Who Benefits from Confusion Induction during Learning? An Individual Differences Cluster Analysis

Blair Lehman[1], Sidney D'Mello[2], and Art Graesser[1]

[1] University of Memphis, Memphis, TN 38152, USA
[balehman, graesser]@memphis.edu
[2] University of Notre Dame, Notre Dame, IN 46556, USA
sdmello@nd.edu

Abstract. Recent research has indicated that learning environments that intentionally induce confusion to promote deep inquiry can be beneficial for learning if students engage in confusion resolution processes and if relevant scaffolds are provided. However, it is unlikely that these environments will benefit all students, so it is necessary to identify the student profiles that most benefit from confusion induction. We investigated how individual differences (e.g., prior knowledge, interest, attributional complexity) impacted confusion and learning outcomes in an environment that induced confusion via false system feedback (e.g., negative feedback after a correct response). A *k*-means cluster analysis revealed four clusters that varied on cognitive ability and cognitive drive. We found that students in the high cognitive ability + high cognitive drive cluster reported more confusion after receiving false feedback compared to the other clusters. These students also performed better on tasks requiring knowledge transfer, but only when they were meaningfully confused.

Keywords: confusion, individual differences, cluster analysis, false feedback, intelligent tutoring systems, learning

1 Introduction

Recent research has shown that intelligent tutoring systems (ITS) are an effective and comparable alternative to novice as well as accomplished (or expert) human tutors [1]. ITSs are effective because they are interactive, provide immediate feedback, and provide individualized instruction, which are similar to the techniques used by human tutors [2-4]. ITSs must attend to both student cognition and affect in order to provide effective, individualized instruction. Recently many ITSs have adopted this approach and provide individualized instruction that focuses on the affective states of the student in addition to their cognitive states (e.g., [5-9]).

Confusion is one affective state that is particularly important to the learning process. Confusion is an epistemic or knowledge affective state [10-11] that occurs when students confront contradictions, anomalies, and discrepant events that create impasses and when students are uncertain about how to proceed [12-14]. In other words, confusion signals that there is something wrong with the state of one's knowledge [15]. Increased experiences of confusion have been linked to learning at deeper levels [16-17]. Importantly, it is not the mere experience of confusion that presumably benefits learning; instead it is the effortful cognitive activities inspired by confusion resolution (e.g., reflection, deliberation) that underlie improvements in learning [14,18]. However, all experiences of confusion are not expected to be beneficial for learning. Learning is unlikely to occur when students are unable to resolve their confusion either due to a lack of motivation, ability, or instructional scaffolds. This type of unresolved or hopeless confusion should be contrasted with productive confusion, which can eventually be resolved [18].

It has been suggested that ITSs can capitalize on the benefits of confusion by adaptively responding to natural occurrences of confusion. For example, UNC-ITSpoke is a novel ITS that provides adaptive feedback and instruction based on the correctness and level of certainty in a student's spoken response [8]. Similarly, the *Affective AutoTutor* provides motivational and supportive statements to help students persist in the learning task when it senses that they are confused [19]. Both systems have been shown to be more effective than non-affective counterparts, but only for a subset of students. This suggests that affective response strategies must take into consideration individual differences, an idea that is at the core of this paper.

A somewhat different approach to *reactively* capitalizing on opportunities afforded by naturally occurring confusion, is a *proactive* approach in which learning environments create learning opportunities through confusion induction. We have experimented with this approach and had some success with confusion induction through the presentation of system breakdowns [20], contradictory information [21-22], and false system feedback [23]. Space limitations preclude a detailed discussion of these studies, however, they all revealed that confusion induction and regulation was a successful learning strategy, but only for a subset of students. It is important, then, to understand the individual differences that influence the incidence of confusion itself, attempts at confusion resolution, and learning outcomes associated with these processes. In line with this, the present paper investigates the impact of individual differences in a learning environment that induces confusion via false feedback.

Our focus is on the analysis of a data set collected from a study in which students attempted to learn research methods while interacting with an animated tutor agent [23]. Students diagnosed the flaws in research case studies and received feedback (accurate or inaccurate) on the quality of the flaw diagnosis. The false feedback was expected to trigger confusion, which would inspire deeper processing, and the learning environment provided explanatory texts to aid confusion resolution. We found that students learned the most when they received false feedback and were successfully confused by the feedback. The previous paper [23] did not analyze individual differences associated with successful learning in this environment. To address this issue, we investigated whether individual differences impacted (1) the effectiveness of false feedback as a method of confusion induction and (2) learning gains in a false feedback learning environment. The individual difference measures included in the present paper were prior knowledge, confidence in the ability to learn from a computer tutor, perceptions of research methods (interest, willingness to put in effort to learn), the School Failure Tolerance scale (SFT, [24]), the Attributional Complexity scale (ACS, [25]), and the Theory of Intelligence scale (TOI, [26]). These measures were selected because they assess preferences for challenging material and responses to academic challenges like those posed by confusion inducing stimuli.

2 Method

2.1 Participants

Participants (called students for the remainder of the paper) were 167 undergraduate students from a mid-south university in the US who received course credit for participation. Data from eleven students was not included in the present analyses because they did not complete the individual difference measures (described below). There were 115 females and 41 males in the sample, 62% of which were African-American, 32% Caucasian, 4% Hispanic, and 2% Asian.

2.2 Design and Manipulation

The experiment had a within-subjects design with four conditions, one on each research method topic (control group, experimenter bias, random assignment, replication): *positive-positive*, *positive-negative*, *negative-negative*, and *negative-positive*. Students completed two learning sessions in which they received accurate feedback and two sessions of false feedback. It was not guaranteed, however, that each student would be in all four conditions due to the fact that condition assignment was partially dependent upon student responses. Order of feedback condition, order of topics, and assignment of topics to conditions were counterbalanced across students with a Graeco-Latin Square.

False feedback was delivered during dialogues with an animated tutor agent over the course of identifying flaws in research case studies. Each study contained one subtle methodological flaw pertaining to one of four topics. The four feedback conditions were based on student response quality (*positive*: correct and *negative*: incorrect) and tutor agent feedback (*positive*: "Yes, that's right" and *negative*: "No, that's not right"). Students who responded correctly either received accurate, positive feedback (*positive-positive*) or inaccurate, negative feedback (*positive-negative*). Students in the *negative-negative* condition received accurate, negative feedback, whereas those in the *negative-positive* condition received inaccurate, positive feedback. It should be noted that all misleading information presented via false feedback was corrected at the end of each dialogue and participants were fully debriefed at the end of the experiment.

2.3 Procedure

The experiment occurred over two phases: (1) knowledge assessments and learning sessions and (2) individual difference measures.

Knowledge Tests. Research methods knowledge was assessed with a multiple-choice definition test and flaw identification task. The definition test consisted of eight multiple-choice questions. There was one question pertaining to each topic that was discussed in the learning sessions. In addition, there were four questions that pertained to topics not covered in the learning sessions (construct validity, correlational studies, generalizability, measure quality). The definition test was presented before and after all of the learning sessions had been completed (pretest and posttest, respectively). Two versions of the test were created and order of presentation was counterbalanced across students.

The flaw identification task consisted of a description of a previously unseen study and students were asked to identify flaw(s) in the study by selecting as many items as they wanted from a list of eight research methods topics. The list included four topics that could potentially be flawed (i.e., discussed in the learning sessions) and four distractor topics (i.e., not discussed in the learning sessions). Students also had the option of selecting that there was no flaw, although each study contained one flaw. Near and far transfer versions of studies were presented to students. The near transfer studies differed from the studies discussed in the learning sessions on surface features, whereas the far transfer studies differed on both surface and structural features. Each topic discussed during the learning sessions had one near and one far transfer study, resulting in eight transfer studies in all.

Learning Sessions. First, students signed an informed consent, completed a brief demographics questionnaire, and completed the pretest. Students then read a short introductory text on research methods. Next, students completed a survey about their perceptions of learning research methods (PLRM). These questions assessed student *interest* in and willingness to put in *effort* when learning about research methods and student *confidence* in the ability to learn from a computer tutor.

Students then began the first of four learning sessions. Each learning session consisted of four phases: manipulation, assumption check, remediation, and post-remediation. For the present paper only the manipulation and remediation phases are relevant and the others are not discussed here. The manipulation phase began with students reading a description of the study that was being discussed. Next, students were presented with a forced-choice question to diagnose the flaw in that study. When discussing the study with replication as its flaw, for example, the tutor agent asked the student "Was this a good or bad replication?" Students then selected one of the three response options: *target* (correct), *thematic miss* (incorrect but generally related to the concept), and *irrelevant distractor* (incorrect and not related to the concept). Students also rated whether they were confident or not confident in the correctness of their response prior to receiving feedback. The majority of students (80%) were confident in the correctness of their response [23]. The tutor agent then provided feedback about the quality of the response. Based on the condition, the feedback delivered could either be accurate or inaccurate, regardless of the actual quality of the response.

After receiving feedback, students were prompted to make a *post-feedback confusion judgment*. Students were prompted to indicate whether a classmate would be confused or not confused at this point in the learning session. The confusion prompt was phrased in this manner to avoid potential biases due to students' negative perceptions of being in a state of confusion [21]. Reports of confusion were found to be significantly related to increased student processing time after feedback [23]. Student processing time was assessed by asking students to indicate when they were ready to proceed with the learning session after receiving feedback.

In the remediation phase students were presented with an explanatory text to potentially alleviate their confusion. The texts were adapted from the electronic textbook that accompanies the *Operation ARA!* ITS [27]. Longer text reading times were considered to indicate greater depth of processing [28], which is ostensibly related to increased effort to resolve confusion. Post-feedback confusion judgments and explanatory text read times served as the learning process measures.

Individual Difference Measures. In addition to the PLRM (see above), students also completed three individual difference measures after the posttest: SFT [24], ACS [25], and TOI [26]. The SFT consists of three subscales: prefer difficult material, experience negative affect after failure, and take action after failure. These subscales describe the type of material students generally prefer (difficult vs. easy; *prefer difficult*) as well as the affective states that they experience (negative vs. positive; *negative affect*) and how they respond after failure (take action vs. avoid; *take action*).

The ACS consists of seven subscales. Only four of the subscales were used in the present analyses due to reliability issues within the current sample (see below). The four subscales used were motivation, metacognition, complex contemporary external explanations, and use of temporal dimension. These subscales assess the degree to which students look for (*motivation*) and monitor their own behavior for (*metacognition*) multiple explanations and prefer complex external explanations that are either temporally close (*contemporary*) or distant (*temporal*) from an event. The TOI has two subscales that represent either a theory that intelligence can be increased through effort and training (*incremental mindset*) or that people have a certain level of intelligence that cannot be altered (*entity mindset*). Reliability (Cronbach's alpha) for the nine subscales included in the analyses ranged from .616 to .915.

3 Results and Discussion

The analyses are divided into two sections. First, we conducted a *k*-means cluster analysis to group students with similar characteristics. Second, we investigated differences between clusters for the learning process and learning outcome measures.

3.1 Cluster Analysis

We used a *k*-means clustering method to group the 156 students into clusters. Students were grouped based on 14 attributes that included their pretest score; self-reported ACT score; interest, effort, and confidence from the PLRM; and the nine subscales from the SFT, ACS, and TOI. The *k* value was set to 4 based on an exploratory factor analysis and a hierarchical cluster analysis. We also experimented with *k's* of 3 and 5; however, the clusters were most distinct with $k = 4$.

ANOVAs indicated that 10 out of the 14 measures used to create the clusters significantly discriminated between clusters (*p's* < .05). *Incremental mindset* (TOI) was only marginally significant ($p < .1$), while *entity mindset* (TOI), *confidence* (PLRM), and *negative affect* (SFT) did not discriminate between clusters (*p's* > .1).

We correlated the individual clusters (dummy coded) and the 10 aforementioned measures in an attempt to name the clusters. Table 1 shows the pattern of correlations and the *N* for each cluster. *Cognitive Ability* (CA) and *Cognitive Drive* (CD) appeared to be the latent factors that distinguished the clusters. CA included pretest and ACT scores, whereas CD encompassed characteristics related to interest, effort, motivation, determination, and persistence. Thus the four clusters were named High CA + High CD (cluster 3), High CA + Low CD (cluster 1), Low CA + High CD (cluster 2), and Low CA + Low CD (cluster 4).

Table 1. Patterns in correlation matrix used for cluster naming

	Cluster 1 High CA + Low CD (*N* = 12)	**Cluster 2** Low CA + High CD (*N* = 68)	**Cluster 3** High CA + High CD (*N* = 32)	**Cluster 4** Low CA + Low CD (*N* = 44)
Cognitive Ability				
Pretest Score			+	-
ACT Score	+	-	+	-
Cognitive Drive				
PLRM: Interest		+	+	-
PLRM: Effort	-	+		
SFT: Prefer Difficult		+		-
SFT: Action	-	+	-	
ACS: Motivation	-		+	-
ACS: Metacognition	-			
ACS: Contemporary	-	+	+	
ACS: Temporal	-			

Notes. +'s or –'s indicate positive or negative correlations at $p < .10$.

3.2 Differences between Clusters

Next, we investigated differences between clusters for the learning process and learning outcome measures. Analyses were conducted separately for each type of learning session: positive-positive, positive-negative, negative-negative, and negative-positive. The High CA + Low CD cluster was not included in the present analyses due to the low *N* of 12. We conducted non-parametric Kruskal-Wallis tests with Mann-Whitney U post hoc tests when the variables were not normally distributed and ANOVAs with Bonferroni post hoc tests otherwise.

There were no significant cluster differences for the accurate feedback learning sessions (positive-positive, negative-negative). Thus, the discussion will focus on the false feedback learning sessions (positive-negative, negative-positive).

Learning process measures. There were marginally significant differences between clusters for the post-feedback confusion judgments in both false feedback learning sessions: positive-negative: $\chi^2(2, N = 119) = 5.47$, $p = .065$; negative-positive: $\chi^2(2, N = 99) = 4.56$, $p = .102$ (see Table 2). For the positive-negative sessions the High CA + High CD cluster reported significantly more confusion than the Low CA + Low CD cluster ($p = .034$). The other cluster comparisons were not significant. For the negative-positive sessions, the High CA + High CD cluster reported more confusion than the Low CA + High CD cluster ($p = .045$) and was the only significant cluster difference. These findings suggest that students must know enough and be sufficiently driven to recognize that there is a discrepancy in the system feedback.

Table 2. Descriptives for learning process measures

Measure	High CA + High CD	Low CA + High CD	Low CA + Low CD
Confusion (Proportion)			
Positive-Negative	.704	.475	.636
Negative-Positive	.630	.381	.412
Text Read Time M(SD) in secs			
Positive-Negative	75.5 (36.7)	68.5 (41.8)	78.2 (45.2)
Negative-Positive	97.9 (45.8)	78.2 (52.6)	62.6 (46.9)

There was a significant cluster difference in explanatory text reading times for the negative-positive sessions, $F(2, 96) = 3.55$, $p = .032$ but not for the positive-negative sessions ($p = .528$) (see Table 2). For the negative-positive sessions, the High CA + High CD cluster read for longer than Low CA + Low CD cluster ($p = .027$). The other cluster comparisons were not significant.

Learning outcome measures. Student performance on the definition posttest was assessed by selection of the correct answer option. For both transfer tasks student performance was assessed with hits (correctly identifying the presence of a flaw). There were no significant differences on the definition posttest for either of the false feedback learning sessions (*p's* > .1).

However, there were significant cluster differences on the flaw identification task (see Table 3). For the near transfer task, there were significant differences between clusters for the positive-negative sessions, $\chi^2(2, N = 118) = 6.24$, $p = .044$. The High CA + High CD ($p = .033$) and Low CA + High CD ($p = .026$) clusters performed better than the Low CA + Low CD cluster. The High CA + High CD and Low CA + High CD clusters did not significantly differ. There was not a significant difference between clusters for the negative-positive sessions ($p = .568$).

Table 3. Proportion of correct flaw detection for the flaw identification task

Measure	High CA + High CD	Low CA + High CD	Low CA + Low CD
Near Transfer			
Positive-Negative	.538	.466	.273
Negative-Positive	.500	.583	.471
Far Transfer			
Positive-Negative	.315	.169	.182
Negative-Positive	.545	.226	.318

There were significant differences between clusters for the negative-positive sessions for the far transfer task, $\chi^2(2, N = 97) = 7.32$, $p = .026$. The only significant cluster difference was that the High CA + High CD cluster performed better than the

Low CA + High CD cluster (p = .008). There was not a significant cluster difference for the positive-negative sessions (p = .248).

These findings show that false feedback can promote learning at a deeper level, but that false feedback was most beneficial for a particular group of students (i.e., High CA + High CD). It is interesting, however, that the High CA + High CD cluster only performed better on the near transfer task when in the positive-negative learning sessions and the far transfer task when in the negative-positive learning sessions. We hypothesized that the increased performance on the transfer tasks could be related to the increased effort to resolve confusion (i.e., longer text read times) by the High CA + High CD cluster when in the false feedback learning sessions.

To address this hypothesis, we explored cluster differences on the transfer tasks when students were divided into those who read the text more quickly and read more slowly via a median split. There were no significant cluster differences when students read more quickly (*p's* > .05). However, when students read for longer, the High CA + High CD cluster performed better than the Low CA + Low CD cluster on the near transfer task, $\chi^2(2, N = 61) = 6.92, p = .031$, and better than the Low CA + High CD cluster on the far transfer task, $\chi^2(2, N = 62) = 5.88, p = .053$, for the positive-negative sessions. A similar pattern was found for the far transfer task in the negative-positive sessions, $\chi^2(2, N = 48) = 6.72, p = .035$, with the High CA + High CD cluster outperforming the Low CA + High CD cluster. These findings suggest that effortful attempts at confusion resolution were needed to perform well on the transfer tasks.

4 General Discussion

Recent research has focused on developing ITSs that promote learning through adaptive scaffolding based on both student cognition and affect [5-9]. It is also important, however, to determine the individual differences (e.g., interest, prior knowledge, learning styles) that influence the effectiveness of these affect-aware learning interventions because there is no one-size-fits-all approach to learning. As a step in this direction, we investigated the relationship between individual differences, confusion, and learning within a learning environment that proactively induces confusion as a means to promote deep inquiry.

A cluster analysis on a number of individual difference measures indicated that students differed with respect to cognitive ability and cognitive drive. We found that students with a combination of high cognitive ability and high cognitive drive benefited the most from the current learning environment. These students were successfully confused by the false feedback (induction) and performed better on the transfer tasks (learning). It is critically important to note that the high cognitive ability and high cognitive drive cluster did not simply learn more than the other clusters in all learning sessions. This cluster of students only outperformed the other clusters on transfer tasks when they received false feedback. Moreover, these students only outperformed the other clusters on the difficult far transfer task when they received false feedback and read the text for longer in an effort to resolve their confusion.

Despite these promising findings, some critics might object to the use of false feedback due to the potential for negative impacts on learning. This is a valid concern for more authentic learning contexts and for this reason it is important to understand which students do and do not benefit from this method of confusion induction. However, it is important to note that previous analyses showed that inaccurate feedback did not negatively impact learning in the present experimental research [23].

Now that we have identified which students benefited from false feedback in the present learning environment, the next step is to determine how to help other students benefit from experiences of confusion during learning. There are two aspects of the learning environment that can be targeted. First, false feedback is not the only method of confusion induction. It may be the case that productive confusion is triggered by different stimuli for different students (e.g., system breakdowns [20], contradictory information [21-22]). Second, presentation of an explanatory text may not have been the most appropriate method of confusion remediation for all students. Students who are lower in cognitive ability and cognitive drive may need more adaptive, targeted scaffolding (e.g., critical information [8] or encouragement [19]). Or perhaps, it is simply better to avoid confusing these students and rely on more explanation-focused pedagogical approaches. Future research will need to differentially adapt both confusion induction and remediation strategies for different individual differences to maximize learning for all students.

Acknowledgments. This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of NSF.

References

1. VanLehn, K.: The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46, 197--221 (2011)
2. D'Mello, S., Lehman, B., Person, N.: Expert tutors feedback is immediate, direct, and discriminating. In: Murray, C., Guesgen, H. (eds.) Proceedings of the 23rd Florida Artificial Intelligence Research Society Conference, pp. 595--560. AAAI Press, Menlo Park (2010)
3. Graesser, A., Person, K., Magliano, J.: Collaborative dialogue patterns in naturalistic one-to-one tutoring. Applied Cognitive Psychology, 9, 495--522 (1995)
4. Lepper, M., Woolverton, M.: The wisdom of practice: Lessons learned from the study of highly effective tutors. In: Aronson, J. (ed.) Improving Academic Achievement: Impact of Psychological Factors on Education, pp. 135--158. Academic Press, Orlando (2002)
5. Arroyo, I., Woolf, B. Cooper, D., Burleson, W., Muldner, K., Christopherson, R.: Emotion sensors go to school. In: Dimitrova, V., Mizoguchi, R., Du Boulay, B., Graesser, A. (eds.) Proceedings of 14th International Conference on Artificial Intelligence in Education, pp. 17--24. IOS Press, Amsterdam (2009)
6. Conati, C., Maclaren, H.: Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction, 19(3), 267--303 (2009)
7. D'Mello, S., Craig, S., Fike, C., Graesser, A.: Responding to learners' cognitive-affective states with supportive and shake-up dialogues. In: Jacko, J. (ed.) Human-Computer Interaction. Ambient, Ubiquitous, and Intelligent Interaction, pp. 595--604. Springer, Berlin/Heidelberg (2009)

8. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication, 53, 1115--1136 (2011)
9. Robison, J., McQuiggan, S., Lester, J.: Evaluating the consequences of affective feedback in intelligent tutoring systems. In: Muhl, C., Heylen, D., Nijholt, A. (eds.) Proceedings of International Conference on Affective Computing & Intelligent Interaction, pp. 37--42. IEEE Computer Society Press, Los Alamitos (2009)
10. Pekrun, R., Stephens, E.: Academic emotions. In: Urdan, T. (ed.) APA educational psychology handbook, vol. 2, pp. 3--31. American Psychological Association, Washington, DC (2012)
11. Silvia, P.: Confusion and interest: The role of knowledge emotions in aesthetic experience. Psychology of Aesthetics, Creativity, and the Arts, 4, 75--80 (2010)
12. Carroll, J., Kay, D.: Prompting, feedback and error correction in the design of a scenario machine. International Journal of Man-Machine Studies 28, 11--27 (1988)
13. D'Mello, S., Graesser, A.: Confusion. In: Pekrun, R., Linnenbrink-Garcia, L. (eds.) Handbook of Emotions and Education. Taylor & Francis, New York (in press)
14. VanLehn, K., Siler, S., Murray, C., Yamauchi, T., Baggett, W.: Why do only some events cause learning during human tutoring? Cognition and Instruction, 21, 209--249 (2003)
15. Piaget, J.: The origins of intelligence. International University Press, New York (1952)
16. Craig, S., Graesser, A., Sullins, J., Gholson, B.: Affect and Learning: An exploratory look into the role of affect in learning. Journal of Educational Media, 29, 241--250 (2004)
17. Graesser, A., Chipman, P., King, B., McDaniel, B., D'Mello, S.: Emotions and learning with AutoTutor. In: Luckin, R., Koedinger, K., Greer, J. (eds.) Proceedings of the 13th International Conference on Artificial Intelligence in Education, pp. 569--571. IOS Press, Amsterdam (2007)
18. D'Mello, S., Graesser, A.: Dynamics of affective states during complex learning. Learning and Instruction, 22, 145--157 (2012)
19. D'Mello, S., Lehman, B., Sullins, J., Daigle, R., Combs, R., Vogt, K., et al.: A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In: Kay, J., Aleven, V. (eds.) Proceedings of the 10th International Conference on Intelligent Tutoring Systems, pp. 245--254. Springer, Berlin/Heidelberg (2010)
20. D'Mello, S., Graesser, A.: Inducing and tracking confusion and cognitive disequilibrium with breakdown scenarios. (in review)
21. D'Mello, S., Lehman, B., Pekrun, R., Graesser, A.: Confusion can be beneficial for learning. Learning and Instruction. (in press)
22. Lehman, B., D'Mello, S., Strain, A., Mills, C. Gross, M., Dobbins, A., Wallace, P., et al.: Inducing and tracking confusion with contradictions during complex learning. International Journal of Artificial Intelligence in Education. (in press)
23. Lehman, B., D'Mello, S., Graesser, A.: False feedback can improve learning when you're productively confused. (in review)
24. Clifford, M.: Failure tolerance and academic risk-taking in ten- to twelve-year-old students. British Journal of Educational Psychology, 58, 268--294 (1988)
25. Fletcher, G., Danilovics, P., Fernandez, G., Peterson, D., Reeder, G.: Attributional complexity: An individual differences measure. Journal of Personality and Social Psychology, 51, 875-884 (1986)
26. Dweck, C.: Self theories: Their role in motivation, personality and development. Taylor & Francis/Psychology Press, Philadelphia (1999)
27. Halper, D., Millis, K., Graesser, A., Butler, H., Forsyth, C., & Cai, Z.: Operation ARA: A computerized learning game that teaches critical thinking and scientific reasoning. Thinking Skills and Creativity, 7, 93--100 (2012)
28. Craik, F., Tulving, E.: Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 104, 268--294 (1975)