Analytics of the Effects of Video Use and Instruction to Support Reflective Learning

Dragan Gašević
Athabasca University
Athabasca, AB, Canada
dgasevic@acm.org

Negin Mirriahi
University of New South Wales
Sydney, NSW, Australia
negin.mirriahi@unsw.edu.au

Shane Dawson
University of South Australia
Adelaide, SA, Australia
shane.dawson@unisa.edu.au

ABSTRACT
Although video annotation software is no longer considered as a new innovation, its application in promoting student self-regulated learning and reflection skills has only begun to emerge in the research literature. Advances in text and video analytics provide the capability of investigating students' use of the tool and the psychometrics and linguistic processes evident in their written annotations. This paper reports on a study exploring students' use of a video annotation tool when two different instructional approaches were deployed – graded and non-graded self-reflection annotations within two courses in the performing arts. In addition to counts and temporal locations of self-reflections, the Linguistic Inquiry and Word Counts (LIWC) framework was used for the extraction of variables indicative of the linguistic and psychological processes associated with self-reflection annotations of videos. The results indicate that students in the course with graded self-reflections adopted more linguistic and psychological related processes in comparison to the course with non-graded self-reflections. In general, the effect size of the graded reflections was lower for students who took both courses in parallel. Consistent with prior research, the study identified that students tend to make the majority of their self-reflection annotations early in the video time line. The paper also provides several suggestions for future research to better understand the application of video annotations in facilitating student learning.

Categories and Subject Descriptors
J.1 [Administrative Data Processing] Education; K.3.1 [Computer Uses in Education] Computer-assisted instruction (CAI)

General Terms
Algorithms, Measurement, Human Factors

Keywords
Learning analytics, self-reflections, text analysis, metacognition

1. INTRODUCTION
Video technologies have been well adopted across multiple disciplines such as medical education [18], teacher education [7], and the performing arts [10, 22] primarily with the goal of supporting and developing students' meta-cognitive capacity. Commonly, these technologies are deployed in the education context to provide opportunity for students to reflect on and monitor their learning of new practical skills. This instructional strategy of asking students to review video recordings of their presentations, performances, or interactions with others allows students to critically reflect on their skills. In so doing, students indicate what they believe they have mastered as well as identifying the possible areas for improvement or to establish future learning goals [18]. The introduction of video annotation software as a core part of the learning design has been argued to develop self-directed learning skills [11]. These captured and shared time specific annotations provide opportunity for individual reflection, instructor feedback and social learning with peers. Given the nature and deployment of these technologies, user interactions can be extensively captured for post-hoc analyses. Despite the growth in video annotations and the user interaction logs that these tools collect, the majority of research on the use of videos for self-reflection have tended to rely on self-reports [1, 7, 8, 18, 26]. While these reports provide valuable insights into the role of video and self-reflection for improving meta-cognitive skills, there is insufficient evidence to determine precisely when, how and to what extent video and video annotations are most effective for developing meta-cognitive skills. In particular, it is challenging to capture objective data on the students' use of annotation software and analyse students' psychological and linguistic processes due to the vast amount of text [35]. As such, new analytical methods are required that can provide new insights into the role video annotations play in developing student self-reflection.

To date, there is limited research that incorporates automated linguistic and psychometric analysis methods for examining students' self-reflections when using video annotation software in courses with different pedagogical approaches, such as graded and non-graded reflections. Hence, this study uses the Linguistic Inquiry and Word Count (LIWC) framework [35] to investigate the type and frequency of psychological and linguistic terms most frequently observed when reflective assessments using video annotation software are graded compared to when they are optional and non-graded. Patterns in the number of annotations students make and their temporal location in the recorded videos are also explored.

LAK '14, March 24 - 28 2014, Indianapolis, IN, USA

http://dx.doi.org/10.1145/2567574.2567590

2. BACKGROUND
2.1 Video-based Annotation Technologies
Video annotation software for streaming educational media is not a novel technology, with the tool being available for over a decade in various forms. One of the earlier prototypes, the web-based Microsoft Research Annotation System (MRAS), provided the basic functionality for students to view streaming video and annotate time-stamped notes that are then available for future retrieval [1]. The MRAS software highlighted the potential for moving video from a passive medium to a more collaborative learning resource. However, the complexity for adoption and integration lay in how users would embrace such a change in practice where the use of annotations has traditionally been embedded in a pen and paper paradigm. Studies comparing the use of MRAS with traditional pen and paper note-taking in a live lecture showed both a preference for the use of the annotation software and a greater number

of annotations were made when using the tool.

Since the founding conception of video annotation software, there have been numerous developments and feature inclusions. For example, while the Media Annotation Tool (MAT) provided similar features to the MRAS, it also has the capacity to automatically generate an annotative learning cycle that consists of the initial notes, comments from peers, concluding notes, teacher feedback, and final reflections [8]. Pilot study results from interviews and surveys with teacher education students using MAT suggested that the video annotations were effective in aiding student learning. The authors noted that peer and instructor feedback on the video were the most valued area of the annotative learning cycle. However, the study did not provide any data related to the analysis of the types of reflections made - student, peer, or instructor [7]. Similarly, a study of medical students use of video annotation software, Digital Video Digital University (DiViDu), indicated that 90% of the participants felt that reviewing and reflecting on recordings of their simulated patient consultations made them aware of areas requiring further improvement [18]. Again, direct analysis of the types of annotations made and how these contributed to student learning was lacking. However, these studies and many others do point to the value these previously passive technologies can now bring to the learning process for improving student engagement and self-reflection, bounded within a more social context.

While most studies on the use of video annotation software gathered data from students' self-reports [1, 7, 8, 18, 26], few have leveraged trace data about students' use of the software and applied analytics to understand students' interaction with the technology. Previous studies on lecture video recordings have shown that trace data from log reports can objectively reveal patterns in students' access of recordings over a semester [3, 32] and, when aggregated with assessment scores, whether viewing the video lectures has any pedagogical impact [28]. Furthermore, a study using text mining techniques to analyse students' chat messages in a live video streaming environment, provides an alternative approach to traditional qualitative labour-intensive content analysis [17].

With advancements in text and data mining techniques uncovering the copious amount of logged information in lecture video recordings, research on video annotation tools are moving towards the use of similar analytical strategies. A study investigating students' note-taking behaviours of lecture videos using a video annotation software called Interactive Shared Education Environment (ISEE), analyzed logged data to discover patterns in the length and frequency of the notes [29]. The analytics revealed that students who used the annotation feature in ISEE had fewer words per note than their peers using traditional methods. While these findings differ to the earlier study on MRAS, the authors concluded that the use of ISEE enables better note taking with fewer words as the software reduces the non-learning cognitive load that is associated with the traditional modes of note taking. Interestingly, the authors also observed that both groups of students had a higher number of notes clustered in the first third of the video.

2.2 Issues in Educational Technology Use

The use of educational technology by learners is driven by individual differences (e.g., motivation) and instructional factors. Not only is the use of technology expected by today's students [16], but research also indicates that the use of technology is positively associated with student engagement and enhanced learning outcomes [5]. However, despite the expectations and educational benefits associated with technology, a common finding of several empirical studies – mainly conducted in blended learning contexts (i.e., mix of online and face-to-face) – is that a majority (over 60%) of learners can be classified as limited users of the offered educational technology. This trends holds even when the instructional design is well-informed by the available evidence regarding the benefits of the employed technology [25]. Such usage patterns of educational technology have been explained from two perspectives. First, self-regulation is not sufficiently developed in order to guide students through the process of effective use of technology [24]. Thus, better external regulation is needed to guide the students' technology-enhanced learning processes. Second, research on educational technology acceptance indicates that the extent a technology is adopted is best predicted by the instructional norm rather than other well-established technology acceptance constructs such as attitudes towards the technology and social norm [27]. This finding resonates well with the work of Huon et al. in their investigation of student use of the institutionally available resources (technologies) and academic performance. The authors noted that "students adapt their learning strategies to the resources available, with an apparent emphasis on learning what will be assessed rather than exploring for understanding" [19].

In spite of a growing interest in video annotation software, currently available research knowledge of different instructional design has received much less attention in the literature. To address this research gap, the study presented in this paper, aimed to investigate the effects of two different instructional design approaches using video annotation software on: i) patterns in creating video annotations; and ii) linguistic and psychological processes observable in annotations.

2.3 Self-regulated Learning and Annotations

Self-regulated learning is a foundational theory in modern educational psychology [2]. It recognizes that learners create knowledge, and select and manage cognitive operations in the form of study tactics they apply to learn. Dynamic feedback loops are the core of self-regulated learning [2, 15]. Based on their analyses of *external conditions* (e.g., type of assessment used in a course or time available) and *internal conditions* (e.g., prior knowledge of the domain), learners set learning goals. The goals can be modeled as a vector of standards for cognition (e.g., rate of learning), effort budgeted for learning, and features of products of cognition (e.g., recall of information).

Learners often use invalid standards in judging the degree of their own learning [2]. This leads to both *inaccurate* and *overconfident* judgments. Poor judgment of learning and overconfidence can lead to underachievement. In contrast, students with high accuracy tend towards improved retention and high academic performance. Individuals with poor judgment in learning predominantly use *improvised* and *ineffective* study tactics, such as, rereading instead of self-testing [1]. However, these issues can be addressed through the provision of regular and timely feedback [2, 15]. Feedback improves the accuracy of judgments of learning when it stimulates active cognitive engagement, e.g., involving learners in recalling previously studied information, summarizing information or generating keywords to capture a body of information.

Self-reflection has been demonstrated to be an effective approach for developing meta-cognitive monitoring. The primary intent of self-reflections are to help students internalize the expected standards of the subject matter. In this context the accuracy of judgment of (own) learning can be considered as essential for effective learning. For instance, students will (should) not stop studying if a particular skill is not mastered up to the expected standards [38]. Summary about the information studied or reflection about their own performance followed by (instructors') feedback has been empirically proven to be the most effective strategy to improve student judgment of learning [12].

Research on self-reflections produced in video annotation software are highly grounded in the self-regulated learning literature. Coding schemes for content analysis of self-reflections in video annotations directly focus on the inquiry of metacognitive monitoring skills. For example, Hulsman et al. proposed a scheme for coding self-reflections: observation, explanation (i.e., motive/effect), and planning (i.e., goals) [18]. To our knowledge, no study has attempted to identify the specific types of linguistic and psychological processes that are activated in self-reflective writing across courses with different instructional designs.

2.4 Linguistic Analysis of Self-Reflections

Previous studies incorporating video analytic techniques have focused on exploring students' viewing and usage patterns [3, 28, 32]. However, in order to obtain an understanding of students' self-regulated learning and self-reflection processes, an analysis of the textual data is required. Recent studies investigating students' use of video annotation software for note-taking of lecture videos have begun to uncover patterns in the frequency of notes [29] and the types of themes emerging in synchronous chat messages [17]. However, research on students' reflective notes when viewing a recording of their own performance or practice has predominantly been limited to manual content analysis [18], rather than leveraging the affordances of automated textual analysis. Word or linguistic analysis of written text can reveal the cognitive and psychometric processes related to different forms of thought, such as reflection on past events, and advances in language processing software allow for massive data sets of text to be analyzed quickly [35]. There are a number of text analysis software available that have been used in social science research, such as QSR NVivo, TAMS Analyzer, TextSTAT, and LIWC [36]. The latter, LIWC, was used in the present study due to its capability to calculate the number of words related to a variety of psychological and linguistic constructs [35] that can therefore provide insight into students' self-regulated learning and reflection processes.

While prior studies have used LIWC to analyze various forms of students' written assignments, such as critical thinking essays [4], self-introductions [33], and career-related narratives [23], only a few have investigated the linguistic and psychometric terms most commonly present in reflective writing [20, 30]. Although LIWC can process word counts associated with 74 linguistic and psychological variables, we incorporate only a subset of these in our textual analysis. The selected subset is based on the findings of research using LIWC to analyse reflective writing [20, 30], research on students' emotions in different learning contexts [31], and those most relevant to the context of our case study. Table 1 lists the 31 variables and their abbreviations used in this study.

First person singular and plural were selected due to research reporting that they were more frequently used in self-reflection writing than third person pronouns [30]. Similarly, tense variables and common verbs were also included in the analysis as patterns emerging in prior studies have shown that reflective writing focuses on actions through verb use and, in particular, future behaviour derived from present observations [20]. In addition, words related to cognitive processes, including insight and exclusive terms, are commonly observed terms in self-reflections [30], and hence, were also included in the study. Words with greater than six letters were also analyzed as their use indicates the use of complex language [35]. Although previous research suggests that complex language is not commonly observed in reflective prose [30], its occurrence can be interpreted as an indicator of more complex cognitive processes. Affective words were studied as certain emotions are more apparent in graded tasks compared to optional or self-regulated learning activities [31]. Lastly, we included additional variables related to perceptual processes and biological processes, as the reflective text analyzed in our study concerned the performing arts discipline and the use of such terms would be worthwhile to observe.

Table 1: LIWC variables used in the study

Main category	Sub categories
	Linguistic Processes
Words	Word count (WC), Words/sentence (WPS), Words > 6 letters (sixltr)
Personal Pronouns	1st Person singular (i), 1st person plural (we)
Verbs	Common verbs (verb)
Tense	Past tense (past), Present tense (present), Future tense (future)
	Psychological Processes
Affective (affect)	Positive emotion (posemo), Negative emotion (negemo), Anxiety (anx), Anger (anger), Sadness (sad)
Cognitive Processes (cogmech)	Insight (insight), Inclusive (incl), Exclusive (excl)
Perceptual Processes (percept)	See, hear, feel
Biological Processes (bio)	Body
Relativity (relativ)	Motion, space, time

2.5 Research Objectives and Predictions

An earlier study by Peden and Carroll reported [30] high effect sizes of the interaction of the type of assignment and the LIWC variables (.41 < partial η^2 <.77, i.e., the values of partial η^2 indicate the percentage of variance in the type of assignment and its associated error that is accounted for by the type of assignment[6]). In the present study, we predicted similar effect sizes – which requires students' use of self-reflections about their own performance with available recordings in the video annotation software (i.e., graded assignment or not [18]) – on the quantity of self-reflections and the linguistic and psychological processes evident within them. Likewise, we predicted similar high effect sizes of the course instructional design on the quantity of reflections and linguistics and psychological processes in self-reflections of students who were enrolled in two different courses in parallel, which had different considerations of self-reflections and composition of assignments related to self-reflections on video-recorded material of their own performance. However, we expected a smaller effect size for those students with parallel enrolments than for those students who were in only one of the two different courses. We anticipated that the standards used for metacognitive monitoring learned in one course could (at least slightly) influence the metacognitive monitoring (and thus products of learning – in this case, self-reflections) in the other course. We predicted more affective processes in the graded self-reflections. Especially, we predicted an increase in negative emotions, given that their established association with deep learning, usually driven by the state of cognitive disequilibrium [9]. Finally, we wanted to explore if there are any patterns in the distribution of self-reflection across different temporal positions of the used videos.

3. METHOD

3.1 Study Design

The research design was through a case study approach. This methodology was most appropriate given that the research was situated in a single setting [14]. That is a research-intensive higher education institution in North America. Secondary data collected from the log files from video annotation software, CLAS [12, 35], were used for exploration, observation, and analysis. The study also has the characteristics of a natural experiment, in which the researchers had no control over the assignment of the participants to the experimental conditions [13]. Rather, the data were collected in an ecologically valid setting where students were enrolled in courses incorporating the video annotation software. The courses under investigation differed in their instructional design.

3.2 Materials

The primary data collected and analysed for this study was extracted from the trace data logged by CLAS, a web-based and locally hosted video annotation application. Similar to the other annotation software mentioned earlier in this paper, CLAS has *two forms of annotation features*. The first, *time-stamped annotations* allow students and instructors to make time-stamped notes that correspond to a specific segment of the video that can be accessed later for review [11, 34]. This feature is simply referred to as a "*time-stamped annotation*" in the reporting of the findings. The second annotation feature is not related to a specific video time stamp, rather it allows users to write a general comment about the video. It is referred to as a "*general annotation*" in this paper.

As the CLAS application does not support downloading of the videos, students are required to view and annotate while using the software. Furthermore, CLAS is locally hosted and designed to capture various trace data, such as video access times, duration of access, frequency of access, textual data from the annotations, and time stamps [28]. For the purpose of this study, data collection and analysis focused only on the content and time stamps of the annotations made through the CLAS interface. The trace data on duration and frequency of access were not included in the study as this data was seen to be an inaccurate measure of the time students spent viewing the videos. As the data was derived from a natural experiment, the researchers could not control viewing behaviours, and therefore, students may have left the video playing or accessed it several times without actually viewing it.

3.3 Sample

With institutional ethics approval, all data pertaining to students' and instructors' annotations made through CLAS in the 2012-2013 academic year were extracted from the system log files. To protect users' anonymity and privacy, all personal identifying information was removed and replaced with random codes in order to represent annotations made by the same user across all videos in the system. Preliminary observation revealed that out of 11 courses using CLAS in the 2012-2013 academic year, students in four of the courses actively engaged with the annotation features for self-reflection purposes. For each record in the log files, we had a code of the course the annotation was made in. Time and date stamps further showed that annotations in two of these courses, Courses 1 and 2, were made in the first semester, while those in the two other courses occurred in the following semester. Furthermore, the codes representing individual students revealed that some of the students were enrolled in both Course 1 and 2 in parallel. Reviewing the annotations made by those with an instructor role helped infer the instructional design as the textual data showed signs of graded feedback in Course 2 but not in Course 1.

The case study reported in this paper focused on the textual data logged in CLAS for the two courses offered in the first semester of the 2012-2013 academic year (Courses 1 and 2). To analyze comparisons, data pertaining to students who were enrolled in both Courses 1 and 2 in parallel compared to those who were only enrolled in either the course with non-graded annotations (Course 1) or graded annotations (Course 2) were differentiated into separate samples as presented in Table 2.

3.4 Variables and Measures

Independent variable. The assignment of the study participants to the two different experimental conditions (enrollment in Course 1 and Course 2) was used as the independent variable to study the effect of both independent and paired (repeated) samples.

Dependent variables. We used two groups of dependent variables:
- *Annotation counts in each course: i)* count of all annotations created by students; *ii)* count of students' time-stamped annotations in each of the temporal quartiles (Q1-Q4) relative to the duration of the videos; and *iii)* count of general annotations.
- *LIWC variables* as introduced in Table 1. In addition, we also wanted to study the length of self-reflections in each course as a ratio of word counts per self-reflection annotation (WC/Ann).

Table 2: Study sample and their distribution across the two courses considered in this study

Course 1 (non-graded annotations)	Course 2 (graded annotations)
Group Course 1a: Students who took only Course 1, n=22	***Group Course 2a***: Students who took only Course 2, n=32
Group Course 1b and 2b[1]: Students who took both courses in parallel, n=8	

[1]Group Course 1b refers to the variables of the eight students collected in Course 1; group Course 2b refers to the variables of the same eight students collected in Course 2.

3.5 Data Analysis

All dependent variables were tested for normality using the Kolmogorov–Smirnov and Shapiro–Walk tests. This was further explored using P-P plots. All dependent variables were non-normally distributed, and hence, non-parametric tests were employed to test our predictions as outlined in Section 2.5. To test for the effect of course instructional design (i.e., independent variable), nonparametric tests for both independent (Mann-U Whitney test) and paired/repeated (Wilcoxon signed-rank test) samples were carried out as a result of the complexity of the natural experiment, on which our study was based. In particular, the Mann-U Whitney test was used to test the effect of instructional design on the dependent variables for the students who took either Course 1 or Course 2 (i.e., Course 1a vs. Course 2a). The Wilcoxon signed-rank test was used to test the effect of instructional design on the dependent variables for the students who took both Course 1 and Course 2 in parallel (i.e., Course 1b vs. Course 2b). We also checked for significant differences in the dependent variables between the students within the same course (i.e., Course 1a vs. Course 1b and Course 2a vs. Course 2b) by using the Mann-U Whitney test. Finally, to compare for differences among temporal positions of the time-stamped (quartiles Q1-4) and general annotations, we used Friedman's ANOVA followed by the Wilcoxon signed-rank test for pairwise comparison between the variables. The results of all the tests were considered significant if $p < .05$, except for Wilcoxon signed-rank test when used as a post-hoc test after Friedman's ANOVA; in this case, the Bonferroni correction (i.e., p was divided by the number of comparisons) was used.

Given the non-normal distribution of dependent variables, descriptive statistics were presented as medians and interquartile range (see Tables 3-7). The following results are reported for the performed tests: U – Mann-U Whitney test, T – Wilcoxon signed-rank test, z – standardized test, p –statistical significance, and Pearson's r – effect size. We followed Cohen's interpretation of r as effect size (i.e., .1 – small, .3 – medium, and .5 – large) and r^2 as a proportion of the variability of the dependent variable explained by the independent variable [6].

The linguistic analysis of self-reflections was performed by using the LIWC2007 software, which produced values for each annotation available in the sample in a CSV file. These variables were aggregated at the student level per course in Excel. Finally, the data was imported into SPSS (v. 19) for statistical analysis.

4. RESULTS

Table 3 outlines the descriptive statistics for the counts of self-reflection annotations across the two courses, including the statistics for the students who took only one course (Courses 1a and 2a) and who took both courses (Courses 1b and 2b). The statistical tests (Column A/B) confirmed our predictions that the students from the course with graded self-reflections (Course 2) had significantly higher counts of self-reflections (annotation total). This difference holds for both the students who took only one of the two courses (Course 1a vs. Course 2a) and for those who took both courses in parallel (Course 1b vs. Course 2b). Consistent with our prediction, the effect size for students with the parallel enrollment (Course 1b vs. Course 2b, r=.63) was smaller than with the students with only one course (Course 1a vs. Course 2a, r = .81). Both effect sizes are considered as large according to Cohen's interpretation of r. This finding results from the students who took both courses had a higher count of self-reflections (Course 1b) than their peers (Course 1a) who only took Course 1. This is in spite of the finding revealed by additional Mann-U Whitney tests (not reported in Table 3) that there were no significant differences between the students within the same course (i.e., Course 1a vs. 1b and Course 2a vs. 2b).

The mean value of effect sizes for the students who took both the courses (Course 1b vs. Course 2b, r=.58) and took only one course (Course 1a vs. Course 2a, r=.62) is similar in the counts of self-reflections associated with time-stamped (Q1-Q4) and general annotations of individual videos. The largest concentration of the time-stamped annotations was in the first quartile in Course 2 (i.e., graded self-reflections), unlike Course 1 (non-graded self-reflections) where the distribution was balanced across the temporal quartiles of videos and generally.

Table 3: Counts of self-reflection annotations across the two courses

	Course 1 (n=30)	Course 1a (n=22)	Course 1b (n=8)	Course 2 (n=40)	Course 2a (n=32)	Course 2b (n=8)	A / B
Annotation total	11.00 (8.00, 17.25)	10.50 (7.50, 8.75)	14.00 (8.25, 16.50)	55.00 (48.00, 60.00)	55.00 (47.25, 59.75)	58.00 (53.00, 62.25)	A: U=689.50, z=5.95, p<.001, r=.81 B: T=36.00, z=2.52 p=.012, r=.63
Annotation position Q1	2.00 (1.00, 3.00)	2.00 (1.00, 3.00)	3.00 (2.00, 3.00)	21.50 (17.00, 30.00)	21.50 (17.00, 30.00)	21.00 (16.00, 26.25)	A: U=703.00, z=6.20, p<.001, r=.84 B: T=36.00, z=2.52 p=.012, r=.63
Annotation position Q2	2.50 (1.00, 5.00)	2.00 (1.00, 5.00)	4.00 (2.00, 5.00)	7.50 (4.25, 10.00)	7.00 (4.25, 9.00)	10.00 (4.00, 10.75)	A: U=587.00, z=4.16, p<.001, r=.57 B: T=36.00, z= 2.53, p=.011, r=.63
Annotation position Q3	2.00 (1.00, 4.00)	2.00 (1.00, 4.00)	2.00 (1.25, 2.75)	6.00 (3.00, 10.75)	6.00 (3.00, 9.00)	8.50 (1.00, 12.75)	A: U=539.00, z=3.31, p=.001, r=.45 B: T=26.50, z= 2.11, p=.034, r=.53
Annotation position Q4	2.50 (1.00, 4.25)	3.00 (1.75, 4.00)	2.00 (0.25, 5.00)	6.50 (3.25, 8.00)	6.50 (3.00, 8.00)	6.00 (4.00, 8.75)	A: U=535.00, z=3.24, p=.001, r =.44 B: T=27.00, z= 2.21, p=.027, r=.55
Annotation general	2.00 (0.00, 3.00)	2.00 (0.00, 3.00)	2.50 (0.25, 3.00)	8.00 (8.00, 14.75)	8.00 (8.00, 14.00)	11.50 (4.25, 15.75)	A: U=686.00, z=5.96, p<.001, r=.81 B: T=34.50, z= 2.31, p=.021, r=.58

Legend: All variables given as median (25th percentile, 75th percentile). A – Course 1a vs. Course 2a (Mann-U Whitney test); B – Course 1b vs. Course 2b (Wilcoxon signed-rank test)

Given that there were no observed significant differences between the students inside either course (Course 1a vs. 1b and Course 2a vs. 2b), Friedman's ANOVA tests were carried out on the entire sample of both courses to test for differences between counts of self-reflections across temporal quartiles and general self-reflections about the videos. In Course 1, there were no significant differences in the count of time-stamped annotations between the four quartiles of the videos and general annotations ($\chi^2(4)$=8.51, p=.075). However, in Course, 2, a significant difference was observed ($\chi^2(4)$=79.22, p<.001). Post-hoc Wilcoxon signed-rank pairwise tests (10 of them with p < .005) revealed significant differences between self-reflections in: Q1 vs. Q2 (T=2.45, z=6.93, p<.001, r=.78), Q1 vs. Q3 (T=2.35, z=6.65, p<.001, r=.74), Q1 vs. Q4 (T=2.71, z=7.67, p<.001, r=.86), and Q1 and general self-reflections (T=1.80, z=5.09, p<.001, r=.57). That is, the students had significantly more self-reflection annotations in the first quartile of their videos and the effect size of this difference was large according to Cohen's interpretation of r.

Table 4 outlines the descriptive statistics of the variables indicative of the LIWC linguistic processes in the self-reflections. The results of the tests confirmed significant differences in (almost all) of the linguistic processes activated by the different instructional designs of the two courses. Moreover, our prediction was also confirmed that the effect size for the students who took both courses in parallel (Course 1b vs. 2b) was lower than for those who took one course only (Course 1a vs. 2a) except for the variables WC/Ann (word count per annotation) and words related to the pronoun, "we". However, while there was no observed difference in the word count per self-reflection (WC/Ann) between students who took only one of the courses (Course 1a vs. 2a), the median value was considerably higher for Course 2a (i.e., graded self-reflections) than for Course 1a. As expected, the students who took only Course 1(a) used significantly more "we" words than those who took Course 2(b). This finding is due to Course 1 videos being used for students' group performances, while Course 2 used videos of solo performances as objects for self-reflections. There was also no significant difference in the use of "we" words in students who took both courses in parallel (Courses 1b and 2b), in spite of the median value for Course 2b (Mdn=0.68) being considerably lower than for Course 1b (Mdn=11.24).

Table 4: Linguistic processes across the two courses

	Course 1 (n=30)	**Course 1a (n=22)**	**Course 1b (n=8)**	**Course 2 (n=40)**	**Course 2a (n=32)**	**Course 2b (n=8)**	**A** **B**
WC	266.00 (158.25, 479.25)	304.00 (108.00, 550.25)	241.00 (178.75, 280.50)	1613.50 (1282.75, 2155.25)	1671.50 (1218.50, 2212.25)	1593.00 (1502.00, 1921.50)	U = 697.00, z=6.07, p<.001, r=.83 T=36.00, z=2.52, p<.012, r=.63
WPS	225.00 (120.13, 423.29)	272.75 (99.75, 547.71)	209.36 (134.63, 275.63)	1209.76 (987.03, 1641.23)	1179.25 (936.29, 1641.25)	1306.54 (1090.02, 1661.01)	U =674.00, z=5.67, p<0.001, r=.77 T=36.00, z=2.52, p<.012, r=.63
WC/ Ann	19.60 (14.37, 36.03)	22.56 (13.17, 48.39)	16.66 (14.74, 20.17)	30.09 (25.39, 37.95)	30.94 (25.42, 39.79)	29.30 (24.50, 33.13)	U = 446.00, z=1.67, p=.098, r=.23 T=36.00, z=2.52, p<.012, r=.63
Sixltr	173.82 (119.43, 269.81)	173.82 (119.43, 285.31)	190.37 (97.73, 271.33)	888.00 (740.56, 1036.67)	892.25 (691.57, 1044.11)	878.75 (814.26, 972.09)	U = 683.00, z= 5.83, p<.001, r=.79 T=36.00, z=2.52, p<.012, r=.63
i	13.94 (6.27, 31.03)	13.94 (6.27, 23.31)	32.52 (2.27, 64.76)	419.10 (175.00, 531.36)	396.89 (175.00, 524.67)	513.33 (162.58, 578.06)	U = 704.00, z= 6.20, p<.001, r=.84 T=28.00, z=2.37, p=.018, r=.59
we	10.20 (0.00, 20.91)	8.01 (0.00, 21.98)	11.24 (0.00, 18.46)	00.00 (0.00, 2.67)	0.00 (0.00, 2.67)	0.68 (0.00, 2.65)	U = 158.00, z=-3.11, p=.002, r=-.42 T=3.00, z= -1.86, p<.063, r=-.47
verb	141.77 (88.28, 209.99)	130.01 (86.99, 209.90)	177.89 (98.35, 237.07)	818.45 (711.28, 985.74)	782.90 (711.28, 982.04)	915.86 (721.823, 985.74)	U = 693.00, z=6.00, p<.001, r=.82 T=36.00, z=2.52, p<.012, r=.63
past	15.24 (6.11, 24.77)	17.19 (6.55, 24.91)	11.42 (3.53, 22.50)	178.17 (128.58, 237.87)	191.69 (128.58, 237.87)	157.64 (60.93, 236.33)	U = 685.00, z= 5.86, p<.001, r=.80 T=36.00, z=2.52, p<.012, r=.63
present	79.87 (43.72, 136.35)	75.13 (35.75, 100.71)	122.46 (71.02, 160.42)	434.86 (351.50, 528.77)	407.05 (337.50, 502.55)	490.30 (396. 00, 577.45)	U = 688.00, z=5.93, p<.001, r=.81 T=36.00, z=2.52, p<.012, r=.63
future	0.00 (0.00, 8.30)	0.00 (0.00, 7.44)	0.00 (0.00, 17.88)	44.55 (27.47, 64.19)	44.55 (29.42, 64.00)	38.75 (13.65, 70.53)	U = 684.00, z=5.88, p<.001, r=.80 T=36.00, z=2.52, p<.012, r=.63

Legend: All variables given as median (25th percentile, 75th percentile). A – Course 1a vs. Course 2a (Mann-U Whitney test); B – Course 1b vs. Course 2b (Wilcoxon signed-rank test)

Table 5 presents the descriptive statistics and results of the tests for the LIWC variables related to cognitive and perceptual processes (see Table 1) across the two courses. As predicted, the statistical differences for both cognitive and perceptual processes were significant with large effect sizes. In contrast, the effect size of the cognitive and perceptual processes of the students who took both courses in parallel (Course 1b vs. 2b) was lower than for the students who took one course only (Course 1a vs. 2a). Although in both courses, the use of words indicative of cognitive processes (i.e. cogmech, insight, incl, and excl; see Table 1) accounts for over 50% of the overall word count (c.f. WC in Table 4), it is noteworthy that the course with graded self-reflections (Course 2) maintained this proportion (i.e. the self-reflections in Course 2 had 5.5 times more words about cognitive processes than in Course 1). As well, the perceptual words follow a similar growing trend in Course 2. However, in Course 1, the ratio of words related to perceptual process (precept) is lower (19%) than in Course 2 (12%) relative to the total word count (c.f. WC in Table 4).

Table 6 reports the descriptive statistics and the results of the tests for the LIWC variables indicative of biological and relative processes (see Table 1). The predicted differences and effect sizes were also found for these two groups of variables, similar to the previous two. In the biological processes group, two particularly important patterns emerged. First, the increase in the counts of words indicative of biological processes is considerably higher than in any previous case between the two courses. Second, the only significant difference observed between the students who took Course 1 only (Course 1a) and who took both courses (Course 1b) was for the *body* variable (U = 94.50, z=.31, p=.014, r=-.06). Although the effect size was small (r=-.06), it seems that the students in Course 1b paid more attention to their bodies when reflecting on their group performance in Course 1.

Table 5: Cognitive and perceptual processes

	Course 1 (n=30)	Course 1a (n=22)	Course 1b (n=8)	Course 2 (n=40)	Course 2a (n=32)	Course 2b (n=8)	A	B
cogmech	155.48 (112.87, 226.82)	143.33 (100.28, 229.22)	172.92 (137.16, 226.18)	860.41 (690.78, 1002.34)	851.36 (701.82, 994.17)	874.04 (492.52, 1064.65)	U=686.00, z=5.88, p<.001, r=.80	T=36.00, z=2.52, p<.012, r=.63
insight	13.62 (5.01, 29.29)	11.25 (3.76, 21.53)	26.75 (15.42, 33.99)	115.60 (79.21, 167.53)	114.39 (76.84, 166.61)	135.56 (85.56, 184.78)	U=669.00, z=5.58, p<.001, r=.76	T=35.00, z=2.38, p=.017, r=.60
incl	39.95 (39.95, 58.22)	36.30 (20.38, 61.24)	43.01 (25.71, 55.18)	177.75 (141.14, 235.44)	182.05 (143.70, 240.41)	155.47 (83.79, 223.10)	U=684.00, z=5.85, p<.001, r=.80	T=34.00, z=2.24, p<.025, r=.56
excl	27.50 (13.22, 38.33)	27.50 (9.78, 41.89)	28.48 (15.34, 35.91)	119.13 (78.66, 175.94)	119.13 (81.09, 167.34)	123.43 (39.21, 191.61)	U=680.00, z=5.60, p<.001, r=.76	T=36.00, z=2.52, p<.012, r=.63
percept	51.44 (21.43, 88.44)	49.75 (18.91, 02.35)	59.43 (30.63, 81.37)	199.28 (164.22, 265.36)	192.86 (161.11, 258.42)	237.94 (172.14, 280.97)	U=674.00, z=5.67, p<.001, r=.77	T=36.00, z=2.52, p<.012, r=.63
see	7.51 (0.00, 14.71)	6.97 (0.00, 17.08)	7.51 (1.44, 12.65)	71.58 (38.42, 100.58)	67.04 (34.19, 100.25)	84.63 (53.68, 105.67)	U=664.00, z=5.50, p<.001, r=.75	T=36.00, z=2.52, p<.012, r=.63
hear	32.44 (8.40, 52.98)	30.93 (5.70, 57.00)	43.52 (12.11, 52.91)	36.21 (28.25, 67.63)	36.21 (28.17, 67.63)	35.48 (28.71, 77.49)	U=422.00, z=1.23, p=.22, r=.17	T=23.00, z=.70, p=.484, r=.18
feel	7.90 (0.00, 16.30)	10.33 (0.00, 16.30)	5.63 (0.58, 27.92)	81.00 (57.2, 109.33)	73.16 (53.40, 109.33)	95.63 (68.57, 108.02)	U = 698.00, z=6.10, p<.001, r=.83	T=28.00, z=2.37, p=.018, r=.59

Legend: All variables given as median (25^{th} percentile, 75^{th} percentile). A – Course 1a vs. Course 2a (Mann-U Whitney test); B – Course 1b vs. Course 2b (Wilcoxon signed-rank test)

Table 7 reports the descriptive statistics and results of the tests of the LIWC variables indicative of affective processes (see Table 1). The results of these tests confirmed our prediction about the overall significant increase of affective processes in the course with the graded self-reflections. However, their effect sizes were only slightly above the threshold (r≥.5) for interpretation as a large effect size according to Cohen [6]. In this case, the effect size for students who took only Course 2 (Course 1a vs. Course 2a) was almost the same as for those who took both courses (Course 1b vs. Course 2b). It is also indicative that the increase in the count of negative emotions (negemo; see Table 1) was much higher than positive ones (posemo; see Table 1) in Course 2; yet, the proportion of negative emotions with respect to all other LIWC categories remained rather low in Course 2. However, the analyses did not reveal significant differences in positive emotion, anger, and sadness for the students who took both courses in parallel (Course 1b vs. Course 2b).

5. DISCUSSION

5.1 Interpretation of the Results

As predicted, the count of self-reflection annotations was considerably higher for Course 2, presumably as direct result of the graded self-reflections in this course. The increase in the median value for Course 2 was precisely five times (see Table 3) that of Course 1 where student annotations were not assessed. The instructional design of graded self-reflections contributed towards the increased numbers of annotations. This is well evidenced in the observed proportion of variance (r^2) shared by the independent variable (i.e., course design) and the dependent variable (annotations total), 65.61% and 39.69% – N.B., r^2 calculated based on the values of r reported in Table 3 – depending if the students took only one of the two courses or both, respectively. Future research should further extend this initial work and consider the potential impact of individual differences, such as metacognitive awareness, motivation, and prior knowledge, on the timing, quantity, and quality of students' self-reflections.

The present study identified a clear trend in the declining count of time-stamped annotations over the duration of a video. Quartile 1 (i.e., first 25% of video content) was consistently where the highest count of self-reflections were made. A similar trend was also reported by Mu (2010) in their usability study of video annotations [29]. This observed trend may be explained by i) the students observed a (set of) common problem(s) in their performance that reoccurred later, ii) adaptation to the instructional requirements by making a strategic decision to stop creating self-reflections judging that the course requirements were satisfied [19], which can often be a result of inaccurate judgment of learning [2], and/or iii) students become cognitively overloaded by creating self-reflections, which leads to their overall decline as their video watching progresses. Increased metacognitive monitoring – induced by graded self-reflections – requires additional memory resources and that might be exhausted, especially for the students who have lower prior knowledge [37]. All these three explanations warrant further research and validation.

Table 6: Biological and relative processes across the courses

	Course 1 (n=30)	Course 1a (n=22)	Course 1b (n=8)	Course 2 (n=40)	Course 2a (n=32)	Course 2b (n=8)	A / B
bio	6.79 (0, 15.19)	5.85 (0.00, 15.29)	6.79 (1.56, 15.38)	161.11 (116.21, 254.60)	157.34 (110.04, 244.97)	202.08 (125.92, 270.71)	U=704.00, z=6.21, p<.001, r=.85 / T=36.00, z=2.52, p<.012, r=.63
body	0.00 (0.00, 2.57)	0.00 (0.00, 0.00)	5.21 (0.00, 11.37)	143.87 (102.65, 220.79)	140.70 (102.65, 201.67)	182.64 (78.80, 268.21)	U=704.00, z= 6.31, p<.001, r=.86 / T=36.00, z=2.52, p<.012, r=.63
relativ	123.19 (78.73, 208.38)	109.12 (65.84, 203.09)	165.54 (104.57, 242.40)	726.86 (617.48, 898.98)	728.62 (607.46, 925.56)	704.61 (646.82, 840.00)	U=695.00, z=6.04, p<.011, r=.82 / T=36.00, z=2.52, p<.012, r=.63
motion	17.32 (9.56, 26.79)	17.61 (10.83, 25.66)	16.21 (7.07, 27.68)	103.96 (77.32, 128.62)	112.07 (77.16, 133.29)	86.89 (77.32, 105.04)	U = 691.00, z= 5.97, p<.011, r=.81 / T=28.00, z= 2.37, p=.018, r=.59
space	66.36 (38.35, 118.86)	53.12 (30.07, 118.86)	81.78 (62.39, 129.87)	376.38 (314.98, 481.37)	370.84 (298.18, 484.90)	390.84 (365.25, 481.37)	U = 693.00, z=6.00, p<.001, r=.82 / T=36.00, z=2.52, p<.012, r=.63
time	50.04 (23.03, 86.60)	46.79 (20.19, 86.60)	63.92 (28.10, 96.41)	252.20 (195.87, 313.65)	268.29 (195.87, 313.65)	221.92 (190.42, 350.32)	U = 687.00, z= 5.90, p<.001, r=.80 / T=36.00, z=2.52, p<.012, r=.63

Legend: All variables given as median (25th percentile, 75th percentile). A – Course 1a vs. Course 2a (Mann-U Whitney test); B – Course 1b vs. Course 2b (Wilcoxon signed-rank test)

Table 7: Affective processes across the two courses

	Course 1 (n=30)	Course 1a (n=22)	Course 1b (n=8)	Course 2 (n=40)	Course 2a (n=32)	Course 2b (n=8)	A / B
affect	108.41 (58.55, 150.49)	104.34 (60.75, 148.96)	117.57 (45.15, 157.96)	239.70 (202.86, 301.15)	235.40 (202.86, 288.55)	243.62 (192.66, 332.67)	U=579.00, z=4.00, p<.001, r=.54 / T=33.00, z=2.10, p=.036, r=.53
posemo	92.98 (50.39, 125.26)	93.98 (50.39, 119.55)	92.09 (45.15, 145.93)	159.88 (123.93, 209.72)	167.79 (123.55, 215.74)	150.70 (126.50, 191.90)	U=597.00, z=5.46, p<.001, r=.74 / T=28.00, z=1.40, p=.161, r=.35
negemo	8.79 (0.00, 20.67)	8.91 (0.00, 19.87)	6.81 (0.93, 21.68)	69.51 (39.71, 98.87)	69.51 (36.78, 94.32)	72.10 (47.32, 124.24)	U=662.00, z=5.43, p<.001, r=.74 / T=35.00, z=2.38, p=.017, r=.60
anx	0.00 (0.00, 1.95)	0.00 (0.00, 0.34)	0.00 (0.00, 7.57)	27.39 (12.48, 45.96)	25.15 (12.48, 44.22)	31.13 (10.98, 51.62)	U=654.00, z=4.48, p<.001, r=.61 / T=27.00, z=2.20, p=.028, r=.55
anger	0.00 (0.00, 0.61)	0.00 (0.00, 0.00)	0.00 (0.00, 9.38)	7.25 (0.78, 15.54)	7.83 (1.17, 15.54)	4.56 (0.00, 27.81)	U=595.00, z=2.29, p=.02, r=.31 / T=19.00, z=.85, p=.398, r=.21
sad	0.00 (0.00, 8.33)	0.00 (0.00, 8.43)	0.00 (0.00,4.41)	7.25 (0.27, 12.77)	7.92 (1.32, 12.78)	4.02 (0.00, 15.94)	U=595.00, z=4.86, p<.001, r=.66 / T=12.00, z=1.21, p=.225, r=.30

Legend: All variables given as median (25th percentile, 75th percentile). A – Course 1a vs. Course 2a (Mann-U Whitney test); B – Course 1b vs. Course 2b (Wilcoxon signed-rank test)

As predicted, the effect size of the graded self-reflections on the students who took both courses in parallel was lower than for those who took only one course. Students, who took both courses in parallel, had higher values of most of variables in Course 1 (i.e., Course 1b) than their peers who were only enrolled in Course 1 (i.e., Course 1a). This finding may be attributed to the students appreciating the value of self-reflections and internalizing how to metacognitively monitor themselves. Consequently this may impact upon their further use of, and learning with, the video annotation software. An alternate explanation is that the students simply "mirrored" their expected learning in the course with graded self-reflections. In the future we plan to study the effects on self-reflections when the grading persists in the follow-up courses vs. when it is not an assessable component.

Indicators of the linguistic and psychological processes were also higher in the course with the graded self-reflections. This is characterized in the use of more complex language (e.g., WPS – longer sentences and Sixltr – longer words). Complexity of language adoption is also an indicator of more complex cognitive processes and critical thinking. These processes are more readily associated with essays than individual self-reflections [4]. The features of self-reflections – characterized by the increased use of "i" and references to cognitive processes as a way to express metacognitive monitoring – were preserved. That is, the grading did not influence the students' focus from metacognitive monitoring of their own performances to the attempts to make their self-reflections look good for grading purposes solely. Namely, the study reported in [30] showed differences in the LIWC variables between essays and self-reflection assignments. Determining an optimal mix of the metacognitive monitoring and presentation of their own performance seems an important question for future research on the effects of graded self-reflections in video annotation software.

It was expected that the higher level of metacognitive monitoring would result in a higher level of the use of words referring to biological and relative processes, given the nature of the courses considered in our study (i.e. performing arts). This prediction was validated in the graded self-reflections for Course 2. It was also observed that there was an increase in the number of references to the body processes in Course 1 for those students who took Course 2 in parallel. Further exploration of the data is required to determine if this result was due to i) the performance standards and self-monitoring learned in Course 2, ii) the need to pay attention to more details requested by the graded self-reflections; or most likely iii) a combination of both i) and ii). Again this would be an important research question to study in the future.

As expected, a higher number of references to affective processes were present in the graded self-reflections (Course 2). This finding was anticipated for negative emotions, as these types of emotions are often associated with a state of cognitive disequilibrium, which has been shown to drive deep learning [9]. Indicators of deep learning engagement are difficult to identify if students are in a state of constant cognitive equilibrium. It is the transition from the state of disequilibrium to the state of equilibrium that leads to observable indicators of deep learning. This is characterized by a shift from negative to positive emotions [31]. In the graded self-reflections course, it is clear that the positive emotions are dominant, with a "necessary" dose of negative emotions, probably activated with deeper cognitive engagement. Future research can focus on (automated) identification of situations when negative emotions are dominant in self-reflections of individuals, as they are precursors of boredom (if frustration is not solved), that typically leads to learning disengagement [9].

5.2 Implications for Practice

As noted earlier, video annotation technologies provide a mechanism for students to make reflective and self-monitoring notes linked to key segments of a video they are viewing as a method of developing their metacognitive skills [11]. In particular, when viewing recordings of themselves, students can describe their behaviour in the video and make note of areas of improvement or future goals, and indication of a full reflection cycle [18]. These notes or annotations provide copious amounts of textual data that can be time-consuming for instructors to review and analyze to ensure their students are critically reflecting and enhancing their meta-cognitive skills. As demonstrated in the present study, quantitative text analysis software, such as LIWC, can quickly analyze a large amount of written text and provide word counts related to linguistic and psychometric processes [35], thereby reducing the time required for instructors to read and analyze each reflective annotation. The provision of these forms of textual analytics data to students can also help them become aware of their own written reflective language and assist in scaffolding the development of their metacognitive skills. However, in order for the word counts to have any significant meaning for instructors and students, it is necessary to identify the terms and their frequency that would be most indicative of the self-reflective and metacognitive process, such as those identified in the present study.

The study also demonstrated that the analytics related to the temporal positions of time-stamped annotations has the potential to inform instructors of patterns in students' annotation positioning. These findings can be translated to inform and improve the design and intervention of learning tasks and feedback.

5.3 Implications for Research

Translation of the self-reflections into grades is an important research challenge. While in this study we had access to the self-reflection grades available from within the annotations, the instructors may have developed their own grading schema. This makes the research on associations between grades and the variables used in this study deeply contextualized. As a result, it is difficult to generalize and replicate these findings in future studies. In our on-going research, we have attempted to incorporate existing grading schemes for analyzing reflections such as those identified in the work of Hulsman et al. and Kember et al. [18, 21]. However, these schemes were inapplicable in this instance due to the "micro" nature of the self-reflections created in the video annotation software. Therefore, future research should aim to develop an instrument that will guide and inform the process of instructional intervention (feedback), assessment, and understanding of self-reflections.

6. ACKNOWLEDGMENTS

This research is in part supported by Australian Office of Learning and Teaching (Innovation and Development Grant), Social Sciences and Humanities Research Council of Canada (Insight Grant) and Natural Sciences and Engineering Research Council of Canada (Discovery Grant). We also thank Thomas Dang for data extraction and preparation and Danijela Gasevic for her advice about the study design.

7. REFERENCES

[1] Bargeron, D., Gupta, A., Grudin, J. and Sanocki, E. 1999. Annotations for streaming video on the Web: system design and usage studies. *Computer Networks*. 31, 11 (1999), 1139–1153.

[2] Bjork, R.A., Dunlosky, J. and Kornell, N. 2013. Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*. 64, 1 (2013), 417–444.

[3] Brooks C, Epp, C D, Logan G and Greer J 2011. The Who, What, When, and Why of Lecture Capture. *Proceedings of the First International Conference on Learning Analytics and Knowledge* (2011), 86–92.

[4] Carroll, D.W. 2007. Patterns of student writing in a critical

thinking course: A quantitative analysis. *Assessing Writing*. 12, 3 (Jan. 2007), 213–227.

[5] Chen, P.-S.D., Lambert, A.D. and Guidry, K.R. 2010. Engaging online learners: The impact of Web-based learning technology on college student engagement. *Computers & Education*. 54, 4 (May 2010), 1222–1232.

[6] Cohen, J. 1992. A power primer. *Psychological bulletin*. 112, 1 (Jul. 1992), 155–159.

[7] Colasante, M. 2011. Using video annotation to reflect on and evaluate physical education pre-service teaching practice. *Australasian Journal of Educational Technology*. 27, 1 (2011), 66–88.

[8] Colasante, M. and Fenn, J. 2009. "mat": A New Media Annotation Tool with an Interactive Learning Cycle for Application in Tertiary Education. *World Conference on Educational Multimedia, Hypermedia and Telecommunications* (2009), 3546–3551.

[9] D'Mello, S. and Graesser, A. 2012. Dynamics of affective states during complex learning. *Learning and Instruction*. 22, 2 (Apr. 2012), 145–157.

[10] Daniel, R. 2001. Self-assessment in performance. *British Journal of Music Education*. 18, 03 (Oct. 2001).

[11] Dawson, S., Macfadyen, L., Evan, F.R., Foulsham, T. and Kingstone, A. 2012. Using technology to encourage self-directed learning: The Collaborative Lecture Annotation System (CLAS). *ascilite Conference* (2012).

[12] Dunlosky, J. and Lipko, A.R. 2007. Metacomprehension A Brief History and How to Improve Its Accuracy. *Current Directions in Psychological Science*. 16, 4 (2007), 228–232.

[13] Dunning, T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press.

[14] Eisenhardt, K.M. 1989. Building theories from case study research. *Academy of management review*. 14, 4 (1989), 532–550.

[15] Hattie, J. and Timperley, H. 2007. The Power of Feedback. *Review of Educational Research*. 77, 1 (Mar. 2007), 81–112.

[16] Hawkins, B.L. and Rudy, J.A. 2008. *EDUCAUSE core data service fiscal year 2007 summary report,*.

[17] He, W. 2013. Examining students' online interaction in a live video streaming environment using data mining and text mining. *Computers in Human Behavior*. 29, 1 (Jan. 2013), 90–102.

[18] Hulsman, R.L., Harmsen, A.B. and Fabriek, M. 2009. Reflective teaching of medical communication skills with DiViDU: Assessing the level of student reflection on recorded consultations with simulated patients. *Patient Education and Counseling*. 74, 2 (Feb. 2009), 142–149.

[19] Huon, G., Spehar, B., Adam, P. and Rifkin, W. 2007. Resource use and academic performance among first year psychology students. *Higher Education*. 53, 1 (Jan. 2007), 1–27.

[20] Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V. and Whittaker, S. 2013. Echoes from the past: how technology mediated reflection improves well-being. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), 1071–1080.

[21] Kember, D., McKay, J., Sinclair, K. and Wong, F.K.Y. 2008. A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education*. 33, 4 (Aug. 2008), 369–379.

[22] Leijen, Ä., Lam, I., Wildschut, L., Robert-Jan Simons, P. and Admiraal, W. 2009. Streaming video to enhance students' reflection in dance education. *Computers & Education*. 52, 1 (Jan. 2009), 169–176.

[23] Lengelle, R., Meijers, F., Poell, R. and Post, M. 2013. The effects of creative, expressive, and reflective writing on career learning: An explorative study. *Journal of Vocational Behavior*. 83, 3 (Dec. 2013), 419–427.

[24] Lust, G., Elen, J. and Clarebout, G. 2013. Regulation of tool-use within a blended course: Student differences and performance effects. *Computers & Education*. 60, 1 (Jan. 2013), 385–395.

[25] Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J. and Clarebout, G. 2011. Tool-use in a blended undergraduate course: In Search of user profiles. *Computers & Education*. 57, 3 (Nov. 2011), 2135–2144.

[26] Magenheim, J., Reinhardt, W., Roth, A., Moi, M. and Engbring, D. 2010. Integration of a Video Annotation Tool into a Coactive Learning and Working Environment. *Key Competencies in the Knowledge Society*. Springer. 257–268.

[27] McGill, T.J. and Klobas, J.E. 2009. A task–technology fit view of learning management system impact. *Computers & Education*. 52, 2 (Feb. 2009), 496–508.

[28] Mirriahi, N. and Dawson, S. 2013. The pairing of lecture recording data with assessment scores: a method of discovering pedagogical impact. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (2013), 180–184.

[29] Mu, X. 2010. Towards effective video annotation: An approach to automatically link notes with video content. *Computers & Education*. 55, 4 (Dec. 2010), 1752–1763.

[30] Peden, B.F. and Carroll, D.W. 2008. Ways of Writing: Linguistic Analysis of Self-Assessment and Traditional Assignments. *Teaching of Psychology*. 35, 4 (Oct. 2008), 313–318.

[31] Pekrun, R., Goetz, T., Titz, W. and Perry, R.P. 2002. Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist*. 37, 2 (2002), 91–105.

[32] Phillips, R., Maor, D., Cumming-Potvin, W., Roberts, P., Herrington, J., Preston, G., Moore, E. and Perry, L. 2011. Learning analytics and study behaviour: A pilot study. *In Williams, P. Statham, N. Brown & B. Cleland (Eds.), Changing Demands, Changing Directions. Proceedings ascilite Hobart 2011* (2011), 997–1007.

[33] R. L. Robinson, Navea, R. and Ickes, W. 2013. Predicting Final Course Performance From Students' Written Self-Introductions: A LIWC Analysis. *Journal of Language and Social Psychology*. (Feb. 2013).

[34] Risko, E.F., Foulsham, T., Dawson, S. and Kingstone, A. 2013. The Collaborative Lecture Annotation System (CLAS): A New TOOL for Distributed Learning. *IEEE Transactions on Learning Technologies*. 6, 1 (2013), 4–13.

[35] Tausczik, Y.R. and Pennebaker, J.W. 2009. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*. 29, 1 (Dec. 2009), 24–54.

[36] Taylor, J. 2012. The assessment of critical evaluation, leadership and reflection skills through participation in online discussions. *Psychology Teaching Review*. 18, 2 (2012), 52–58.

[37] Valcke, M. 2002. Cognitive load: updating the theory? *Learning and Instruction*. 12, 1 (Feb. 2002), 147–154.

[38] Winne, P.H. and Hadwin, A.F. 1998. Studying as self-regulated learning. *Metacognition in educational theory and practice*. D.J. Hacker, J. Dunlosky, and A.C. Graesser, eds. Lawrence Erlbaum Associates Publishers. 277–304.