance are used together often, and it may be that reassurance is used after advice, or vice versa.

This section describes the creation and evaluation of an algorithm for recommending the types of Emotional Support and slant in feedback for each personality type and score, which also considers the order of strategies. We first describe how we can evaluate how well an algorithm describes the data it was based on. We then present the evaluation of two algorithms, generated in different ways. We finish by presenting a final algorithm which takes elements from both, and investigate the best order to give different types of emotional support in.

10.1 Evaluation (Validation) Measure for Algorithms

To calculate how well the algorithm describes the emotional support data gathered in Section 3, we used the DICE scoring measure (van Deemter et al., 2012). For each response by a human participant, a DICE score is calculated as follows:

$$DICE = \frac{2 \times ES_u}{(NS_r + NS_a)}$$

Where ES_u is the number of Emotional support strategies that the participant's response and algorithm have in common, NS_r is the total number of Emotional Support statements in the participant's response, and NS_a is the total number of Emotional Support statements that the algorithm recommends. For example, if the algorithm recommends emotional support feedback of length 2, containing one Advice statement (A) and one Reassurance statement (R), and the participant's response contains one emotional support statement, namely A, then: $ES_u = 1$, $NS_r = 1$, and $NS_a = 2$.

The DICE score for this response would be as follows:

$$DICE = \frac{2 \times 1}{(1 + 2)} = 0.666\ldots$$

For each trait level and score, the DICE scores for each individual response was calculated and then averaged. These score averages were then averaged again to give the DICE average for the trait (and level, where appropriate). The closer the score is to 1, the better the algorithm is at describing the data it is being compared to.

Baseline DICE score

Once the DICE score for the adaptive algorithm was calculated, we required a DICE score for a non-adaptive algorithm to compare it to, to check how well our algorithm performed compared to a baseline. To establish a baseline, we created an algorithm with a random recommendation of Emotional Support for each trait level and score combination: we provided a 50% chance of each type of emotional support being selected. We ran this algorithm ten times, calculating the DICE score each time by comparing the random emotional support with the experimental data, and averaged the resulting DICE scores. The resulting DICE score for this non-adaptive baseline is 0.33. An alternative baseline algorithm of always providing all emotional support categories provides similar results with a DICE score of 0.38.

10.2 Algorithm based on Statistical Analysis for Emotional Support

The statistical analysis resulted in a set of recommendations (see Table 20) which indicated how many strategies to use, and the types of Emotional Support to use for each personality trait and score. We transformed this into a statistical algorithm using the following procedure:

- Where the number of statements recommended was higher than the number of strategies recommended, we used multiple statements of the same type. For example, for high conscientiousness, there are two statements recommended for 90%, and one strategy, Praise, so we treated this as two Praise statements.
- Where the number of statements recommended was lower than the number of strategies recommended, we used the strategies with the highest means. There was only one instance of this, low Emotional Stability at 45%, which recommended 2 statements picked from Advice, Emotional Reflection or Reassurance (see Table 20). We used Advice and Reassurance, as these had the highest means.

- Where the statistical algorithm could make no recommendation for the type of emotional support to give, we treated this as no emotional support. This only occurred once, for high Conscientiousness and a score of 70%.

We were interested to see how well this algorithm performed compared to the actual data from the responses. For Agreeableness, Extraversion and Openness to Experience, the statistical analysis showed that the significant differences found were not strong enough to recommend different strategies for their trait levels. We thus DICE scored the responses independent of trait level in these cases. For Conscientiousness and Emotional Stability, there were clear differences in the recommendations from the statistical analysis and so we treated them independently. Table 21 shows the resulting algorithm and the calculated the DICE score for each of the trait levels.

The DICE score is modest for the statistical algorithm, with a mean DICE score of 0.52, though this still substantially outperforms the two baseline scores (see Section 10.1) of 0.33 and 0.38. The DICE score is sensitive to participants' responses which contained no emotional support at all (these have a DICE score of 0), and we hypothesised that these were bringing the DICE score averages down. As emotional support was always recommended, we calculated an *adjusted* DICE score, which only

Table 21
Algorithm based on statistical analysis of all trait experiments.

Trait	Level		Score (%) 10	30	45	55	70	90	Avg DICE (Avg Adj)
Extraversion	High and Low	Strategy	A R	A R	A R	A	A	P	
		DICE	0.54	0.64	0.61	0.34	0.34	0.66	0.52
		DICE Adjusted	*0.66*	*0.73*	*0.68*	*0.47*	*0.42*	*0.75*	*0.62*
Agreeableness	High and Low	Strategy	A R	A R	A R	A	A	P	
		DICE	0.53	0.59	0.54	0.41	0.24	0.67	0.50
		DICE Adjusted	*0.66*	*0.69*	*0.64*	*0.55*	*0.38*	*0.86*	*0.63*
Conscientiousness	High	Strategy	A R	A R	A R	A A	X	P P	
		DICE	0.61	0.54	0.50	0.35	0.25	0.54	0.46
		DICE Adjusted	*0.65*	*0.60*	*0.66*	*0.48*	*0.00*	*0.56*	*0.49*
	Low	Strategy	A R	A R	A R	A A	A A	P P	
		DICE	0.42	0.55	0.64	0.35	0.39	0.40	0.46
		DICE Adjusted	*0.64*	*0.73*	*0.68*	*0.49*	*0.41*	*0.53*	*0.58*
Emotional Stability	High	Strategy	A R	A R	A R	A R	P P	P P	
		DICE	0.57	0.69	0.64	0.52	0.28	0.62	0.55
		DICE Adjusted	*0.68*	*0.72*	*0.64*	*0.52*	*0.35*	*0.65*	*0.59*
	Low	Strategy	A ER R	A ER R	A R	A R	A P	P P	
		DICE	0.71	0.64	0.57	0.53	0.58	0.55	0.60
		DICE Adjusted	*0.75*	*0.71*	*0.67*	*0.66*	*0.58*	*0.69*	*0.68*
Openness to Experience	High and Low	Strategy	A R	A R	A R	A R	A P	P P	
		DICE	0.50	0.55	0.52	0.42	0.53	0.58	0.52
		DICE Adjusted	*0.69*	*0.68*	*0.63*	*0.54*	*0.64*	*0.68*	*0.65*
								Overall DICE Average for Algorithm:	0.52
								Adjusted:	*0.61*

compared the algorithm with the responses that contained emotional support. These are the adjusted scores shown in italics in Table 21. This saw a small improvement of the DICE scores as the mean adjused DICE score of the algorithm as a whole is 0.61.

The DICE score is also affected by the case where we do not know which emotional support category to recommend, which was treated in the scoring as a recommendation to give no support at all. We could have substituted this for a recommendation of a random strategy, however this would be arbitrary. A further problem with the statistical method is the total number of emotional support statements to give. Where there were no statistically significant effects for number of statements (e.g. conscientiousness), the recommendations (see Table 20) used a rounded average. However, this mean tended to be around 1.5, which is rounded to 2, causing a poor fit when tested against the original data as we may be recommending to give more emotional support than the majority of the participants originally gave. So we investigated an alternative approach to generate the algorithm, described in the next section.

10.3 Algorithm based on Data analysis

To generate the alternative algorithm for each trait, we took the median number of statements given (NS) for each trait level and score, and then picked the most commonly used strategies for a response with this NS. Where the median was a decimal, we used the mode NS as a guide to decide whether to round up or down.

When deciding on the most common feedback type, we identified three alternative approaches. For example, if the median NS was 3, we could:

- Find the most common single strategy, then the most common pair of strategies containing that single, then find the most common triple containing that pair, or
- Find the most common triple overall, or
- Find the three most common strategies overall

We computed the results for all three approaches, and the results were the same for all traits and scores.

In this part of the analysis, we do not consider the order that the strategies were used in. So, in other words, P & A would count the same as A & P.

We could have used the mode NS to generate the algorithm, however this raises problems reconciling the number of statements to give to use when there are multiple modes (for example, even numbers of participants gave 1, 2 or 3 statements). The issue is further complicated by the mode being 0 when the number of statements given varies considerably (2,3,4,5,6 statements) leaving 0 as the most common, despite more participants wanting to give emotional support than those wishing to omit it. This is not representative of what most participants wanted, so we decided that the median was the best compromise.

When we evaluated the statistical algorithm, we used an adjusted score. However, in this approach, it raises some difficulties. The participants who gave no Emotional Support were factored into the generation of the algorithm as their total number of statements was zero, affecting the calculation of the median (which was used to decide how many statements to give). It seems unfair to take them into account when generating the algorithm, but not when evaluating it. So we decided to ignore the results from partici-

Table 22
Alternative algorithm from data analysis of all trait experiments, only considering those instances where emotional support was given. Bold = recommendation performs better than the statistical algorithm. Italic = recommendation performs worse than the statistical algorithm.

Trait	Level		Score (%) 10	30	45	55	70	90	Avg adj DICE
Extraversion	High and Low	Strategy	A R	A R	A R	A	**A P**	P	0.65
		DICE	0.66	0.73	0.68	0.47	**0.59**	0.75	
Agreeableness	High and Low	Strategy	A R	A R	*R*	A	A	P	0.62
		DICE	0.66	0.69	*0.54*	0.55	0.38	0.86	
Conscientiousness	High	Strategy	*A ER*	*A*	*A ER*	**A P**	**P**	**P**	0.63
		DICE	*0.59*	*0.57*	*0.60*	**0.64**	**0.58**	**0.80**	
	Low	Strategy	A R	A R	A R	**A R**	**A**	**P P**	0.60
		DICE	0.64	0.73	0.68	**0.54**	**0.51**	**0.53**	
Emotional Stability	High	Strategy	A R	*R*	A R	A R	**P**	**P**	0.59
		DICE	0.68	*0.42*	0.64	0.52	**0.46**	**0.83**	
	Low	Strategy	A ER R	A ER R	A R	*A*	*P*	**P**	0.67
		DICE	0.75	0.71	0.67	*0.59*	*0.51*	**0.81**	
Openness to Experience	High and Low	Strategy	A R	A R	A R	*A*	**P**	**P**	0.68
		DICE	0.69	0.68	0.63	*0.53*	**0.65**	**0.87**	
						Overall DICE Average for Algorithm:			0.62

pants who gave no Emotional Support when calculating the median, as the analysis (described in Section 9) always recommended giving some Emotional Support for every trait level and score.

Table 22 shows the alternative algorithm and DICE scores. The overall mean DICE score for the algorithm is 0.62, which is slightly better than the score generated by the statistical algorithm. However, although some of the differences in this algorithm perform better than those recommended in the statistical one (the bold items), some perform worse (the italic items) when comparing the DICE adjusted scores. For example, for high conscientiousness, A and ER are recommended, which is contrary to what our statistical analysis recommends and has a lower DICE than A R, which the statistics recommended.

To decide on the Slant to employ, we compared the statistical recommendation with the most common slant employed for every response, and there was no difference between the two. We thus only recommend to use a negative slant for Low Conscientiousness at 30% and 70% in our final algorithm.

10.4 Overall algorithm

We now have two algorithms: one based on statistical analysis (Table 21) and the other based on data analysis (Table 22). Both have their strengths and weaknesses for certain trait levels and scores, with one performing better than the other. We thus decided to combine the algorithms. Where recommendations differed, we picked the best scoring one and used that in our final algorithm.

We now investigate the order to provide combinations of strategies in, when they are used together: Advice and Reassurance; Advice and Praise; and Advice, Reassurance and Emotional Reflection.

Table 23
Final algorithm, with the order of Emotional Support Strategies included.

Trait	Level		Score (%) 10	30	45	55	70	90	Avg adj DICE
Extraversion	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	A neutral	P A neutral	P neutral	0.65
Agreeableness	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	A neutral	A neutral	P neutral	0.64
Conscientiousness	High	Strategy Slant	R A neutral	R A neutral	R A neutral	P A neutral	P neutral	P neutral	0.66
	Low	Strategy Slant	R A neutral	R A **negative**	R A neutral	R A neutral	A **negative**	P P neutral	0.60
Emotional Stability	High	Strategy Slant	R A neutral	R A neutral	R A neutral	R A neutral	P neutral	P neutral	0.64
	Low	Strategy Slant	ER R A neutral	ER R A neutral	R A neutral	R A neutral	P A neutral	P neutral	0.69
Openness to Experience	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	R A neutral	P neutral	P neutral	0.68
							Overall DICE Average for Algorithm:		0.65

Advice and Reassurance were only given together for scores up to 55%. We thus considered the data from all five trait experiments for 10% – 55%. Reassurance followed by Advice is the most popular for scores of 30% (58% for R then A) and 55% (66% R then A). For a score of 10%, A then R was more popular, however this was very close (52% for A then R and 48% for R then A), so we decided to always give Reassurance followed by Advice.

Praise and Advice were only given together for a score of 70%. Following a similar procedure, we found that 81% of participants use Praise followed by Advice, so this is what we recommend.

The combination of Advice, Emotional Reflection and Reassurance is only used for scores of 10% and 30%, so following the same procedure as before, we found that ER, R, A is the most popular order overall, so we decided on this for our recommendation.

Our final algorithm including order is shown in Table 23. We now have an algorithm which describes the adaptations that we observed with reasonable accuracy. Next, we describe a qualitative study to investigate whether the recommendations from the algorithm are appropriate for learners. We also investigate what to recommend for a learner who has normal conscientiousness and emotional stability, by using the recommendations for the three traits which do not distinguish between trait levels.

11 EVALUATION & REFINEMENT OF ALGORITHM

This section uses a study to evaluate and refine the algorithm developed in Section 10, which recommends different types of Emotional Support and slant in feedback tailored to learner personality and progress. While we know how well our algorithm describes the experimental data gathered, we do not know whether the feedback it recommends is appropriate (i.e. is it good for the learner's well-being, and does

it reflect what a teacher would do in a real-world setting?).

There are multiple ways in which the algorithm's appropriateness could be tested. One approach would be to observe the effect of the feedback on learners following a short learning exercise. However, at this stage, the algorithm can only adapt to one polarized trait at once, and finding learners with the exact personality type required for evaluation would be difficult. Additionally, as the feedback recommended by the algorithm is untested, there could be ethical implications of this on the learner's well-being. A second approach is to evaluate the feedback choices recommended by the algorithm with real life teachers. This could be done via a further quantitative study. However, finding a large enough sample of teachers for robust statistical analysis would be difficult. Furthermore, this approach would not lead us to fully understand why they came to their decisions.

So we decided that a qualitative study would be more suitable. Focus groups were chosen as they provide a good way of discovering opinions, and why people hold them. Focus groups also allow participants to reach a consensus where opinions originally differ, and the ensuing discussion provides insights.

A second objective is to resolve conflicts in the algorithm for each score where the level of a trait did not seem to matter. For example, for a score of 55%, there was no difference in the recommendation of Emotional Support between high and low levels for Extraversion (recommended: A), Agreeableness (recommended: A), Openness to Experience (recommended: R A) and Emotional Stability (recommended: R A) (see Table 24). If the trait level truly does not matter, we would expect the same recommendation for all of these traits (like we see for a score of 45%, where R A is recommended in all cases). Thus, we investigated this in the focus groups as a secondary objective.

11.1 General rules considering all traits and deviations

First, we considered those trait and score combinations where the trait level did not matter:

- For 10%, 30% and 45%, R and A is given for all of these scores.
- For 90%, P is given.
- For 55%, there is an unexpected difference between traits: for Extraversion and Agreeableness, A is given for high and low levels, whilst for Openness to Experience and Emotional Stability, R and A are given for high and low levels (see the red items in Table 24).
- For 70%, there is an unexpected difference between traits. For Extraversion, A and P are given. For Agreeableness, A is given. For Openness to Experience, P is given (see the red items in Table 24).

Next, we considered those trait and score combinations where trait level did matter (the blue items in Table 24):

- For 10% and 30%, for low Emotional Stability, ER is added to the standard R and A.
- For 55%, for high Conscientiousness, P and A are given, while R A is recommended for low Conscientiousness.
- For 70%, for high Conscientiousness and high Emotional Stability, P is given. For low Conscientiousness, A is given. For low Emotional Stability, A and P are given.
- For 90%, for low Conscientiousness, an additional P statement is given.

Table 24
Algorithm for providing emotional support resulting from the work in this paper. Red=differences in traits that did not matter between levels, that we were interested in resolving. Blue = differences between trait levels that we were interested in discussing.

Trait	Level		Score (%) 10	30	45	55	70	90
Extraversion	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	**A** neutral	**P A** neutral	P neutral
Agreeableness	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	**A** neutral	**A** neutral	P neutral
Openness to Experience	High and Low	Strategy Slant	R A neutral	R A neutral	R A neutral	**R A** neutral	**P** neutral	P neutral
Conscientiousness	High	Strategy Slant	R A neutral	R A neutral	R A neutral	**P A** neutral	**P** neutral	**P** neutral
	Low	Strategy Slant	R A neutral	R A **negative**	R A neutral	**R A** neutral	**A** **negative**	**P P** neutral
Emotional Stability	High	Strategy Slant	**R A** neutral	**R A** neutral	R A neutral	**R A** neutral	**P** neutral	P neutral
	Low	Strategy Slant	**ER R A** neutral	**ER R A** neutral	R A neutral	**R A** neutral	**P A** neutral	P neutral

We investigate whether teachers believe that the algorithm's adaptations to learner personality and progress (highlighted in blue in Table 24) are appropriate. We also investigate which of the differences highlighted in red in Table 24 teachers think are important to keep, and which are deemed inappropriate (and may just have been caused by slight differences in the median when the algorithm was generated). We hope that some of the conflicts may be resolved, leaving us with a clearer picture, and facilitating the creation of an algorithm which can adapt to more than one polarized trait at once.

11.2 Focus Groups

Four focus groups (FG1, FG2, FG3 and FG4) were held, in which participants provided their opinions in a group setting. Fifteen participants took part in the focus groups, selected through convenience sampling. There were eight women and seven men, four participants were 18-25 years old and eleven participants aged 26-40. The eight participants of FG1 and FG2 were training to be primary school teachers, the seven participants of FG3 and FG4 had experience as teaching assistants or lecturers in Computing Science in higher education. As materials we used: the 10 stories expressing polarized personality traits (see Table 5), the emotional support categories and definitions (Dennis et al., 2013), the emotional support statements for each category (see Table 6), and the algorithm predictions (see Table 24).

We defined research questions (RQ1 – RQ6): these are provided in the results section below in bold. We investigated RQ6 in FG1 and 2. In FG3, we investigated all research questions except RQ3(d) and

RQ3(e) (as time did not permit it), and in FG4 we investigated all research questions.

In FG1 and FG2, participants were shown the stories for low and high Extraversion, and low and high Agreeableness, and examples of praise and advice. Participants then gave their opinion on what emotional support to give for a score of 70% only. In FG3 and FG4, participants were introduced to the scores, personality traits and types of emotional support. They were shown the stories about the learners and examples of emotional support. They were informed that we were assuming that the learners were equal in ability. A group discussion was held, in which participants gave their opinion on each of the research questions.

11.3 Results

In this section, we will summarize the discussion for each of the research questions, and briefly indicate our conclusion for each.

Appropriateness

RQ1. Is the general adaptation pattern of the usage of different emotional support types to learner scores appropriate? Participants of FG3 and FG4 agreed that it was appropriate. They said that they "would expect that" and "that makes sense" to them. *We conclude that the general adaptation to score is appropriate.*

RQ2. Is the adaptation of different Emotional Support types to trait level for Emotional Stability and Conscientiousness, but not for the other traits, appropriate? All participants in FG3 and FG4 agreed that this is sensible. For conscientiousness, they said that "[it] makes sense [to adapt to conscientiousness] as it is how they approach the work in the first place". For emotional stability, they thought that someone with low emotional stability would require "much more reassurance" whereas someone who is more emotionally stable "would require more advice". One participant in FG3 said that extraversion would matter if there was a group of learners. However, another participant stated that whilst in group work Extraversion would matter, "I think it makes sense in this context if it is just one teacher to one person because the degree of how many friends they have would not make a difference". *We conclude that it makes sense to adapt to levels of Emotional Stability and Conscientiousness, but not Extraversion, Openness to Experience and Agreeableness.*

RQ3. Are the differences in feedback between high and low Conscientiousness appropriate?
a) for Emotional Support with a score of 55%: All participants in FG3 agreed that praising the high person with 55% made sense. "The advice definitely for both low and high, but the praise for high makes sense [...] as someone who is highly conscientious will need it." FG4 also agreed with giving praise for a learner with high conscientiousness for this score, as "you know that they would have tried". *We conclude that this adaptation is appropriate.*

b) for Emotional Support with a score of 70%: All participants in FG3 agreed that this was appropriate and that "it makes sense", as if a learner had high conscientiousness, they would have "worked harder and deserve praise". On the other hand, if they were low in conscientiousness, they are "used to coasting

through" and the advice may make them work harder, and you don't want to "praise bad behaviour". In FG4, one participant said they would still praise, however other participants said that "they liked the idea of a learner with low conscientiousness getting some advice" as they could probably do better. However, in FG4, two participants leaned towards both Advice and Praise for this score. *We conclude that this adaptation should be kept as Advice, as the majority of participants felt it was appropriate.*

c) for Emotional Support with a score of 90%: Participants in FG 3 and 4 thought that this may be due to surprise that the learner who was not conscientious managed such a good score, or that this particular learner who is not normally conscientious has put extra effort in this time. "If someone who doesn't normally work hard comes out with a score that is this high, then I'd give them a lot of praise as they might think 'next time I'll work hard again'." When asked if this was ethical, participants said "Yes as this may motivate them to be more motivated in the future". *We conclude that this adaptation is appropriate.*

d) for negative Slant with a score of 30%, and low conscientiousness: FG4: All participants agreed this was ok -"I think that using [negative slant] for someone that has not worked hard will make them see that we expected more from them". *We conclude that this adaptation is appropriate.*

e) for negative Slant with a score of 70%, and low conscientiousness: FG4: Most participants thought this was ok, and they understood why it had happened. However, one participant did not as they had concerns about being "negative for a good mark". Others disagreed as "70% is closer to the pass rather than the top mark, so negative slant is ok". *We conclude that this adaptation is appropriate.*

RQ4. Are the differences between high and low Emotional Stability appropriate?
a) For Emotional Support with a score of 10% and 30%: In FG3 and FG4, participants agreed with the algorithm as "These are the people that I would reassure". However, one participant in FG3 worried about giving ER without knowing they are really upset. One participant said they would like to add more perspective (reassurance). One participant in FG4 also added that "Advice can also help to mitigate worry in learners". *We conclude that this adaptation is appropriate.*

b) For Emotional Support with a score of 70%: For low emotional stability, participants in FG3 and FG4 disagreed with Advice being given as well as Praise for 70%, as "they may take this very negatively" and that "they have done well". They would prefer to have just Praise for this score. *We conclude that this adaptation is not appropriate, and learners with low Emotional Stability should receive Praise only.*

Conflicts

RQ5. For a score of 55%, independent of trait level for Extraversion, Agreeableness, Openness to Experience and Emotional Stability, is it appropriate that A is sometimes used, and sometimes R and A, and which of these would be better for learners who are "normal" on all traits? In both FG3 and FG4, the majority of participants felt that there should be no difference between the traits. However one participant in FG3 noted that a person who is very introverted or very extroverted may require different feedback than a normal person and that this may cause differences between the traits. In FG3, participants leaned towards giving Advice and Reassurance for this score as "you want to reassure [the

learner] as they have passed, but you also want to give them advice about how to improve in the future". In FG4, participants were more divided. One participant "wouldn't give reassurance at all because for some people 55% might be the best that they can do". However, they would definitely give "Advice [. . .] if they were looking to improve their grade for the next time". *We conclude that there should not be a difference between Agreeableness, Extraversion and Openness to Experience for a score of 55% and the majority of participants decided to recommend R A for a "normal" person.*

RQ6. For a score of 70%, independent of trait level for Extraversion, Openness to Experience and Agreeableness, is it appropriate that A is sometimes used, P or P and A, and which of these would be better for learners who are "normal" on all traits? In FG1 and FG2, participants were shown the four stories representing Extraversion and Agreeableness at high and low levels. For each of these traits, participants were completely divided whether to give Praise or Advice for a score of 70%. So, they did not really make a differentiation between the traits as such. It was more a difference of opinion on whether a score of 70% should be praised or not, as it can be improved on, but is also a good score. One participant was concerned with giving praise – "I think saying 'I'm proud of you' to someone I didn't know would be weird". However, they would be fine with saying "well done", so they agreed that this was due to the text of the statement rather than the category of emotional support. In FG2, one participant stated "If I could pick both [Praise and Advice] I would, but if I had to pick one, I'd pick praise". In FG4, participants said that "it made no sense" that different strategies were offered for Openness, Extraversion and Agreeableness. In FG3 this issue was not discussed, with discussion centring on what to do for a "normal" learner. Most participants in FG3 and FG4 leant towards praising this mark for a 'normal' learner - "you'd probably just say 'well done' ". "Unless they weren't particularly happy with it, I'd say 70% is a good score and I would give praise". *We conclude that there should not be a difference between Agreeableness, Extraversion and Openness to Experience for a score of 70% and the majority of participants decided to recommend P for this score for a "normal" person.*

11.4 Refined Algorithm

Based on the results of the focus groups, we refined the algorithm (shown in Table 24) to take into account the perspectives of the participants. The final algorithm is shown in algorithm 1. The only case where we had a conflict in the algorithm was for slant at scores of 30% and 70%, when a learner has both low Conscientiousness (where a negative slant is recommended) and high or low Emotional Stability (where a neutral slant is recommended). In our final algorithm, we decided to not use the negative slants for learners with low Emotional Stability and low conscientiousness, as negative slanting is likely to have an adverse effect on neurotic learners. If a learner has high Emotional Stability but low Conscientiousness, we used the negative slants on scores of 30% and 70% as this learner is likely to be stable enough for it to be appropriate. However, this requires further investigations in the future.

Algorithm 1: Refined Algorithm for Emotional Support

Input : $score$ the score the learner achieved; $emotional\ stability$ the learner's level of emotional stability; $conscientiousness$ the learner's level of conscientiousness

Output: $slant$ the slant to use in feedback; ES the categories of emotional support to use in feedback

```
1  begin
2      switch score do
3          case 10%
4              slant = neutral;
5              if emotional stability = low then
6                  ES = ER R A;
7              else
8                  ES = R A;
9              end
10         end
11         case 30%
12             if emotional stability = low then
13                 ES = ER R A;
14             else
15                 ES = R A;
16             end
17             if conscientiousness = low ∧ emotional stability ≠ low then
18                 slant = negative;
19             else
20                 slant = neutral;
21             end
22         end
23         case 45%
24             slant = neutral;
25             ES = R A;
26         end
27         case 55%
28             slant = neutral;
29             if conscientiousness = high then
30                 ES = P A;
31             else
32                 ES = R A;
33             end
34         end
35         case 70%
36             if conscientiousness = low then
37                 ES = A;
38             else
39                 ES = P;
40             end
41             if conscientiousness = low ∧ emotional stability ≠ low then
42                 slant = negative;
43             else
44                 slant = neutral;
45             end
46         end
47         case 90%
48             slant = neutral;
49             if conscientiousness = low then
50                 ES = P P;
51             else
52                 ES = P;
53             end
54         end
55     endsw
56 end
```

12 CONCLUSIONS AND FUTURE WORK

We have found that learner personality is considered by humans when adapting slant and emotional support in feedback to learners, particularly conscientiousness and emotional stability. We have generated and evaluated an algorithm which encapsulates these adaptations. These results are encouraging but there are several limitations. The feedback was provided without much history or context for the given learner. So far, we only informed participants that the teacher expected the learner to pass. However, prior learner performance is likely to change the expectations of the teacher. For example, it may be that the learner in question really struggles with the topic, and would require extra praise for a much lower score than the algorithm recommends if they suddenly improve. This would likely mean that the algorithm should change the recommendations for emotional support and slant as a learner progresses through a course.

We have also only examined adaptation to one polarized trait at once. Although the algorithm can adapt to polarized levels of emotional stability and conscientiousness at the same time, it may be that the recommendations would differ if participants see more than one polarized trait in a story. Additionally, the other three traits which did not seem to matter may interact with conscientiousness and emotional stability. For example, a neurotic learner who is an extrovert may receive different feedback to an introverted neurotic learner. Investigating this will require further story development.

We have investigated the adaptations in feedback to six scores, roughly mapped to grade boundaries (e.g. A-F in the UK). However, we presented a percentage score to participants, thus we do not know what to do for an exact pass (we provided a marginal pass and a marginal fail). Furthermore, although we know that the type of emotional support changes as grades increase, we do not know the exact point at which this change should happen. Further experiments could be devised where participants see a random percentage score.

The methodology could also be adapted to deal with performance on multiple topics. This would enable an intelligent conversational agent to provide feedback on an end of term report, for example.

Our research has taken place with a western perspective (using participants the US). The adaptations that we have discovered are unlikely to be appropriate in other cultures which have different perspectives. The methodology we have used in this paper could be used to examine how adaptations occur in other cultures, however this would require translation and re-validation of the stories.

When the Emotional Support statements were selected, we chose the ones that were most reliably categorized, as this provided us with the best way of investigating how the use of the categories differed. However, we do not know how emotionally supportive our statements actually are. It may be that our requirement for validated categories led to the elimination of the statements that are in fact best at supporting negative affective states. Further investigations need to establish how effective our remaining emotional support statements are at mitigating negative affect, and new ones developed if necessary. Additionally, we have only examined the use of emotional support category, however it may be that some statements in those categories were used more than others, and that this depends on learner score and personality. A future study could investigate this and discover which statements in each category are thought of as the best by participants.

A final limitation is that we have not empirically investigated the effect of the adapted feedback on the motivational levels of real learners in a longitudinal study.

REFERENCES

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., and Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. *IJAIED*, 24(4):387–426.

Assor, A., Kaplan, H., Kanat-Maymon, Y., and Roth, G. (2005). Directly controlling teacher behaviors as predictors of poor motivation and engagement in girls and boys: The role of anger and anxiety. *Learn Instr*, 15(5):397 – 413.

Bandura, A. (1994). *Self-efficacy*. Wiley Online Library.

Bandura, A. (2012). On the functional properties of perceived self-efficacy revisited. *J Manage*, 38(1):9–44.

Barbee, A. P., Cunningham, M. R., Winstead, B. A., Derlega, V. J., Gulley, M. R., Yankeelov, P. A., and Druen, P. B. (1993). Effects of gender role expectations on the social support process. *J Soc Issues*, 49(3):175–190.

Beal, C. and Lee, H. (2005). Creating a pedagogical model that uses student self reports of motivation and mood to adapt its instruction. In *Workshop on Motivation and Affect in Educational Software, AIED 05*.

Blanchard, E. and Frasson, C. (2004). An autonomy-oriented system design for enhancement of learner's motivation in e-learning. In *Intelligent Tutoring Systems*. Springer Berlin Heidelberg.

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., and Lester, J. (2008). Balancing cognitive and motivational scaffolding in tutorial dialogue. In *Intelligent Tutoring Systems*, pages 239–249. Springer.

Brave, S., Nass, C., and Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *IJHCI*, 62(2):161–178.

Burleson, W. and Picard, R. W. (2007). Gender-specific approaches to developing emotionally intelligent learning companions. *Intelligent Systems, IEEE*, 22(4):62–69.

Calvo, R. A. and Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *J Eng Educ*, 99(4):427–438.

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* World Book Co.

Cerri, S., Clancey, W., Papadourakis, G., and Panourgia, K.-K., editors (2012). *Intelligent Tutoring Systems: Proceedings of the 11th International Conference, ITS 2012*, volume 7135.

Cianci, M., Klein, H. J., and Seijts, G. H. (2010). The effect of negative feedback on tension and subsequent performance: The main and interactive effects of goal content and conscientiousness. *Appl Psychol*, 95(4):618–630.

Cobb, S. (1976). Social support as a moderator of life stress. *Psychosom Med*, 38(5):300–314.

Colquitt, J. A. and Simmering, M. J. (1998). Conscientiousness, goal orientation, and motivation to learn during the learning process: A longitudinal study. *J Appl Psychol*, 83(4):654.

Costa, P. T. and McCrae, R. R. (1992). *NEO personality Inventory professional manual*. Odessa, FL: Psychological Assessment Resources.

Csikszentmihalyi, M. (1988). *The flow experience and its significance for human psychology*, pages 15–35. Optimal experience. Cambridge University Press, Cambridge, UK.

Cutrona, C. E. (1996). *Social support in couples: Marriage as a resource in times of stress*, volume 13. Sage Publications.

Cutrona, C. E. and Russell, D. W. (1990). Type of social support and specific stress: Toward a theory of optimal matching. In *Social Support: an interactional view*. John Wiley and Sons.

Deci, E. L. and Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. *Adv Exp Soc Psychol*, 13(2):39–80.

Deci, E. L. and Ryan, R. M. (1985). *Self-Determination*. Wiley Online Library.

Deci, E. L. and Ryan, R. M. (2002). *Self-determination research: Reflections and future directions*, pages 431–441. Handbook of self-determination theory research. University of Rochester Press, Rochester,NY.

Del Soldato, T. and Du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *IJAIED*, 6:337–378.

Dennis, M. (2014). *Adapting Feedback to Learner Personality to Increase Motivation*. PhD thesis, University of Aberdeen.

Dennis, M., Masthoff, J., and Mellish, C. (2012a). Adapting performance feedback to a learner's conscientiousness. In *UMAP*, volume 7379 of *LNCS*, pages 297–302. Springer.

Dennis, M., Masthoff, J., and Mellish, C. (2012b). The quest for validated personality trait stories. In *Proceedings of IUI 2012*, pages 273–276, New York, NY, USA. ACM.

Dennis, M., Masthoff, J., and Mellish, C. (2012c). Towards a model of personality, affective state, feedback and learner motivation. In *UMAP Workshops '12*.

Dennis, M., Masthoff, J., and Mellish, C. (2013). Does learner conscientiousness matter when generating emotional support in feedback? In *ACII 2013*, pages 209–214.

Dennis, M., Masthoff, J., Pain, H., and Mellish, C. (2011). Does self-efficacy matter when generating feedback? In *Artificial Intelligence in Education*, volume 6738 of *LNCS*, pages 444–446. Springer Berlin Heidelberg.

Desmarais, M. and Baker, R. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted Interaction*, 22(1-2):9–38.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annu Rev Psychol*, 41(1):417–440.

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., el Kaliouby, R., Picard, R., et al. (2008). Autotutor detects and responds to learners affective and cognitive states. In *Workshop on Emotional and Cognitive Issues, ITS*.

El-Bishouty, M., Chang, T.-W., Graf, S., Kinshuk, and Chen, N.-S. (2014). Smart e-course recommender based on learning styles. *J. of Computers in Education*, 1(1):99–111.

Eysenck, H. J. (2013). *The Structure of Human Personality (Psychology Revivals)*. Routledge.

Fayard, J. V., Roberts, B. W., Robins, R. W., and Watson, D. (2012). Uncovering the affective core of conscientiousness: The role of self-conscious emotions. *J Pers*, 80(1):1–32.

Fogg, B. J. and Nass, C. (1997). Silicon sycophants: the effects of computers that flatter. *Int J Hum-Comput St*, 46(5):551–561.

Gilliland, A. L. (2011). After praise and encouragement: Emotional support strategies used by birth doulas in the usa and canada. *Midwifery*, 27(4):525–531.

Goldberg, L. (1993). The structure of phenotypic personality traits. *Am Psychol*, 48:26–34.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., and Gough, H. C. (2006). The international personality item pool and the future of public-domain personality measures. *J Res Pers*, 40:84–96.

Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *J Res Pers*, 37(6):504–528.

Graham, S. and Weiner, B. (1996). Theories and principles of motivation. In *Educational Psychology*, chapter 4, pages 63–84. MacMillan.

Hone, K. (2006). Empathic agents to reduce user frustration: The effects of varying agent characteristics. *Interact Comput*, 18(2):227–245.

Jackson, G. T. and Graesser, A. C. (2007). Content matters: An investigation of feedback categories within an its. *Fr Art Int*, 158:127.

John, O. P. and Srivastava, S. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Johnson, D., Gardner, J., and Wiles, J. (2004). Experience as a moderator of the media equation: the impact of flattery and praise. *Int. J Hum-Comput St*, 61(3):237–258.

Judge, T. A., Erez, A., Bono, J. E., and Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *J Pers Soc Psychol*, 83(3):693–

710.

Keller, J. and Suzuki, K. (2004). Learner motivation and e-learning design: A multinationally validated process. *J Educ Media*, 29(3):229–239.

Klein, J., Moon, Y., and Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interact Comput*, 14(2):119–140.

Lane, H., Cahill, C., Foutz, S., Auerbach, D., Noren, D., Lussenhop, C., and Swartout, W. (2013a). The effects of a pedagogical agent for informal science education on learner behaviors and self-efficacy. In *Artificial Intelligence in Education*, volume 7926 of *LNCS*, pages 309–318. Springer Berlin Heidelberg.

Lane, H., Yacef, K., Mostow, J., and Pavlik, P., editors (2013b). *Artificial Intelligence in Education: PROC of AIED 2013.*, volume 7926.

Lee, E.-J. (2008). Flattery may get computers somewhere, sometimes: The moderating role of output modality, computer gender, and user gender. *Int. J. of Human-Computer Studies*, 66(11):789–800.

Magai, C. and McFadden, S. (1995). *The Role of Emotions in Social and Personality Development*. Plenum Press.

Major, D. A., Turner, J. E., and Fletcher, T. D. (2006). Linking proactive personality and the big five to motivation to learn and development activity. *J. of Applied Psychology*, 91(4):927.

Martin, G. N.n, M., Alvarez, A., Fernandez-Castro, I., and Urretavizcaya, M. (2011). Adapted feedback supported by interactions of blended-learning actors: A proposal. In *AIED*, volume 6738 of *LNCS*, pages 205–212.

Masthoff, J. (1997). *An agent-based interactive instruction system*. PhD thesis, Technical University, Eindhoven, The Netherlands.

Masthoff, J. (2006). The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. *5th Wkshop on User-Centred Design and Evaluation of Adaptive Systems*, 1:460–469.

McCrae, R. R. and John, O. P. (1992). An introduction to the five-factor model and its applications. *J Pers*, 60(2):175–215.

McQuiggan, S., Mott, B., and Lester, J. (2008). Modeling self-efficacy in intelligent tutoring systems: An inductive approach. *UMUAI*, 18(1-2):81–123.

Meyer, D. K. and Turner, J. C. (2002). Discovering emotion in classroom motivation research. *Educ Psychol*, 37(2):107–114.

MT (2012). Amazon mechanical turk. http://www.mturk.com.

Nguyen, H. and Masthoff, J. (2009). Designing empathic computers: the effect of multimodal empathic feedback using animated agent. In *Persuasive*, page 7.

Nkambou, R. (2006). Towards affective intelligent tutoring system. In *Workshop on Motivational and Affective Issues in ITS. ITS 2006*, pages 5–12.

Nunes, M. A. S. N. (2008). *Recommender Systems based on Personality Traits*. PhD thesis, Universite Montpellier.

Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., Zoll, C., and Hall, L. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents. In *AAMAS*, pages 194–201.

Picard, R. W. and Klein, J. (2002). Computers that recognise and respond to user emotion: theoretical and practical implications. *Interact Comput*, 14(2):141–169.

Porayska-Pomsta, K. and Mellish, C. (2013). Modelling human tutors' feedback to inform natural language interfaces for learning. *Int J Hum-Comput St*, 71(6):703–724.

Prendinger, H., Dohi, H., Wang, H., Mayer, S., and Ishizuka, M. (2004). Empathic embodied interfaces: Addressing users' affective state. In *Wks on. Affective Dialog Systems*, pages 53–64. Springer.

Prendinger, H. and Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users'affective states. *Appl Artif Intell*, 19(3-4):267–285.

Robison, J., McQuiggan, S., and Lester, J. (2009a). Evaluating the consequences of affective feedback in intelligent tutoring systems. In *ACII Workshops, 2009.*, pages 1–6. IEEE.

Robison, J., McQuiggan, S., and Lester, J. (2010). Developing empirically based student personality profiles for

affective feedback models. In *ITS*, pages 285–295. Springer.

Robison, J. L., Mcquiggan, S. W., and Lester, J. C. (2009b). Modeling task-based vs. affect-based feedback behavior in pedagogical agents: An inductive approach. In *AIED*, pages 25–32.

Rook, K. and Underwood, L. (2000). Social support measurement and iintervention: comments and future directi. In *Social Support measurement and interventions: A guide for health and social scientists*, pages 311–334. Oxford University Press.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychol Monogr*, 80:1–26.

Rusting, C. L. and Larsen, R. J. (1997). Extraversion, neuroticism, and susceptibility to positive and negative affect: A test of two theoretical models. *Pers Indiv Differ*, 22(5):607–612.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions* 1. *Contemp Educ Psychol*, 25(1):54–67.

Ryan, R. M., Kuhl, J., and Deci, E. L. (1997). Nature and autonomy: An organizational view of social and neurobiological aspects of self-regulation in behavior and development. *Dev Psychopathol*, 9(04):701–728.

Saucier, G. (1994). Mini-markers: A brief version of goldberg's unipolar big-five markers. *Personality Assessment*, 63(1):506–516.

Schunk, D. H. and Ertmer, P. A. (1999). Self-regulatory processes during computer skill acquisition: Goal and self-evaluative influences. *Educ Psychol*, 91(2):251.

Soldz, S. and Vaillant, G. E. (1999). The big five personality traits and the life course: A 45-year longitudinal study. *J Res Pers*, 33(2):208 – 232.

Turner, J. C., Thorpe, P. K., and Meyer, D. K. (1998). Students' reports of motivation and negative affect: A theoretical and empirical analysis. *J Educ Psychol*, 90(4):758.

van Deemter, K., Gatt, A., Sluis, I. v. d., and Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Sci*, 36(5):799–836.

van der Sluis, I. and Mellish, C. (2010). Towards empirical evaluation of affective tactical nlg. In *Empirical methods in natural language generation*, pages 242–263. Springer.

VanLehn, K., Burleson, W., Girard, S., Chavez-Echeagaray, M. E., Gonzalez-Sanchez, J., Hidalgo-Pontet, Y., and Zhang, L. (2014). The affective meta-tutoring project: Lessons learned. In *ITS*, pages 84–93. Springer.

Varnosfadrani, A. D. and Basturkmen, H. (2009). The effectiveness of implicit and explicit error correction on learnersâĂŹ performance. *System*, 37(1):82 – 98.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., and Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *Int J Hum-Comput St*, 66(2):98–112.

Watson, D., A, C. L., and Tellegen, A. (1998). Development and validation of brief measures of positive and negative affect: the panas scales. *J Pers Soc Psychol*, 54:1063–1070.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychol Rev*, 92(4):548.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and Cognitive Approaches to the Communication of Knowledge*. Morgan Kaufmann Publishers Inc.

Wigfield, A., Eccles, J. S., Roeser, R., and Schiefele, U. (2008). Development of achievement motivation. *Child and adolescent development: An advanced course*, 1:406–434.

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., and Picard, R. (2009). Affect-aware tutors: Recognising and responding to student affect. *Int J Learn Technol*, 4(3/4):129–164.

Zakharov, K., Mitrovic, A., and Johnston, L. (2008). Towards emotionally-intelligent pedagogical agents. In *ITS*, volume 5091 of *LNCS*, pages 19–28. Springer.