

Improving Proctoring by Using Non-Verbal Cues During Remotely Administrated Exams

by

Chia-Yuan Chuang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2015 by the
Graduate Supervisory Committee:

John Femiani, Co-Chair
Scotty Craig, Co-Chair
Jennifer Bekki

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

This study investigated the ability to relate a test taker's non-verbal cues during online assessments to probable cheating incidents. Specifically, this study focused on the role of time delay, head pose and affective state for detection of cheating incidences in a lab-based online testing session. The analysis of a test taker's non-verbal cues indicated that time delay, the variation of a student's head pose relative to the computer screen and confusion had significantly statistical relation to cheating behaviors. Additionally, time delay, head pose relative to the computer screen, confusion, and the interaction term of confusion and time delay were predictors in a support vector machine of cheating prediction with an average accuracy of 70.7%. The current algorithm could automatically flag suspicious student behavior for proctors in large scale online courses during remotely administered exams.

TABLE OF CONTENTS

	Page
LIST OF TABLES	v
LIST OF FIGURES	vi
CHAPTER	
1 INTRODUCTION.....	1
2 BACKGROUND	4
Evidence that Online Cheating is Prevalent	4
Online Proctoring Tools and Services	6
Possible Non-verbal Cues Related to Academic Dishonesty	7
Time delay in testing.....	8
Affective states in learning.....	8
Affective states in testing.....	9
Visual focus of attention (VFOA) in testing.....	10
Research Questions and Hypotheses	11
3 METHODS.....	13
Participants.....	13
Materials.....	13
Learning materials.....	13
Testing materials.....	14
Cheating materials.....	15
Interview materials.....	15
Procedure.....	16
Discussion	18
4 THE ROLE OF STUDENT'S CERTAINTY AND TIME DELAY	20
Response Times in Testing	20
Impasses in Testing	21
Current Study.....	21

CHAPTER	Page
Results	22
Discussion	23
5 DETECTING CHEATING BASED ON TIME DELAY AND HEAD POSE	25
Current Study.....	25
Data Treatment.....	25
Features in VFOA.....	25
Fast Correlation-Based Filter (FCBF).....	27
Results	28
Feature selection of VFOA.....	28
VFOA combined with time delay.	31
Discussion	32
6 ESTIMATING CHEATING INCIDENTS BASED ON NON-VERABAL CUES	35
Current Study.....	35
Data Treatment.....	35
Results	37
Discussion	40
7 GENERAL DISCUSSION	43
Contributions.....	46
Limitations and Future Research	48
Conclusion.....	50
REFERENCES.....	52
APPENDIX	
A EXEMPTION FROM IRB REVIEW	56
B LERANING MATERIALS.....	58
C TESTING MATERIALS	62
D INTERVIEW MATERIALS	68
E INSTRUCTIONS OF THE EXPERIMENT PROCESS.....	75

F	CONSENT INFORMATION	78
---	---------------------------	----

LIST OF TABLES

Table	Page
1. Models of predicting cheating intentions based on personal/situational factors	7
2. Results of the third model of hierarchical logistic regression with predictors subjects, experiences, time delay, and students' certainty ratings.....	23
3. Mean and S.D. for time delay and certainty rating.....	23
4. Eight statistical features of VFOA and their definitions	27
5. Results of Fast Correlation-Based Filter (FCBF) for head pose features	29
6. Results of logistic regression for features of VFOA	31
7. Results for the full hierarchical logistic regression with VFOA and Time Delay	31
8. Results for the final model in logistic regression with VFOA and Time Delay.....	32
9. The results of Logistic Regression Model and SVM	32
10. Results for the initial model by using the amount of confusion as the predictor.....	37
11. Results for the second model by adding time delay and the variation of head pose relative to the screen.....	38
12. Results for the third model by adding the interaction term of confusion and time delay....	38
13. The classification results of the final model with Logistic Regression and SVM	40

LIST OF FIGURES

Figure		Page
1.	An example of the learning video lecture. In this example, the lecture started from declaration of two variables, “a” and “b”, with value 1 and 2. After that, the lecture went through the comparison and addition of two integer variables. Finally, it explained the change of variable values and showed how to concatenate two string variables.	14
2.	An example of possible positions of cheating materials. There was no restriction for participants to put cheating materials during the testing session, Phase C. Participants were allowed to put the cheating materials at any place they liked.	15
3.	The four phases of experimental procedures. The experiment was a repeated-treatment design. Phase A was a learning session. Phase B was an online testing session without forbidden resources. Phases C was an online testing sessions with forbidden resources. Phase D was an interview session. During phase B and phase C, behaviors, such as mouse clicks, webcam videos and time on questions, were recorded.....	16
4.	An example the user interface for online exams with a proctoring system. Participants can see their own recorded video, and the remaining time in the exam. The next and previous button switches between questions.	17
5.	Examples of FACS for the affective state, confusion. (a) Netrual state; (b) eyebrow lower, AU4; (c) lid tightened, AU7. Reteived from http://face-and-emotion.com/dataface/facs/manual/TOC.html	24
6.	An example of coordinate and rotation of head pose. The positive <i>X</i> coordinate pointed to a test taker’s left. The positive <i>Y</i> coordinate pointed to the floor. The positive <i>Z</i> coordinate pointed away from the computer screen.	26
7.	An example of the affective state of confusion. In the pilot study, the participant showed confusion with AU4 (eyebrow lower) on his face while answering a question.....	36

Figure	Page
8. The interaction term of time delay and the amount of confusion. The solid line has the moderator, time delay, with low value, -1. The dash line has the moderator, time delay, with high value, 1. The value of time delay and confusion were whitened in pre-processing (zero mean with standard deviation one).....	39
9. The three proposed factors significantly associated with cheating. All three factors are positively related to cheating behaviors. The most relevant and non-redundant features in the head pose pair with time delay and confusion the head position relative to the monitor. In addition, time delay and the amount of confusion have an interaction term. The amount of confusion has stronger effect when time delay is shorter.	41

CHAPTER 1

INTRODUCTION

Online courses offer students the promise of an “anytime anywhere” education. Academic institutions are turning to online education in order to expand their reach and provide education to a greater volume and more diverse group of students, while at the same time using less faculty labor and less physical infrastructure than traditional face-to-face courses. However, the distributed nature of online courses presents a potential risk of increased academic dishonesty, particularly when students are asked to take exams at remote locations without a proctor in the room (Harmon & Lambrinos, 2008; Kennedy, Nowak, Raghuraman, Thomas, & Davis, 2000; Prince, Fulton, & Garsombke, 2009; Watson & Sottile, 2010). The prevention of academic dishonesty can be addressed to some extent by altering the assessments. For example, multiple versions of an exam could be used, question order could be randomized, or identical exam questions from previous semesters could be avoided (Harmon, Lambrinos, & Buffolino, 2010). However, a need remains to replace the traditional proctor in the room by another system to ensure the qualification of the online degrees offered by institutions (Frank, 2010; Harmon et al., 2010).

A survey of techniques and tools for proctoring remotely administered exams (Frank, 2010) found that the majority of solutions involve recording an exam attempt or streaming a live video to a proctor who will monitor or review the exam sessions from a remote location. In order to make effective use of the data, a trained human observer must devote time and attention to each student’s exam session. A naive approach of reviewing a set of recordings of individual exams may take significantly more effort than it would take a single proctor to monitor students in a traditional classroom setting. The costs of extra labor and technology devoted to proctoring exams must either be paid by the institution or passed on to students who wish to take courses online.

There are several important reasons for addressing issues of academic dishonesty in distance education. The first is an increasing trend in online education, both in terms of student enrollment and corporate market. Student enrollment has increased from 2.9 million in 2004-05 to

4.3 million in 2007-08 (National Science Board, 2012), and corporate market has increased from \$5.2 billion in 2007 to \$6.8 billion in 2010 (Adkins, 2011). Second, surveys of both faculty and students indicate a belief that cheating is more prevalent in online exams when students are not proctored (Kennedy et al., 2000; Watson & Sottile, 2010). Third, empirical studies have demonstrated that given the same online learning materials, scores in un-proctored exams were not only significantly higher than proctored ones (Prince et al., 2009) but also had significantly lower degrees of explanatory power to students' ability (Harmon & Lambrinos, 2008). Therefore, although online education provides opportunities to people who traditionally would not have access to high quality education due to schedule conflicts or physical constraints, these opportunities may be undercut if prospective employers do not trust the diplomas and certificates gained through online courses. The distrust of online degrees can be attributed in part to general beliefs of cheating in online assessments, and skewed test scores by which students' ability are hardly interpreted (Harmon et al., 2010).

Proctoring has been shown not only deter cheating in online assessments but also enhance learning performance in online courses. Wellman (2005) showed that online-module delivery paired with proctored quizzes was more effective in promoting learning when compared to un-proctored quizzes. The proctored group practiced more frequently than the un-proctored group, especially students in the bottom half of performers. In spite of the benefits, it can be impractical to supervise all quizzes in large online courses. Typically only high-stakes exams, such as midterms or final exams, are under surveillance (Luecht, 2006). A motivation of the proposed research is to make proctoring more effective in order to scale the use of online proctoring to low-stake quizzes, not just limited to high-stake exams.

The objective of this dissertation is to improve the effectiveness of online proctoring by modeling and detecting some of the patterns of behavior that may indicate academic dishonesty in online exams. The scope is limited to non-verbal cues associated with online cheating which can be reliably predicted with data collected by instruments that online students likely already have on their computers, such as a webcam, and a mouse. The four outcomes of this research are:

1. To test the relevance of a student's time to answer a question and self-reported uncertainty ratings as factors in predicting cheating.
2. To explore the relevant and non-redundant features of a student's visual focus of attention (VFOA) from webcam video streams and to evaluate the effectiveness of VFOA as factors in predicting cheating.
3. To determine the relevance of facial expressions that show a student's affective display (as observed via webcam recordings) as a factor in predicting cheating. A correlation between an affective display and academic dishonesty would be a strong motivation for future work to automatically detect the relevant facial expressions.
4. To identify the effectiveness of combined affect, VFOA and time delay as factors in predicting cheating, if all of them are significant. The classification accuracy of combined factors will be investigated.

The proposed multi-disciplinary research bridges areas including computer vision, image processing, surveillance, and learning science. Modeling student behaviors and predicting their intent to violate exam rules based on affect falls within the traditional domain of learning science; sensing and inferring the VFOA falls within the domain of computer science and computer vision.

CHAPTER 2

BACKGROUND

The purpose of this chapter is to provide the psychological, computational and educational context for understanding the rest of this document. It begins with a review of the research showing that online cheating is prevalent. Current online proctoring tools and services are examined and drawbacks are listed. Three possible non-verbal cues, time delay, visual focus of attention (VFOA), and affect, are discussed as indicators of academic dishonesty during remotely administered exams. Finally, the research questions and hypotheses are addressed.

Evidence that Online Cheating is Prevalent

Online courses offer a unique venue for academic dishonesty because of the indirect interaction between faculty and students. However just because cheating may be possible does not necessarily mean it occurs. It is important to conduct survey and show the evidence that the overall amount of cheating in online exam is more prevalent than in a traditional in-class environment.

Several authors have used surveys or anonymous polls as a basis to report the general beliefs of easiness of online cheating. For example, Kennedy et al. (2000) surveyed students across all academic areas and showed that 64% of 69 faculty members and 57% of 172 students felt that cheating was easier in online exams. A belief that cheating is easier in online versus traditional courses was also indicated in the research of King, Guyette and Piotrowski (2009), where 73.6% of 121 undergraduate business students agreed that it was easier to cheat in online.

There are also many reports showing that students admit that they are more likely to cheat in online courses. Chapman and his colleagues (Chapman, Davis, Toy, & Wright, 2004) found that 24% of 824 business students indicated that they had cheated on an electronic exam and 42% of them claimed they would cheat on electronic exams if given the opportunity. Students also indicated that electronic testing was one of several important situational determinants related to the probability of cheating. Lanier (2006) surveyed 1,262 students in a state-funded university and found that 41.1% admitted to cheating in online courses. Watson and Sottile (2010) surveyed 635 undergraduate and graduate students across all academic areas and found that students

were significantly more likely to obtain answers from others during an online test or quiz ($t(381) = -6.051, p < 0.001$) versus a traditional exam.

A challenge to the use of surveys to determine the prevalence of academic dishonesty is a possible self-report bias which may lead to under-reporting (Scheers & Dayton, 1987) or over-reporting the probability of cheating (Nelson & Schaefer, 1986). It can be argued that surveys may depend upon students to admit their guilt which may be to their perceived disadvantage and cause them under report dishonest behaviors, especially if the behaviors characterized as low on peer-group acceptability (e. g. breaking into a professor's office to steal an exam). Conversely, it can be argued that subjects perceive the deviant behaviors which are the norm in their group and therefore over report cheating than they actually engaged, especially if the behaviors characterized as high in peer-group acceptability (e.g. cheating on a test) (D. F. Crown & Spiller, 1998).

Empirical experiments that compare the results of proctored and unproctored exams in an online setting also support the notion that cheating is an issue. Harmon and Lambrinos (2010) used an R-squared statistic to analyze the results of proctored (24 students in a testing center) and un-proctored online exams (38 students) in a business school and showed evidence of cheating in un-proctored online exams. In their statistical model, human capital variables, such as GPA and the students' majors were used to explain the variation in test scores. They further investigated exam results within and across the proctored and un-proctored exams and concluded that human capital variables did not explain nearly as much of the variation in test scores in the un-proctored format as they did in the proctored format. The difference in performance has also been observed when the proctoring is done remotely. Prince et al. (2009) proposed an online proctoring tool for monitoring exams remotely among 76 business students. The test scores were compared by t-tests and it showed highly significant differences ($t(150) = 1.976, p < 0.001; d=0.32$) between average test scores for proctored (79%) versus un-proctored (87%) tests.

The evidence of cheating in online exams has been reported indirectly through surveys and has been validated directly through empirical studies. Both indicated that un-proctored, remote

exams were more likely to have a significant amount of cheating than proctored exams.

Furthermore, the difference in cheating has been demonstrated whether proctoring is done in testing centers (Harmon & Lambrinos, 2008) or using online tools (Prince et al., 2009). Both surveys and empirical studies support the notion that proctoring is an important factor in determining whether cheating will occur in remote exams.

Online Proctoring Tools and Services

Recently, Frank (2010) reviewed current commercial remote proctoring systems. Based on his review, current systems can be categorized into two types of proctoring methods: one is online, the other is offline. An online proctoring system means that monitoring is conducted in real time, for example, ProctorU (ProctorU, 2014) and Kryterion (Kryterion, 2014). An off-line proctoring system means that the testing sessions are recorded first and reviewed later, such as Securexam Remote Proctor (Securexam Remote Proctor, 2014) and VProctor (VProctor, 2014). Both monitoring systems, however, may cost more time and money to proctor all students or review all recorded testing sessions than face-to-face courses. For example, if there is a one-hour exam for one hundred students, in an online-exam setting, it will cost one hundred hours for a proctor to monitor or review all testing sessions. Suppose a proctor can monitor four students at the same time, it still costs 25 hours to monitor all. On the other hand, it will cost only one hour for several proctors, for instance, three, to supervise a large in-classroom exam. Similarly, if a proctor's hourly pay is ten dollars, in an online exam, it costs one thousand dollars to remotely proctor all students, or two hundred fifty dollars if a proctor monitors four students at one time. In an in-classroom exam, it only costs thirty dollars if there are three proctors.

An algorithm which can automatically flag suspicious behaviors can not only reduce the load on a proctor but also lower cost of an online education system. Current algorithms, such as audio detection and face detection, are effective at directing proctors attention to some salient parts of a recorded exam but leave rooms for improvement. For example, face detection can only assure that a person is in front of computer but cannot detect a behavior of inappropriately referencing in a textbook or a cell phone.

There is a research opportunity to determine which non-verbal cues associated with cheating behaviors can reliably be detected with current proctoring software systems and generally available computer instruments. Although there are a number of empirical studies demonstrating significant personal or situational factors which may lead to cheating shown as in Table 1 (Murdock, Hale, & Weber, 2001; Sierra & Hyman, 2008; Smith, Davy, Rosenberg, & Haight, 2003), it seems unlikely for schools to run a mass profiling survey of personal/situational constructs which determine the probability of cheating intentions. Even if the profiling study is approved by institutions, there is no established cutoff score to claim the cheating happens. Therefore, in this dissertation, personal or situational factors are excluded.

Table 1.

Models of predicting cheating Intentions based on personal/situational factors

Papers	Significant factors	Method/ Cheating behavior	Model built	Online exams	Sample
(Murdock, Hale, & Weber, 2001)	Academic motivations (self-efficacy, and extrinsic goal), social motivations (participation structure and teacher competence/commitment), and demographic data (grade in school).	Self-reports of cheating intentions	logistic regression analysis	N	Middle school
(Smith, Davy, Rosenberg, & Timothy Haight, 2003)	in-class cheating deterrents, prior cheating, and neutralization	Self-reports of cheating intentions	Structural equation modeling	N	Business program
(Sierra & Hyman, 2008)	Personal expertise, anticipated elation, and internal locus of control	Multi-item vignette measure of cheating intentions (Self-reports of cheating intentions)	structural equation modeling	N	Business program

In the next section, three proposed solutions, time delay, visual focus of attention (VFOA), and affective states, will be reviewed.

Possible Non-verbal Cues Related to Academic Dishonesty

This section reviews three proposed non-verbal cues that may indicate academic dishonesty and that can be determined from recorded webcam videos of a student taking an exam. First, it surveys the research in time delay and testing. Second, it discusses visual focus of attention (VFOA) and how it may be related to online cheating. These visual cues are explored as

a way to detect attempts to use forbidden resources, such as notes, textbooks, calculators, tablets or smartphones in the closed tests relying on remote proctoring (Frank, 2010). Third, it reviews affective states in learning and discusses affective states that may be useful to predict cheating in remote testing.

Time delay in testing. Van der Linden and his colleagues (van der Linden & Guo, 2008; van der Linden & Jeon, 2012) proposed two methods of detecting academic dishonesty based on person-fit analysis: erasure analysis and response time analysis. Person-fit analysis is a technique for determining if testing scores is valid by use of aberrant response patterns as indices (Meijer & Sijsma, 2001). The key to the approach is the availability of a response model that adequately represents regular behavior by the test takers (Meijer & Sijsma, 2001). If the results defy expectations of the model in the research of Van der Linden and Guo (2008) and Vander Linden and Jeon (2012), cheating is the factor which attributes to aberrant responses.

Erasure analysis measures the changing of answers in multiple choice questions. Van der Linden and Jeon (2012) showed that cheating could be identified as irregular behavior of aberrant changing rate of answers from wrong to right. Response time analysis measures the speediness of the test. Van der Linden and Guo (2008) showed that for a well-designed test (e.g. adaptive test), the response times (RTs) should follow the pattern of time intensities of the items in the test. Aberrant response-time patterns could be identified as cheating behaviors such as memorization of items during the test or foreknowledge of some of the items in the test pool.

Based on Van der Linden and Guo (2008) model that used response times as a potentially significant factor for cheating, it is hypothesized that the time a test taker spends on a single question plays a significant role in predicting students' decision to consult a forbidden resource. It is expected that test takers spend a greater amount of time to search for the answer than as opposed to answering the question honestly.

Affective states in learning. There is considerable evidence suggesting a set of basic emotional facial expressions that are innate and that cross cultural boundaries (Ekman & Friesen, 1978; Izard, 1971, 1994). Ekman and Friesen (1978) developed the Facial Action Coding System (FACS) to classify a set of facial expressions into six basic emotions by coding specific features

and muscles of the face. The six basic emotions coded in FACS are happiness, sadness, surprise, disgust, anger and fear (Ekman & Friesen, 1978). However, the adequacy of basing an entire theory of emotions on these basic emotions have been questioned in academic settings (Craig, Graesser, Sullins, & Gholson, 2004; Rozin & Cohen, 2003). Research shows that the six emotions are neither the most frequent nor the most significant emotions among the role of affective state during learning experience (Craig et al., 2004; Pekrun, Goetz, Titz, & Perry, 2002). Instead, confusion is the most significant affective state related to learning (Craig et al., 2004; VanLehn, Siler, Murray, Yamauchi, & Baggett, 2003).

Affective states in testing. One of the contributions of this dissertation is to test the hypothesis that confusion is significantly related to cheating while students are taking online exams. An early study by Bronzaft, Stuart, and Blum (1973), attempted to identify the relationship between anxiety and cheating, but no significant relationship has been found. Indeed, most research on affective states during the testing process focus more on how affective states can influence learning performance, rather than whether affective states can predict cheating (Birenbaum, 2007; Hembree, 1988). In the following section, the relationship among confusion, impasses and cognitive disequilibrium will be explained. After that, confusion will be proposed as a factor to predict cheating in the testing process, based on the theory of cognitive disequilibrium.

A cognitive system is in disequilibrium when individuals are confronted with problems or situations that present obstacles to goals, anomalous events, contradictions, discrepancies, and obvious gaps in knowledge (Graesser, Lu, Olde, Cooper-Pye, & Whitten, 2005). Impasses are the obstructions students encounter in academic settings. During learning, an impasse occurs when a student gets stuck, detects an error, or does an action correctly but expresses uncertainty about it (VanLehn et al., 2003). During testing, an impasse can be seen as the situation when a test taker does not know how to answer a question. Confusion is the physical exhibition while individuals hit impasses that turn the cognitive systems from equilibrium to disequilibrium (Craig, Mello, Witherspoon, & Graesser, 2008; Graesser et al., 2005; Otero & Graesser, 2001). In other words, confusion, impasses, and cognitive disequilibrium are similarly connected phenomena. An impasse is an event that causes a student to experience cognitive disequilibrium, cognitive

disequilibrium is an internal state of uncertainty, and confusion is the outward display that often signals the onset of cognitive disequilibrium.

Many people have found a significant relationship between confusion and learning (Craig et al., 2004; Mello & Graesser, 2011, 2012; Mello, Lehman, Pekrun, & Graesser, 2014; VanLehn, 1998; VanLehn et al., 2003), and it is suspected by the author that confusion plays a similar role in predicting cheating when it occurs during exams. During learning, the transition of cognitive system from disequilibrium to equilibrium has been shown to indicate a better understanding of the learning material (Craig et al., 2008; VanLehn et al., 2003). This strong connection makes confusion a predictor to learning. Similarly, test takers hit impasses during testing when they do not know the answers to some questions. Confusion is therefore exhibited, and cognitive systems are in disequilibrium. In response to an impasse in an exam, students may decide to guess, decide to cheat, remain in the impasse until they arrive at insightful answers, or remain in the impasse until time expires. Cheating is one of the possible solutions to restore the equilibrium in cognitive systems. The author proposes that the affective state, confusion caused by impasses, can provide the first observable indication or a gateway of future cheating behaviors.

Visual focus of attention (VFOA) in testing. The Visual Focus of Attention (VFOA) of a person is generally defined as the particular location in one's visual field where a person focuses in the attentive mode (Koch & Ullman, 1987). When students are taking exams, there is a need to read questions, understand problems, think possible solutions, and choose the correct answers. Test takers, therefore, will spend a fair amount of time focusing their attention on a screen. Current proctoring systems can capture screenshots in remotely administering settings (VProctor, 2014), so it is difficult for test takers to use forbidden resources on their computers without being caught since the resources would show in the screenshots and be visible to a proctor. However, proctors have a very limited view of students' local environments away from the computer screen, so the main cue that an external resource is being accessed will likely be that the VFOA of a test taker deviates from a computer screen. Gazing away from the screen may not always indicate an attempt to use a forbidden resource; another possibility may be that students have gaze aversion

while answering questions in order to reduce their cognitive loads (Doherty-Sneddon & Phelps, 2005).

There was a claim from employees who worked at the remotely proctoring company, ProctorU (2014), saying that the observable patterns of behaviors for normal people versus the people who tried to sneak in a cell phone and looked up information were clear, based on a report in the New York Times by Eisenberg (Eisenberg, 2013). We suspect that the salient cheating behaviors flagged by proctors are the directions of VFOA towards forbidden resources instead of the exam questions. It is because off-screen areas with significant VFOA should have more resources which not only cannot be seen by proctors but also are likely infringement of exam rules. If it is true, an application that can estimate VFOA may assist proctors' perception during supervision of online assessments. This research aims to validate this assertion empirically.

Research Questions and Hypotheses

Previous study investigated cheating based on personal/situational factors (D. Crown & Spiller, 1998; Murdock et al., 2001; Smith et al., 2003). However, it has been demonstrated that the certainty of being caught, or past example of a student caught cheating, is a significant deterrent to cheating (Nagin & Pogarsky, 2003). Although there are numerous remotely proctoring software (ProctorU, 2014; Secureexam Remote Proctor, 2014; VProctor, 2014), there is a need to enhance the process of remotely monitoring procedures, especially in a large online courses. The current study investigates possible non-verbal cues which may indicate online cheating. Based on the previous findings, it is possible that time delay (van der Linden & Guo, 2008), head pose deviated from screen (Eisenberg, 2013), and confusion (Craig et al., 2008) are significant factors related to cheating. Additionally, since there are many possible ways for students to violate exam rules, the scope of this research is limited to methods that are important and detectable through a remote proctoring system. The simplification is that only the cheating behaviors of using forbidden items, such as notes, textbooks, calculators, tablets, or smartphones, are investigated. Although not a comprehensive solution to prevent academic dishonesty, this is an important topic because one of the major motivations for online proctoring systems is to prevent students from accessing forbidden resources during closed-book exams (Frank, 2010).

This research seeks to answer the following research questions:

1. Can time delay while answering a question be a non-verbal cue to reliably predict cheating behaviors during online exams?
2. Can students' self-reported certainty rating have significant relationship related to cheating behaviors during online exams?
3. Can facial expressions associated with confusion be a non-verbal cue to reliably predict cheating behaviors during online exams? If so, the implications are that proctors can be trained to look for these expressions and software could be developed to highlight points in a recording where the student exhibits the expression.
4. Can VFOA be a non-verbal cue to reliably predict cheating behaviors during online exams? If it is a reliable cue then automatic or semi-automatic methods to detect VFOA on exams are warranted, and information on which VFOA patterns might relate to violations can be used to train proctors.

In order to answer the three research questions, there are four null hypotheses for each question:

$H1_0$ A test taker's time delay on each question has no statistically significant relation to cheating decisions during online exams.

$H2_0$ A test taker's certainty rating on each question has no statistically significant relation to cheating decisions during online exams

$H3_0$: VFOA deviated from the computer screen have no statistically significant relation to cheating behaviors during online exams.

$H4_0$: Facial expressions associated with confusion have no statistically significant relation to cheating behaviors during online exams.

The author aims to estimate VFOA automatically and to estimate confusion manually. In the next chapter, the experiment design of this dissertation will be addressed.

CHAPTER 3

METHODS

The goal of this chapter is to simulate an online-testing scenario and to explore the significant non-verbal cues. The section is organized as follows. First, it describes the number of participants and how these participants are recruited. Second, it presents the experimental materials which were used for learning, testing, and interviewing. Last, it gives the details of the design of procedures of the experiment.

Participants

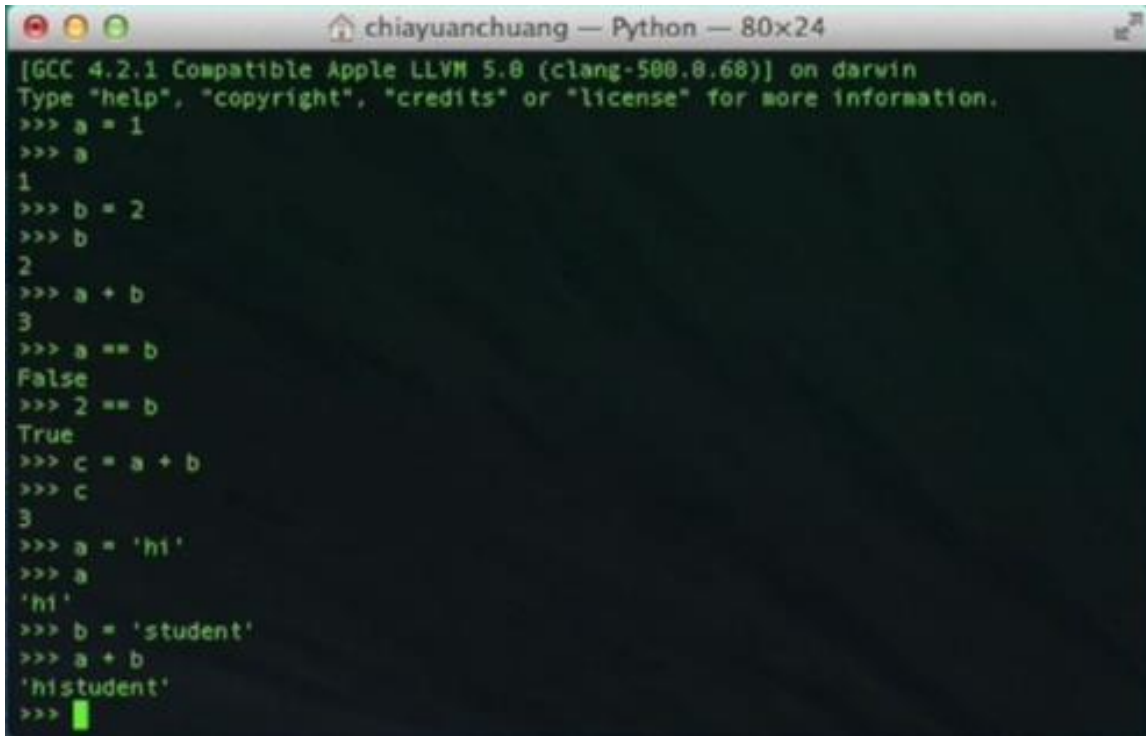
Recruitment was performed via “sona-systems” from the ASU poly subject pool (<http://asup.sona-systems.com>). Forty-two students (28 male, 14 female) took part in the study. They were between the ages of 18 and 36 ($M = 20.93$, $SD = 3.90$). Participants were undergraduate students enrolled in an Introductory Psychology course. They were offered partial course credit in return for their participation in the study. The exemption from Institutional Review Board (IRB) at Arizona State University is shown in the APPENDIX A.

Materials

Learning materials. The learning materials were a 12 minutes video covering the basics of the Python computer programming language. Two pages of printed summary along with the video lecture were provided to participants during the learning phase. In addition, one piece of blank paper was provided to participants. Participants were allowed to take their own notes either on the two pages of summary or on the blank page while watching the video. Python program was selected as the domain because most participants would not have been exposed to it before the study. So, it was less likely they would already know the answers to the test.

The Python lecture started by the introduction of Python interpreter, such as entering and leaving Python interpreter through terminal, different types of variables in Python, declaration of variables and assigning new values to the declared variables. After that, the lecture went over some default operators and functions in Python, for example, modulo and comparison operators and a length function. Finally, participants were taught how to declare and execute their own

functions in a Python file. *Figure 1.* shows a screen shot of the learning video. The details of learning materials are provided in APPENDIX B.



```
[GCC 4.2.1 Compatible Apple LLVM 5.8 (clang-500.8.68)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> a = 1
>>> a
1
>>> b = 2
>>> b
2
>>> a + b
3
>>> a == b
False
>>> 2 == b
True
>>> c = a + b
>>> c
3
>>> a = 'hi'
>>> a
'hi'
>>> b = 'student'
>>> a + b
'histudent'
>>>
```

Figure 1. An example of the learning video lecture. In this example, the lecture started from declaration of two variables, “a” and “b”, with value 1 and 2. After that, the lecture went through the comparison and addition of two integer variables. Finally, it explained the change of variable values and showed how to concatenate two string variables.

Testing materials. Two ten-item multiple choice tests were randomly presented to participants. Both of these tests covered the material presented on the Python programming language, but each test had unique questions and covered different concepts. The first test session was implemented within a typical online exam setting. The second session was implemented in a cheating inducing environment in which participants were encouraged to answer questions by all means even if cheating. The second session was used to ensure that some cheating behaviors were observed. The details of the testing materials in the first session are shown in the TESTING MATERIALS

Phase B of APPENDIX C. The details of the testing materials in the second session are shown in the Phase C of APPENDIX C.

Cheating materials. The notes and documents provided in the learning phase were returned to participants as cheating materials. The cheating materials included one page of self-written notes and two pages of summary of the video lecture (see APPENDIX B). Participants were allowed to put the cheating materials at any place they liked (see *Figure 2*).

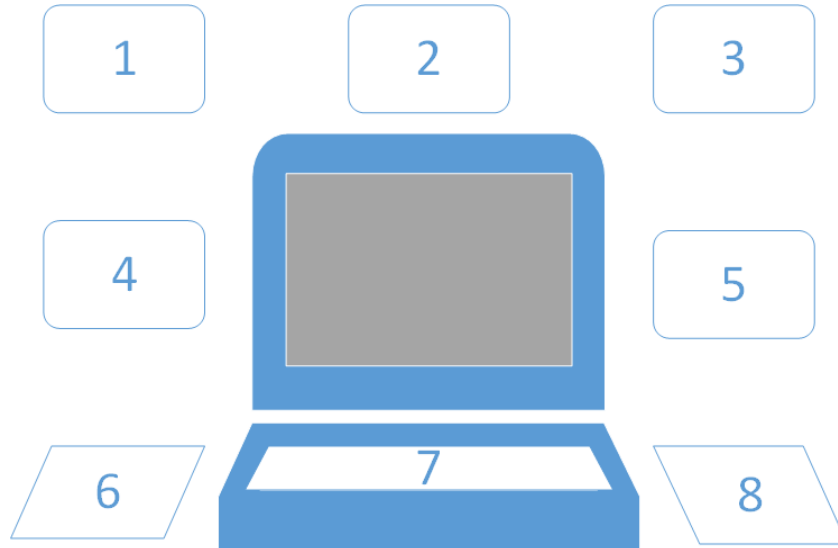


Figure 2. An example of possible positions of cheating materials. There was no restriction for participants to put cheating materials during the testing session, Phase C. Participants were allowed to put the cheating materials at any place they liked.

Interview materials. At the end of each testing phases, participants were interviewed by the experimenter. There were two types of interview questions: one was behavior questions, and the other was demographic questions. The behavior interview required participants to provide a self-reported certainty rating, self-reported cheating, and methods for cheating for each question. The participant's certainty rating consisted of a scale from one to five, where one indicates a guess and five indicates knowledge with high confidence. The demographic interview required participants to provide the information of major, prior knowledge of computer programming, gender, age and the intention of preparation of cheating materials. Prior knowledge was assessed by self-reports with a scale from zero to five, while zero indicates no knowledge on

computer programming and five indicates mastering the topic before the experiment. The details of interview materials are shown in the APPENDIX D.

Procedure

The experiment was a repeated-treatment design, which means that one subject answers multiple questions in two different exam settings. Whether test takers cheated will be coded in the interview process. Figure 3 shows a diagram of the overall process that participant's undertook. After participants arrived at the lab and completed the informed consent procedure, they completed the study which consisted of four phases labeled phase A through D in Figure 3. The instructions to participants in the four phases of experiment process are listed in APPENDIX E.

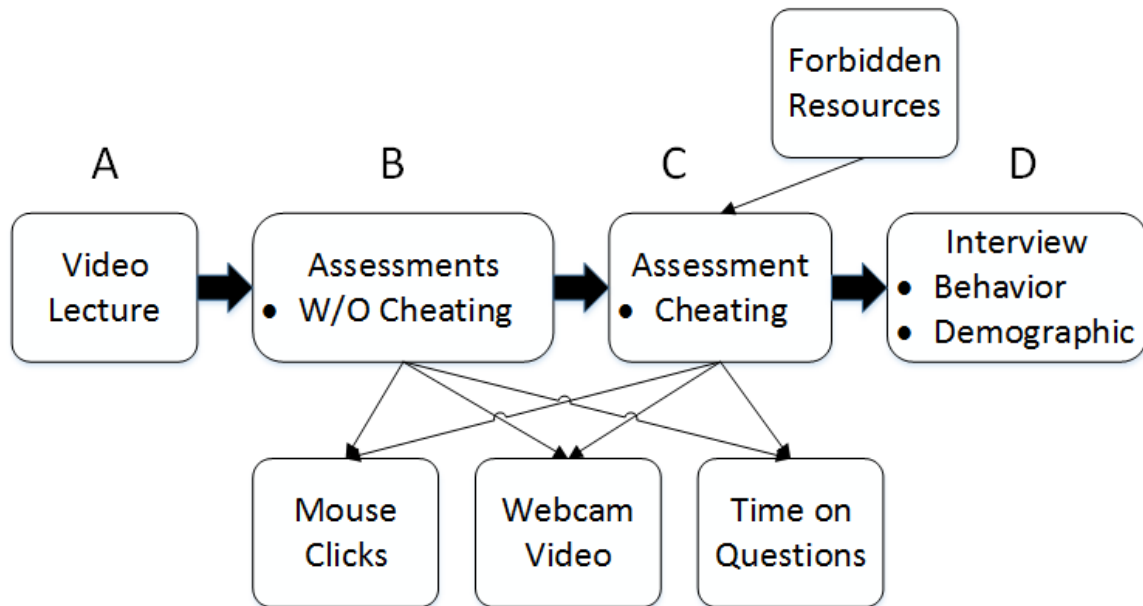


Figure 3. The four phases of experimental procedures. The experiment was a repeated-treatment design. Phase A was a learning session. Phase B was an online testing session without forbidden resources. Phases C was an online testing sessions with forbidden resources. Phase D was an interview session. During phase B and phase C, behaviors, such as mouse clicks, webcam videos and time on questions, were recorded.

Phase A was a learning session. After the informed consent process was completed and initial instructions presented to participants, they started the learning phase of the study. In this phase, participants were asked to watch a 12 minute video lecture on the subject of programming in the Python language. One piece of blank paper and two pages of printed summary were

provided to participants as notes and learning materials. After finishing the video lecture, the experimenter went through the learning documents to make sure that participants not only understood the content in the video lecture, but also were familiar with the content in the learning documents. Therefore, participants could find the answers in the learning materials while using them as cheating materials. The details of consent form is listed in APPENDIX F.

Phase *B* was a replication of a typical online testing setting. Participants took a 13 minute online exam in which forbidden resources such as smart phones and cheat sheets were not allowed. The computer system was also locked into the testing program, by which participants could not leave the exam window until they finished the exam. Only the experimenter knew the special keystroke combination to leave the testing program. So, no online resources could be accessed during the experiments. The time and positions of mouse clicks, webcam videos, and time on questions were recorded during this experiment. However, analysis of the dataset collected in the phase *B* was beyond the scope in this dissertation. The goal of this session was to make participants familiar with the testing environment. An example of user interface in the Phase B is shown in Figure 4.

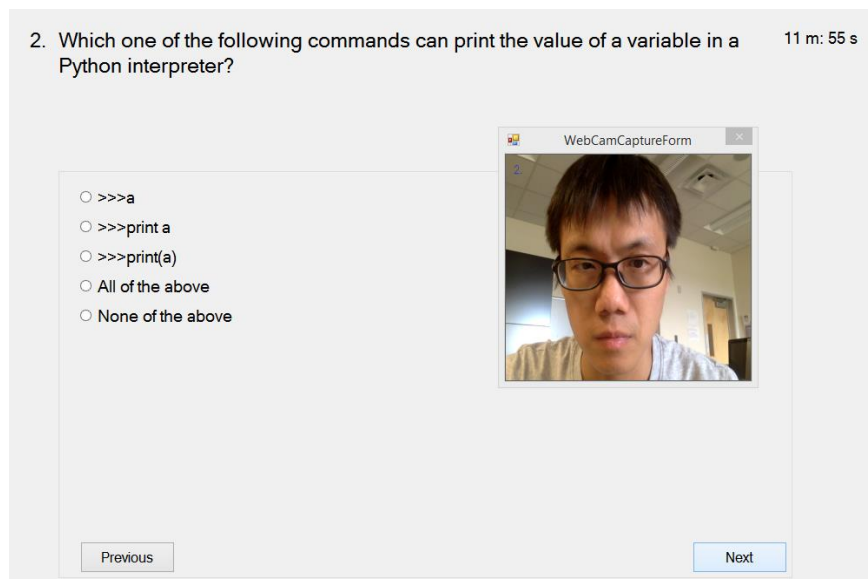


Figure 4. An example the user interface for online exams with a proctoring system. Participants can see their own recorded video, and the remaining time in the exam. The next and previous button switches between questions.

Phase C was a cheating inducing environment. Participants again took a 13 minute exam and were asked to answer the questions as best as they could, including cheating without being caught by an online proctoring system. The forbidden materials were returned to participants. Participants had 5 minutes to arrange their cheating materials before starting the exams. The time and positions of mouse clicks, webcam videos, and time delay in each question were recorded by a proctoring system. In this dissertation, only time delay and webcam videos in the phase C was analyzed.

In phase D, the experimenter asked behavior questions first and conducted a demographic survey later. The experimenter stepped one by one through each assessment item and asked participants' certainty ratings and whether they cheated in each question. If they cheated, they received a follow up questions of how they cheated on the question. The testing behavior questions in the phase C were interviewed first in order to retrieve more solid memory associated with participants' cheating patterns. After that, participants were given a demographic survey, including their major, previous experiences in computer programming, gender, age and the year in school.

Discussion

There are many possible ways for students to violate exam rules, so the scope of the proposed research is limited to the methods that are important and detectable through a remote proctoring system. The first simplification is that only the cheating behaviors of using forbidden items, such as notes, textbooks, calculators, tablets, or smartphones, will be investigated. Although there is no comprehensive solution to prevent academic dishonesty, based on the survey of Frank (2010), one of the major motivations for online proctoring systems is to remotely monitor test takers' behaviors, especially for the misuse of prohibited resources. Therefore, the goal in this research is to enhance the remote proctoring processes and prevent test takers from accessing forbidden resources during closed-book online exams.

The second simplification is that each cheating event is discrete with binary response, which means that students either answer questions honestly or dishonestly. The levels of academic dishonesty students have involved are disregarded. In this study, it is hypothesized that

there are certain common features that reveal significant differences between honest and dishonest behaviors during online exams. One of major goals is to investigate these important features that can classify violation of academic integrity or not automatically.

As a third simplification, it is assumed that the cheating events among subjects are independent and identically distributed. The interactions of eliciting cheating behaviors among different questions to one subject and the interactions of one question to different subjects are excluded in the model proposed in this dissertation. It is because this study is an exploratory experiment, which aims to investigate non-verbal cues, not only significantly but also commonly related to online cheating behaviors, rather than propose a full model, which describes the test-taking process from the beginning to the end of time. The investigation of full process of test-taking behavior will be described as future works in the section of Limitations and Future Research (See).

CHAPTER 4

THE ROLE OF STUDENT'S CERTAINTY AND TIME DELAY

The primary contribution of this chapter is to test the ability of test taking behavior to predict student cheating during online exams. Specifically this chapter tests the impact of a student delay time to answer a question and the student certainty rating for the question. The design of experiment is described in CHAPTER 3. This chapter begins with the review of related research, including response times in testing and impasses in learning. The research questions and hypotheses are addressed in the section of Current Study. Based on the results of hierarchical logistic regression, it was concluded that not only time delay but also uncertainty had positively significant relationship to cheating. Finally, the importance of the two significant factors was discussed.

Response Times in Testing

One of the factors proposed in this paper is time delay in online testing, which may indicate suspicious behaviors in online exams. The traditional approach to detecting aberrant behavior is to use person-fit analysis by which aberrant response patterns that defy some expectation can be an index to validate the integrity of test scores (Meijer & Sijtsma, 2001). Although the methods of person-fit analysis have been well developed for the last two decades, the reasons for aberrant response patterns are still largely unknown. However, cheating is one of the reasons for aberrant response patterns (Meijer & Sijtsma, 2001; Petridou & Williams, 2007).

In (2008), Van Der Linden and Guo used response times (RTs) as an additional source of information on the test taker's behavior. They conducted a simulation study on two cheating behaviors: (1) pre-knowledge of some of the items; (2) attempts to take tests only for the purpose of memorizing the items. However, the two types of cheating behaviors, such as pre-knowledge and memorization of testing items (van der Linden & Guo, 2008), are too difficult for proctors to identify remotely. Since the aim of this dissertation is to assist the online proctoring processes, instead, the focused cheating behavior in this study is the misuse of prohibited resources, which is the primary cheating behaviors proctors monitor and prevent in the remotely proctoring systems (Frank, 2010).

Impasses in Testing

The other factor we focus on is students' certainty rating on the scale from one to five, where one indicates a guess and five indicates knowledge with high confidence. The certainty rating as a factor is inspired by the theory of impasses during learning. The details of literature review of an impasse in learning and how it is related to cheating was described in the CHAPTER 2.

Current Study

Rather than focusing on personal or situational factors (D. F. Crown & Spiller, 1998), this paper examined test taking behaviors. Specifically it investigated if students' delay time to answer a question and their certainty rating for the question impacted their decision to cheat. This study implemented a common metric to define cheating behaviors used by a majority of proctoring systems: a misuse of forbidden resources, such as a smart phones or cheat sheets (Frank, 2010).

Based on Van der Linden and Guo (2008) model that used response times as a potentially significant factor for cheating, it is hypothesized that the time a test taker spends on a single question plays a significant role in predicting students' decision to consult a forbidden resource. It is expected that test takers spend a greater amount of time to search for the answer than as opposed to answering the question honestly.

Additionally, the student's confidence in their ability to answer the question correctly could impact their cheating behavior. When a student encounters a question that they cannot answer or have difficulty answering, their level of certainty will decrease. This inability to answer the question is the equivalent to an impasse (VanLehn et al., 2003) in a learning setting. During testing, there are no learning opportunities remaining to work pass the impasse. If the student is sufficiently motivated to provide a correct answer and resources are readily available with limited monitoring within the online setting, as uncertainty level increase, students are more likely to cheat because it is the only way to resolve the impasse and move forward.

This chapter seeks to answer the following two research questions: (1) Can time delay be an indication reliably predict cheating decisions during online exams? (2) Can student's certainty rating be an indication to reliably predict cheating decisions during online exams? The two null

hypotheses $H1_0$ and $H2_0$ are addressed in the section of Research Questions and Hypotheses (see CHAPTER 2).

Results

A hierarchical logistic regression was conducted using participants, previous experience, time delay and certainty rating as predictor variables to predict student's cheating behavior (criterion variable). The initial model had two predictors, subject and previous experience. Both of these measures were categorical and repeated for each instance of the test item (i.e., repeated ten times for each participant). This indicated that these two variables were not significant predictors of cheating, $\chi^2(2) = 1.49$; $p > 0.05$. The first model had an $R^2 = 0.004$. A second model added time delay as a predictor. This addition resulted in a significant model, $\chi^2(1) = 48.91$; $p < 0.001$. The delta R^2 between the first model and the second model are 0.109 and the second model had an $R^2 = 0.113$. Finally, in the third model, certainty rating is added in addition to the previous predictor variables resulting in a significant change, $\chi^2(1) = 7.81$; $p < 0.01$. The delta R^2 between the second model and the third model were 0.016 and the third model had $R^2 = 0.129$. This indicates that student's certainty rating is a significant factor to predict cheating decisions. Time delay has the strongest predictive power with $\exp(B) = 1.00$; $t(1) = 29.01$; $p < 0.001$, which means an increment of one second will increase about one percent of odds ratio of cheating. Certainty was also significant predictor with negative relationship as certainty decreased likelihood of cheating increased, $\exp(B) = 0.757$; $t(1) = 7.83$; $p < 0.01$. The results of full model is shown in Table 2. The mean and S.D. of time delay and certainty rating for cheating and non-cheating items is provided in Table 3.

Table 2

Results of the third model of hierarchical logistic regression with predictors subjects, experiences, time delay, and students' certainty ratings

<i>Predictor</i>	<i>exp(B)</i>	<i>t</i>	<i>df</i>	<i>p</i>
<i>Participants</i>	1.012	1.583	1	$p > 0.05$
Prior Experiences	0.955	0.311	1	$p > 0.05$
<i>Time Delay</i>	1.000	29.013	1	$p < 0.001^{***}$
<i>Certainty</i>	0.757	7.826	1	$p < 0.01^{**}$

Note: $N = 420$; $\chi^2(4) = 58.208$; $p < 0.001^{***}$; $R^2 = 0.129$

Table 3

Mean and S.D. for time delay and certainty rating

Cheated	Time (sec)		Certainty	
	Mean	SD	Mean	SD
Yes	58.84	33.28	3.51	1.24
No	37.14	26.43	4.08	1.06

Discussion

Both null hypotheses, $H1_0$ and $H2_0$, were rejected and it was concluded that not only time delay but also certainty rating of each question were significant predictors of test takers' cheating decisions. Specifically, the probability of cheating was positively related to time delay but negatively related with the participant's certainty rating. The results also indicated that neither subjects themselves nor their experiences in the test materials significantly predicted cheating decisions. The strongly positive relationship between time delay and academic dishonesty matches the expectation that given the opportunity to cheat, cheaters spend more time in consulting forbidden resources than non-cheaters. The significant relationship between uncertainty rating and cheating behaviors is also compatible with impasse theory (VanLehn et al., 2003).

The strength in this study was that the proposed factors, such as time delay, can be monitored in real time. The factors explored in the previous research were personal/situational

constructs, which ignore the dynamic behaviors of test takers during online testing. Moreover, given the personal/situational factors, it seems unlikely for schools to run a mass profiling survey. The proposed factor, time delay, provided an objectively quantitative measurement which can be easily implemented and coped with current online proctoring systems.

In this study, the classification accuracy, is currently unknown. Further research is still needed to determine the accuracy for classification but the significant features found in this paper are a start toward such as a model of detecting suspicious behaviors during remote testing. The application of this research could be significant time reduction in remote proctoring.

Since current proctoring system can record test takers behaviors during online exams, including facial expressions, it is possible that a test taker's certainty rating of each question can be assessed more objectively based on affective states, for example, confusion (Craig et al., 2008). Craig et al. (2008) found that the physical exhibition of confusion has a significant relationship to observable human facial action units (Ekman & Friesen, 1978), especially for AU 4 and AU 7. AU 4 indicates lower eyebrows and AU 7 demonstrates tightened lids (see Figure 5.). The analysis of confusion and delay time through recorded videos provides an opportunity for proctoring systems to monitor test takers' certainty rating in real time. In the next chapter, the amount of confusion during testing will be analyzed and discussed.

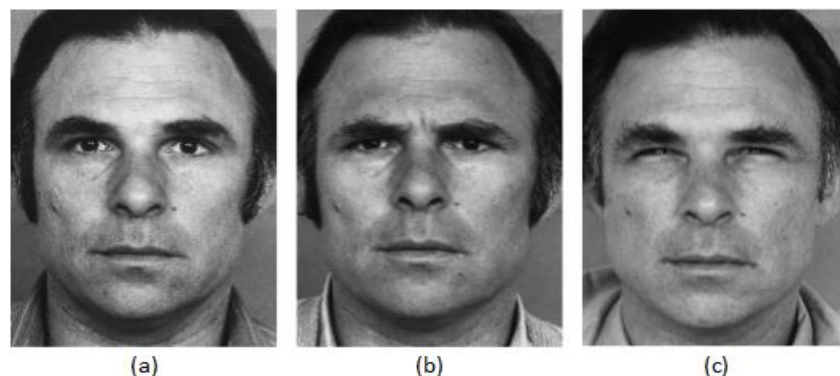


Figure 5. Examples of FACS for the affective state, confusion. (a) Neutral state; (b) eyebrow lower, AU4; (c) lid tightened, AU7. Retrieved from <http://face-and-emotion.com/dataface/facs/manual/TOC.html>

CHAPTER 5

DETECTING CHEATING BASED ON TIME DELAY AND HEAD POSE

Previous research investigated personal/situational factors to predict cheating intentions rather than using a student's behavior during online exam settings (Chuang, Craig, & Femiani, 2015). The objective of this paper is to explore a method for automatic detection of cheating based on testing behaviors. The primary contribution is to propose a model that can be used to flag suspicious activities within recordings of online exams. Similar to CHAPTER 4, the design of experiment is described in CHAPTER 3. The explored predictors were 48 statistical features based on Visual Focus of Attention (VFOA) and time delay during testing. The final model is constructed by logistic regression and support vector machine with Fast Correlation-Based Filter (Yu & Liu, 2003)

Current Study

The current study investigates whether student's behaviors can predict cheating during online testing situations. The first goal of the study will be to find the relevant and non-redundant VFOA features related to cheating. The second goal will be to determine which features are predictive of cheating. Based on previous work, it is hypothesized that factors of time delay (Chuang et al., 2015) and VFOA (Eisenberg, 2013; Koch & Ullman, 1987) can predict cheating behaviors. However, it is currently unknown which, if any, VFOA features will be influential for cheating detection. Based on this hypothesis, three predictions have been tested. The first prediction is that changes in VFOA will be a significant cues related to cheating behavior. Second, VFOA features paired with a time delay will be significantly related to cheating behaviors. Third, VFOA and Time Delay will provide an acceptable capability to detect cheating, meaning that AUC is at least greater than 0.7 (Hosmer Jr & Lemeshow, 2004).

Data Treatment

Features in VFOA. The student VFOA was extracted from the webcam videos recorded during online testing. The recorded videos were segmented into intervals corresponding to individual questions first, and then analyzed by the Constrained Local Model (CLM) proposed by Baltrusaitis, Robinson, and Morency (2012), which uses a generalized adaptive view-based

appearance model (GAVAM). The CLM-GAVAM algorithm was chosen because it has low error in head pose estimation (yaw:3.00°; pitch:3.81°; roll:2.08° in mean absolute error). The total number of video segments was 420 and only 409 events were selected while 11 events were discarded. The 11 events were excluded because subjects' faces were occluded by hand gestures, and therefore there were no head-pose features extracted. Figure 6 shows an example of the coordinates and rotation of head pose.

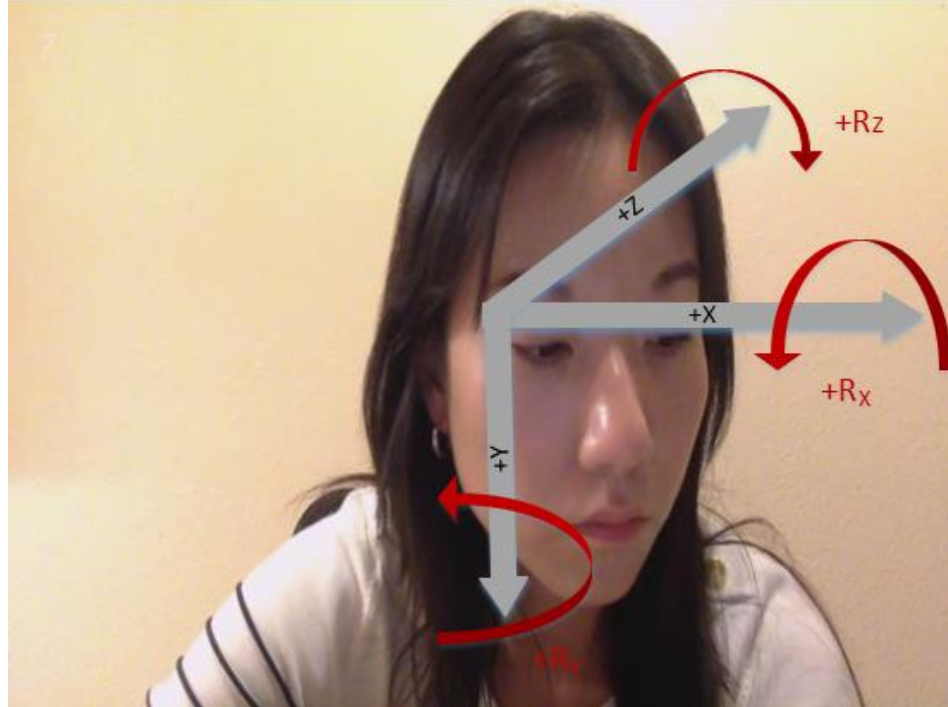


Figure 6. An example of coordinate and rotation of head pose. The positive X coordinate pointed to a test taker's left. The positive Y coordinate pointed to the floor. The positive Z coordinate pointed away from the computer screen.

During the experiments, each individual situated themselves differently in front of their screen. Therefore, head pose rotation was recorded relative to their modes, where the modes were estimated by using robust regression (M-estimator), proposed by Huber (1964). After that, all features were whitened, with zero mean and standard deviation one. The definitions of eight different statistical features are shown in Table 4.

Table 4

Eight statistical features of VFOA and their definitions

Name	Denotation	Definition
Min value	$Min(V)$	Minimum value in a question
Max value	$Max(V)$	Maximum value in a question
Mean value	$Mean(V)$	Average value in a question
Stdev value	$Stdev(V)$	Standard deviation of the values in a question
Range value	$Range(V)$	$Max(V) - Min(V)$
Number of Min 5% value	$Num_{min}(V)$	The number of frames which have lowest 5 % values of $Range(V)$ in a question
Number of Max 5% value	$Num_{max}(V)$	The number of frames which have highest 5 % values of $Range(V)$ in a question
Absolute Slope of value	$Slope_{abs}(V)$	A measure for the average local variability in value V for all frames f_i , where $Slope_{abs}(V) = \frac{1}{N-1} \sum_{i=2}^N \frac{ V(f_i) - V(f_{i-1}) }{f_i - f_{i-1}}$

Note: The value, V , can be six different head pose measurements: P_X , P_Y , P_Z , R_X , R_Y , and R_Z . An image of the six different head poses can be seen in Figure 6

Fast Correlation-Based Filter (FCBF). The VFOA contains total 48 features, including six head poses features multiply by eight statistical features (see the previous section: Features in VFOA.). In order to reduce the high dimensionality of features the Fast Correlation-Based Filter (FCBF) method proposed by Yu and Liu (2003) was used to remove irrelevant and redundant features. The details of FCBF can be seen in the paper Yu and Liu (2003). In this paper, only a summary of this work is described.

FCBF uses a measure called symmetrical uncertainty (Press, Teukolsky, Vetterling, & Flannery, 1988) to rank features. Symmetrical Uncertainty (SU) is an entropy-based measurement of two random variable X and Y . A $SU(X, Y)$ value of 1 indicates that by given one random variable X , values in the other random variable Y could be perfectly predicted (or vice-

versa); a $SU(X, Y)$ value of 0 indicates that the two random variables are totally independent. For details of SU, see Press et al. (1988).

The FCFB works by ranking all features based on their ability to predict the class attribute, and pruning out the features whose predictive value falls below a user-defined threshold. Starting with the highest ranked feature, a list of “predominant” features is built incrementally. At each iteration, the highest ranked feature is added to a list of predominant features and all other features that can be predicted by the highest ranked feature better than by the class attribute, according to the value of SU, are removed from consideration. When all features have either been eliminated or added to the list of predominant features, the algorithm terminates. The predominant list is the selection of relevant and non-redundant features in the subset.

Results

Feature selection of VFOA. Weka (Hall et al., 2009), an open-source tool for machine learning, was used for FCBF feature selection of VFOA. The number of features identified by FCBF depends on a user-defined threshold. In this study, the threshold of FCBF was kept at the system default value, which was negative one times the largest value in double. The results of SU value for each feature are listed in Table 5. In the Table 5., the features in the first column are listed based on $SU(\text{Feature1}, \text{Cheating})$ values in the descending order. The relevant and non-redundant features were $Max(P_X)$, $Num_{min}(P_X)$, and $Stdev(P_Z)$, listed in the third column of Table 5. A logistic regression model was used to predict the binary response variable, cheating based on the three relevant and non-redundant features selected by FCBF. This model indicated that the predictor variables could significantly identify cheating ($\chi^2(3) = 35.865$; $p < 0.001$). All of the three selected features, $Max(P_X)$, $Stdev(P_Z)$ and $Num_{min}(P_X)$, had significantly positive relationship to cheating. The results are shown in Table 6.

Table 5

Results of Fast Correlation-Based Filter (FCBF) for head pose features.

<i>Feature1</i>	<i>SU(Feature1,Cheating)</i>	<i>Feature2</i>	<i>SU(Feature2,Feature1)</i>
<i>Stdev(P_Z)</i>	0.112	*	
<i>Range(P_Z)</i>	0.100	<i>Stdev(P_Z)</i>	0.694
<i>Range(P_X)</i>	0.091	<i>Stdev(P_Z)</i>	0.224
<i>Range(P_Y)</i>	0.084	<i>Stdev(P_Z)</i>	0.331
<i>Range(R_X)</i>	0.079	<i>Stdev(P_Z)</i>	0.383
<i>Stdev(P_X)</i>	0.075	<i>Stdev(P_Z)</i>	0.175
<i>Stdev(P_Y)</i>	0.075	<i>Stdev(P_Z)</i>	0.414
<i>Range(R_Z)</i>	0.074	<i>Stdev(P_Z)</i>	0.219
<i>Stdev(R_X)</i>	0.072	<i>Stdev(P_Z)</i>	0.268
<i>Max(R_X)</i>	0.067	<i>Stdev(P_Z)</i>	0.164
<i>Range(R_Y)</i>	0.065	<i>Stdev(P_Z)</i>	0.285
<i>Max(P_X)</i>	0.065	*	
<i>Stdev(R_Y)</i>	0.058	<i>Stdev(P_Z)</i>	0.258
<i>Min(R_Z)</i>	0.058	<i>Stdev(P_Z)</i>	0.113
<i>Min(R_Y)</i>	0.057	<i>Stdev(P_Z)</i>	0.198
<i>Stdev(R_Z)</i>	0.056	<i>Stdev(P_Z)</i>	0.207
<i>Max(R_Z)</i>	0.055	<i>Stdev(P_Z)</i>	0.155
<i>Max(R_Y)</i>	0.054	<i>Stdev(P_Z)</i>	0.149
<i>Slope_{abs}(P_Y)</i>	0.052	<i>Stdev(P_Z)</i>	0.204
<i>Min(R_X)</i>	0.049	<i>Stdev(P_Z)</i>	0.269
<i>Min(P_Z)</i>	0.049	<i>Stdev(P_Z)</i>	0.099
<i>Slope_{abs}(R_X)</i>	0.049	<i>Stdev(P_Z)</i>	0.133
<i>Num_{min}(P_X)</i>	0.045	*	
<i>Slope_{abs}(P_Z)</i>	0.040	<i>Stdev(P_Z)</i>	0.187
<i>Slope_{abs}(P_X)</i>	0.039	<i>Stdev(P_Z)</i>	0.157
<i>Num_{min}(P_Z)</i>	0.038	<i>Stdev(P_Z)</i>	0.047

Table 5 (continue)

Results of Fast Correlation-Based Filter (FCBF) for head pose features.

<i>Feature1</i>	<i>SU(Feature1,Cheating)</i>	<i>Feature2</i>	<i>SU(Feature2,Feature1)</i>
$Num_{max}(P_Y)$	0.037	$Stdev(P_Z)$	0.042
$Num_{min}(R_Z)$	0.036	$Num_{min}(P_X)$	0.084
$Num_{max}(P_Z)$	0.030	$Stdev(P_Z)$	0.091
$Mean(P_X)$	0	$Stdev(P_Z)$	0
$Min(P_Y)$	0	$Stdev(P_Z)$	0
$Max(P_Y)$	0	$Stdev(P_Z)$	0
$Mean(P_Y)$	0	$Stdev(P_Z)$	0
$Num_{max}(P_X)$	0	$Stdev(P_Z)$	0
$Slope_{abs}(R_Z)$	0	$Stdev(P_Z)$	0
$Num_{min}(P_Y)$	0	$Stdev(P_Z)$	0
$Mean(R_Y)$	0	$Stdev(P_Z)$	0
$Num_{max}(R_Y)$	0	$Stdev(P_Z)$	0
$Slope_{abs}(R_Y)$	0	$Stdev(P_Z)$	0
$Mean(R_Z)$	0	$Stdev(P_Z)$	0
$Num_{min}(R_Y)$	0	$Stdev(P_Z)$	0
$Num_{max}(R_X)$	0	$Stdev(P_Z)$	0
$Max(P_Z)$	0	$Stdev(P_Z)$	0
$Num_{min}(R_X)$	0	$Stdev(P_Z)$	0
$Mean(P_Z)$	0	$Stdev(P_Z)$	0
$Num_{max}(R_Z)$	0	$Stdev(P_Z)$	0
$Mean(R_X)$	0	$Stdev(P_Z)$	0
$Min(P_X)$	0	$Stdev(P_Z)$	0

Table 6

Results of logistic regression for features of VFOA

Predictor	$\exp(B)$	t	df	p
$Max(P_X)$	1.680	10.805	1	$p < 0.01^{**}$
$Num_{min}(P_X)$	0.767	3.886	1	$p < 0.05^*$
$Stdev(P_Z)$	2.182	9.922	1	$p < 0.01^{**}$

Note: $N = 409$; $\chi^2(3) = 35.865$; $p < 0.001$; $R^2 = 0.084$

VFOA combined with time delay. Hierarchical logistic regression was used to test VFOA combined with Time Delay (TD). The initial model had three predictors, $Max(P_X)$, $Num_{min}(P_X)$, and $Stdev(P_Z)$, which were selected based on FCBF (see section 3.1.). The results of the first model indicated that three predictor variables can identify cheating behaviors (See Table 6). The second model added one predictor, time delay (TD). It showed that time delay can significantly predict cheating, $\chi^2(1) = 25.614$; $p < 0.001$. The results of the full hierarchical model are shown in Table 7.

Table 7

Results for the full hierarchical logistic regression with VFOA and Time Delay

Predictor	$\exp(B)$	t	df	p
$Max(P_X)$	1.371	3.787	1	$p > 0.05$
$Num_{min}(P_X)$	0.792	3.122	1	$p > 0.05$
$Stdev(P_Z)$	1.906	7.426	1	$p < 0.01^{**}$
TD	1.817	23.144	1	$p < 0.001^{***}$

Note: $N = 409$; $\chi^2(4) = 61.479$; $p < 0.001$; $R^2 = 0.14$

In Table 7, the model combined four predictors indicated that only $Stdev(P_Z)$, and TD , had significant relationship to cheating, while $Max(P_X)$ and $Num_{min}(P_X)$ did not. Therefore, in our final model, $Stdev(S_Z)$ and TD were chosen as predictors for logistic regression showing significant prediction ($\chi^2(2) = 56.997$; $p < 0.001$). Table 8 shows the results of the final model.

Table 8

Results for the final model in logistic regression with VFOA and Time Delay

<i>Predictor</i>	<i>exp(B)</i>	<i>t</i>	<i>df</i>	<i>p</i>
<i>Stdev(P_Z)</i>	2.005	9.530	1	<i>p</i> < 0.01**
<i>TD</i>	1.902	28.971	1	<i>p</i> < 0.001***

$N = 409$; $\chi^2(2) = 56.997$; $p < 0.001$; $R^2 = 0.130$

To evaluate our model, the data were divided into 40 folds, one fold per test taker, and we performed a leave-one-out cross-validation. A model was fit by using data from 39 test takers and tested on the remaining test taker. The 40-fold cross-validation of logistic regression and Support Vector Machine (SVM) was conducted by the library of LibLinearR (Fan, Chang, Hsieh, Wang, & Lin, 2008) and LibSVM (Chang & Lin, 2011) in R. Accuracy is calculated by the equation (1). Precision is calculated by the formula(2). The results of two classifiers show in Table 9.

$$accuracy = \frac{\text{number of true positive} + \text{number of true negative}}{\text{number of true positive} + \text{false positive} + \text{false negative} + \text{true negative}} \quad (1)$$

$$Precision = \frac{\text{number of true positive}}{\text{number of true positive} + \text{false positive}} \quad (2)$$

Table 9

The results of Logistic Regression Model and SVM

<i>Classifier</i>	<i>Accuracy</i>	<i>Precision</i>	<i>False Positive Rate</i>	<i>AUC</i>
<i>Logistic</i>	70.31%	67.35%	10.84%	75.30%
<i>SVM</i>	70.24%	68.38%	8.00%	73.79%

Discussion

The final model's results showed that the amount of variation of head movement relative to the monitor (*Stdev(P_Z)*) and time delay were both significantly positive predictors of cheating behaviors during online exams. However, time delay was the stronger of the two predictors. The logit model with *Stdev(P_Z)* and time delay explained 13% of the variation for cheating behaviors ($N = 409$; $\chi^2(2) = 56.997$; $p < 0.001$; $R^2 = 0.130$). Additionally, the proposed model was evaluated by 40-fold cross-validation. During the validation, training was performed on data from

39 test takers and tested on the remaining participant. This had an average accuracy of 70.24% with precision 68.38% and false positive rate 8.00% in SVM.

Based on the current findings, the null hypothesis that cheating has no significant relationship to time delay and VFOA was rejected. The final logistic model (Table 8) provided evidence supporting the study's predictions that head position and the amount of time spent on a problem can predict cheating behavior. Specifically, it was found that changes of VFOA, specifically the test taker's movement relative to the distance from monitor ($Stdev(P_z)$), had a significant positive relationship to cheating behaviors. Similarly, time delay had a positive relationship with cheating behavior. These two significant predictors explained 13% of the variation. In support of the third prediction, the 40-fold cross validation results showing that the logistic regression had highest average classification accuracy was 70.31% with 75.30% average area under ROC curve provided evidence for the third prediction that cheating behaviors can accurately be detected with an algorithm. The classification of SVM provided lower false positive rate and higher precision with similar accuracy and AUC, compared to logistic regression model (see Table 9).

This provides a major contribution by providing a methodology for automatically flagging suspicious behaviors in online exams that can potentially be implemented in real time. The standard method of proctoring with human surveillance is extremely resource intensive. The proposed logit model based on time delay and $Stdev(P_z)$ provided an objectively quantitative measurement which can be easily implemented and coped with current online proctoring systems.

One of limitation in this study was the false alarm rate. The current average false alarm rate in cross-validation was 10.84% and 8% in logistic regression and SVM respectively. Because of this false alarm rate, this algorithm should only be viewed as one potential resource to detecting online cheating behaviors. It is suggested, however, that this algorithm would be best implemented as a filter for human review. The proposed work only indicated that time delay and VFOA in the Z direction were significant factors which could build a semi-automatic proctoring system with human in a loop. Proctors should combine more evidence, such as checking the recorded videos and see if test takers actually access forbidden resources. However, this

combined approach would still be an improvement over previous current approaches that use all human proctoring.

The current study implemented $Stdev(P_z)$ and time delay as predictors for cheating behaviors in online testing. However, it is possible that these are not the only factors useful for online cheating detection. Chuang et al. (2015) also found that self-reporting uncertainty levels during exams could have a significant relationship to cheating. In this study, test taker's uncertainty rating was collected through an interview process after an online exam. Nevertheless, the significant finding for uncertainty ratings might suffer from the same problems as the earlier self-reported cheating studies. It relied on a self-report by the test taker at the time of the test. It is highly possible that students will not want to provide this rating or will not provide this rating reliably during testing situation. For this to be a viable method, test taker's certainty rating of each question must be assessed more objectively. Since current proctoring system can record test takers behaviors during online exams, including facial expressions, it is possible that a test taker's certainty rating of each question can be assessed more objectively based on affective states, for example, confusion (Craig et al., 2008). Craig et al. (2008) found that the physical exhibition of confusion has a significant relationship to observable human Facial Action Units (Ekman & Friesen, 1978), AU4 and AU 7. In the next chapter, the research in confusion related to cheating behavior will be discussed.

CHAPTER 6

ESTIMATING CHEATING INCIDENTS BASED ON NON-VERBAL CUES

Previous chapter investigates the variables of Time Delay, Head Pose, and Self-reported Certainty as predictors to cheating behaviors. Since current proctoring systems can record test takers behaviors during online exams, including facial expressions, it is possible that a test taker's certainty rating of each question can be assessed more objectively based on affective states, for example, confusion (Craig, D'Mello, Witherspoon, & Graesser, 2008). This chapter focuses on investigating whether the confusion, can reliably predict cheating. Finally, a model combined three non-verbal cues, time delay, head pose and confusion are formulated and discussed.

Current Study

The current study investigates whether non-verbal cues can predict cheating during online testing situations. The goal of the study is to find the relationship among confusion, time delay and VFOA related to cheating. Based on previous work, it is hypothesized that the factors of time delay (Chaung et al, 2015), the variation of head pose relative to computer screen (see CHAPTER 5), and confusion (Chaung et al, 2015) can predict cheating behaviors. Based on this hypothesis, the prediction is that the combination of the three proposed factors will be a significantly related to cheating behavior.

Data Treatment

One of the non-verbal cues in this study is the affective state, confusion, during online testing. Craig et al. (2008) found that the physical exhibition of confusion has a significant relationship to observable human Facial Action Units (Ekman & Friesen, 1978), AU4 and AU7. AU4 indicates eyebrow lower while AU7 represents lid tightened. An example of confusion represents AU4 and AU7 shows in Figure 7.



Figure 7. An example of the affective state of confusion. In the pilot study, the participant showed confusion with AU4 (eyebrow lower) on his face while answering a question.

In this study, the amount of confusion was manually coded by two judges, who were trained by a certified FACS researcher. After the training, an initial testing from the pilot datasets was conducted to validate the quality of affect coding, specifically in confusion. The training process continued until the FACS trained researcher was confident that the two judges coded confusion systematically.

The initial coding process was conducted from 10% of datasets. The test of reliability of the initial coding showed an internal agreement of Kappa value 0.9. Kappa is a measurement to test the reliability between two judges. Kappa values can range from positive one to negative one, where 1 is perfect agreement, -1 is perfect disagreement, and 0 is agreement by chances. The Kappa statistic is defined as in equation (3). Two judges went through the rest of the data with disagreements determined by discussion.

$$Kappa = \frac{\Pr(Obsered Agreement) - \Pr(Agreement by Chance)}{1 - \Pr(Agreement by Chance)} \quad (3)$$

Results

A hierarchical logistic regression was conducted using the amount of confusion coded in a question ($Num(Conf)$), time delay (TD), the variation of the head pose position relative to the screen $Stdev(P_z)$, and the interaction term of confusion and time delay ($Num(Conf) \times TD$) as predictor variables to predict the student's cheating behavior (criterion variable). The initial model had one predictor, $Num(Conf)$, showing a significant relationship to cheating ($\chi^2(1) = 10.657, p < 0.01$). The results of the initial model showed in Table 10.

Table 10.

Results for the initial model by using the amount of confusion as the predictor

<i>Predictor</i>	<i>exp(B)</i>	<i>t</i>	<i>df</i>	<i>p</i>
$Num(Conf)$	1.400	9.722	1	$p < 0.01^{**}$

Note: $\chi^2(1) = 10.657, p < 0.01, R^2 = 0.026$

The second model added TD and $Stdev(P_z)$ as predictors. The result showed that the two added predictors had significance related to cheating by themselves ($\chi^2(2) = 47.813, p < 0.001$). In the second model, both TD and $Stdev(P_z)$ had significant predictive power to cheating but $Num(Conf)$ did not. The details of the second model showed in Table 11.

The third block added interaction term between amount of confusion and time delay ($Num(Conf) \times TD$). The results showed that the interaction had marginal significance related to cheating ($\chi^2(1) = 3.575, p = 0.59 > 0.05$). The final results of hierarchical logistic regression showed that the four predictors can significantly predict cheating ($\chi^2(4) = 62.045, p < 0.001$), listed in Table 12.

Table 11.

Results for the second model by adding time delay and the variation of head pose relative to the screen

Predictor	$\exp(B)$	t	df	p
$Num(Conf)$	1.400	1.422	1	$p > 0.05$
TD	1.846	25.098	1	$p < 0.001^{***}$
$Stdev(P_z)$	1.928	8.461	1	$p < 0.01^{**}$

Note: $\chi^2(3) = 58.470, p < 0.001, R^2 = 0.133$

Table 12.

Results for the third model by adding the interaction term of confusion and time delay

Predictor	$\exp(B)$	t	df	p
$Num(Conf)$	1.308	4.150	1	$p < 0.05^*$
TD	1.956	28.079	1	$p < 0.001^{***}$
$Stdev(P_z)$	1.788	6.516	1	$p < 0.05^*$
$Num(Conf) \times TD$	0.862	3.939	1	$p < 0.05^*$

Note: $\chi^2(4) = 62.045, p < 0.001, R^2 = 0.141$

The dependent variable, *Cheating*, has a binary response with the value 1 if the individual cheated while answering the question, and 0 otherwise. The logistic regression of cheating estimation is shown in (4).

$$\ln(Cheating) = -0.643 + 0.268 \times Num(Conf) + 0.671 \times TD + 0.581 \\ \times Stdev(P_z) - 0.148 \times Num(Conf) \times TD \quad (4)$$

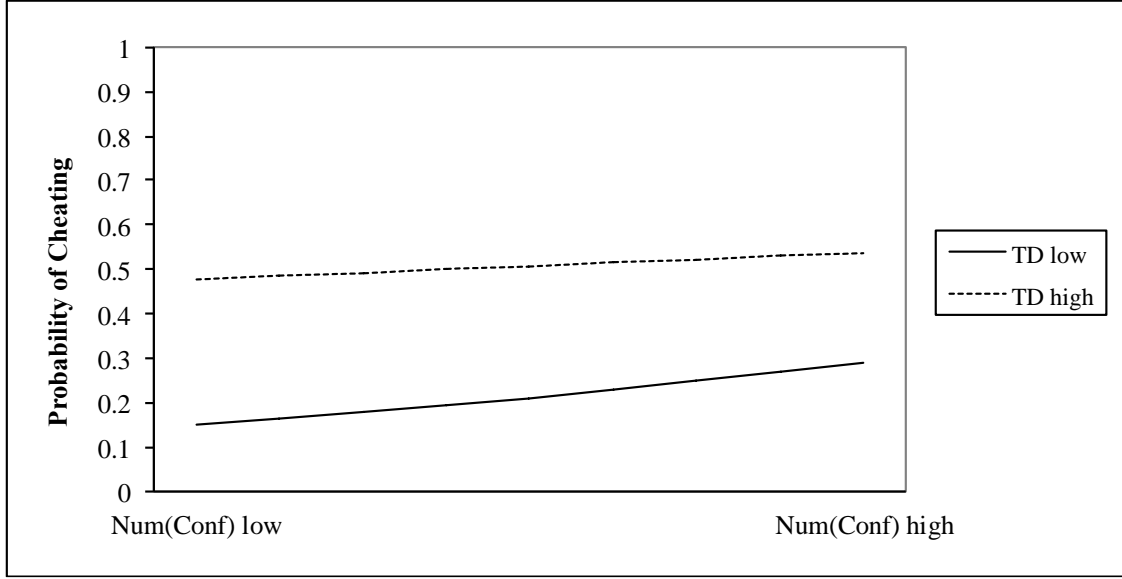


Figure 8. The interaction term of time delay and the amount of confusion. The solid line has the moderator, time delay, with low value, -1. The dash line has the moderator, time delay, with high value, 1. The value of time delay and confusion were whitened in pre-processing (zero mean with standard deviation one).

The plot of the interaction is show in Figure 8. The probability of cheating in logistic regression is calculated by equation (5).

$$Pr(Cheating) = \frac{e^{-0.643+0.268 \times Num(Conf)+0.671 \times TD+0.581 \times Stdev(P_z)-0.148 \times Num(Conf) \times TD}}{1+e^{-0.643+0.268 \times Num(Conf)+0.671 \times TD+0.581 \times Stdev(P_z)-0.148 \times Num(Conf) \times TD}} \quad (5)$$

In the equation (5), the value of $Stdev(P_z)$ is 0, and the values of $Num(Conf)$ range from -1 to 1 with 0.25 intervals. The solid line indicates a time delay value of -1 and the dash line indicates a time delay value of 1. The positive coefficients ($Num(conf)$, TD , and $Stdev(P_z)$); see equation (4)) for the final logistic regression model indicated that increases in the individual variables were associated with more cheating. The negative interaction coefficient between confusion and time delay means that confusion has a stronger effect for detecting cheating when a test taker is faster to answers a question (see Figure 8). However, the confusion becomes less predictive as the amount of time to answer the question increases. This complex interaction effect showed that the model might not be straightforward and linear. Therefore, a non-linear classification, Support Vector Machine with Radial Basis Function (RBF) kernel, was used to validate the accuracy compared to the logit model.

To evaluate our model, the data were divided into 40 folds, one fold per test taker, and we performed a leave-one-out cross-validation. It means that a model is fitted by using data from 39 test takers and tested on the remaining test taker. The 40-fold cross-validation of logistic regression and Support Vector Machine (SVM) were conducted by the library of LibLinearR (Fan et al., 2008) and LibSVM (Chang & Lin, 2011) in R. The predictors in logistic regression and SVM are the amount of confusion, time delay in a question, the variation of head pose relative to the computer screen, and the interaction term between time delay and head pose. The results of two classification show in Table 13.

Table 13

The classification results of the final model with Logistic Regression and SVM

<i>Classifier</i>	<i>Accuracy</i>	<i>Precision</i>	<i>False Positive Rate</i>	<i>AUC</i>
<i>Logistic Regression</i>	68.11%	62.67%	14.04%	74.33%
<i>SVM</i>	70.70%	72.10%	7.1%.	74.40%

Discussion

This study investigated test takers' behaviors during online exams and found that time delay to answer the question, the variation of students' head poses relative to the computer screen, and confusion, had positive significant relationship for predicting cheating behaviors (see Figure 9). The final model of the hierarchical logistic regression (Table 12) showed that the amount of confusion, time delay, and the variation of head pose relative to the screen were significantly associated with cheating behavior. The significantly positive relationship among $Num(Conf)$, TD , and $Stdev(Z)$ to cheating matches the expectation that given the opportunity to cheat, cheaters show more confusion on their faces, spend more time in consulting forbidden resources, and had more variation in their head poses, compared to non-cheaters. The results of confusion and cheating also validated the significant relation between students' uncertainty ratings and academic dishonesty found by Chuang et al. (2015).

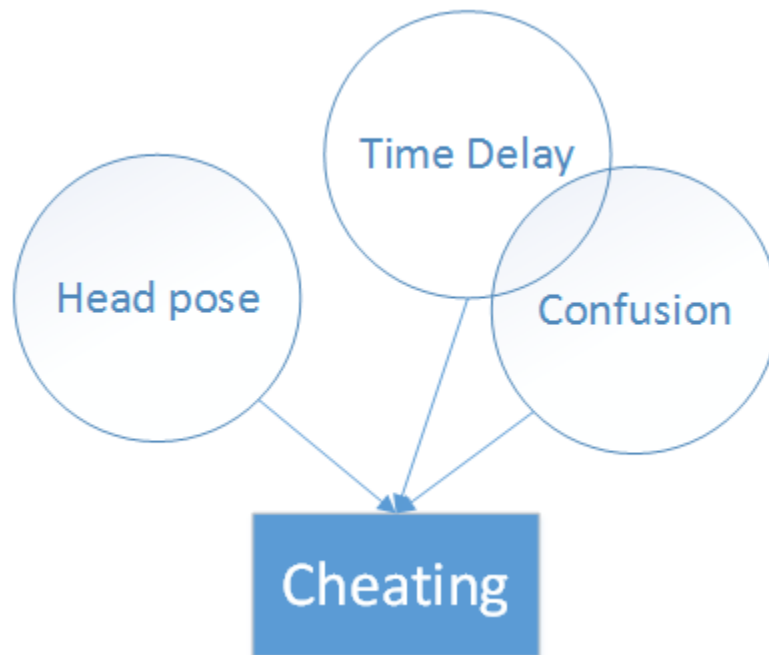


Figure 9. The three proposed factors significantly associated with cheating. All three factors are positively related to cheating behaviors. The most relevant and non-redundant features in the head pose pair with time delay and confusion the head position relative to the monitor. In addition, time delay and the amount of confusion have an interaction term. The amount of confusion has stronger effect when time delay is shorter.

Based on Figure 8, the participant's data indicates that when a test taker is answering quickly as the amount of confusion increased, they are more likely to cheat. However, it is moderated by time delay to answer the question. When there is a significant time delay presented, the amount of confusion has weaker effect to cheating decision. The interaction term between time delay and the amount of confusion indicated that the relationship between cheating behavior and the four predictors in the final model (see Table 12.) was not a linear model. Therefore, in Table 13, the non-linear SVM classification with Radial Basis Function (RBF) kernel was used. It showed higher accuracy and almost half of false positive rates of cheating detection compared to linear logistic regression model.

Previous research showed that the self-reported uncertainty ratings had a significant positive relation to academic dishonesty (Chuang et al., 2015). Based on the strong relationship between the physical exhibition of confusion found by Craig et al. (2004), the internally mental status of uncertainty ratings were assumed to be measured more objectively without relying on

self-reported surveys. For this more objective measure, instead of using a self-report questionnaire survey, confusion was coded based on human observations from the recorded videos. The results of confusion revealed a statistically significant relationship with online cheating, especially while test takers answered a question quickly. This significant finding in this study demonstrates that the measurement of confusion can be not only conducted by the third party objectively but also be implemented paired with currently proctoring systems. Additionally, current software, Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011) demonstrates an ability of measuring a wide array of facial action units in real-time, including AU 4 and AU7. It shows a possibility of automatic analyses of confusion paired with online proctoring systems.

CHAPTER 7

GENERAL DISCUSSION

Based on the results in Table 2, Table 8, and Table 12, the statement of null hypothesis, $H1_0$, where a test taker's time delay on each question would have no statistically significance related to cheating decisions, was rejected. All results showed that time delay had the positively strongest relationship to cheating behavior among all predictors in the three logistic regression models. The results supported the prediction that the longer time a student spent on a question there was a higher likelihood that he or she cheated in the question.

The null hypothesis of $H2_0$ states that a test taker's certainty rating on each question would have no statistically significant relation to cheating decisions during online exams, was rejected. In the Table 2, the students' certainty ratings showed a significantly negative relation to cheating behavior ($\exp(B) = 0.757$; $t(1) = 7.83$; $p < 0.01$). The results indicated that the more uncertainty students felt while answering the question, the higher likelihood that they cheated in the question.

The null hypothesis $H3_0$ states that VFOA deviated from the computer screen would have no statistically significant relation to cheating behaviors during online exams. With an approximation of VFOA by using statistical features of head poses, the hypothesis $H3_0$ was rejected based on the results in Table 7, Table 8, and Table 12. Based on the results in Table 7, and Table 8, the most significant feature of VFOA among 48 different features (see Table 4) the variation of head position relative to the computer screen ($Stdev(P_z)$) has the highest statistical significance related to cheating behavior. This indicated that the more variation of head movement in back-and-forth relative to the monitor, there was a higher likelihood that students cheated in the question. In Table 8, the logistic regression model with $Stdev(P_z)$ and time delay explained 13% of the variation for cheating behaviors ($N = 409$; $\chi^2(2) = 56.997$; $p < 0.001$; $R^2 = 0.130$). The classification of SVM provided lower false positive rates (8%) and higher precision (68.38%) with similar accuracy (70.24%) and AUC (73.79%), compared to logistic regression model (see Table 9).

Based on the results of Table 12, the null hypothesis, $H4_0$, where facial expressions associated with confusion would have no statistically significant relation to cheating behaviors during online exams, was rejected. The positive coefficients of $Num(conf)$ in equation (4) for the final logistic regression model indicated that increases in the amount of exhibited confusion was associated with more cheating. The negative interaction coefficient between confusion and time delay means that confusion has a stronger effect for detecting cheating when a test taker is faster to answers a question (see Figure 8). The final model of SVM (see Table 13) showed the best classification accuracy, 70.7%, with 72.1% precision and 7.1% false alarm rate. The SVM model provided an acceptable capability to detect cheating, showing a AUC value 74.40%, which is greater than 0.7 (Hosmer Jr & Lemeshow, 2004).

The observed evidence in support of hypothesis 2 and hypothesis 4, where uncertainty and confusion had significant relation to academic dishonesty, provides alternative supports for the proposed explanations of the cognitive process while cheating in online exams. When a test taker encounters a question that he or she cannot answer or have difficulty answering, an impasse occurs, similar to the learning process (VanLehn et al., 2003). Therefore, the cognitive system is moved into disequilibrium, meaning that individuals are confronted with problems (Graesser et al., 2005). While individuals' cognitive systems turn from equilibrium to disequilibrium because of hitting impasses, a physical exhibition, confusion, shown upon their faces. At that movement, students, therefore, want to move out of the disequilibrium state and back into a normal state of equilibrium.

During the testing, there are no learning opportunities remaining to work pass the impasse. If the student is sufficiently motivated to provide a correct answer and resources are readily available with limited monitoring within the online setting, students are more likely to cheat. It is because cheating is the only way to resolve the impasse and move forward. Because cheat sheets, notes, and other forbidden resources such as cell phones or printed textbooks are the only resources during online exams, students will move back and forth in order to search for the answer to the question. These physical cheating behaviors lead to not only the larger variation of

head pose movements relative to the monitor, but also longer time to answer a question compared to the honest students.

Hypothesis 1 was based on the significant factor of response time (RTs) found by Van der Linden and Guo (2008). They conducted a simulation study and successfully detected two cheating behaviors based on RTs: (1) pre-knowledge of some of the items; (2) attempts to take tests only for the purpose of memorizing the items. Inspired by their research results, in this dissertation, a significant relationship between time delay and cheating behavior was demonstrated based on Table 2, Table 8, and Table 12. There were two differences between the current study and Van der Linden's and Guo's research (2008). The first one was that the current study was a lab-based experiment, rather than a computer simulation (van der Linden & Guo, 2008), where cheating incidents were not established by real human's cheating behavior. Second, the focused cheating behavior in this dissertation was the misuse of the forbidden resources, not pre-knowledge or memorization of test items (van der Linden & Guo, 2008). The investigated cheating behavior was the dishonest behavior that current human proctors monitored remotely during online exams (Frank, 2010).

Hypothesis 2 was based on the theory of impasse during learning (VanLehn et al., 2003). The current findings revealed that the high uncertainty ratings caused by hitting an impasse during testing had significant relationship to cheating (see Table 2), similar to the self-reported uncertainty related to learning found by McQuiggan et al. (2010).

Hypothesis 3 was based on a report from employees who worked at the remotely proctoring company, ProctorU (2014), saying that the observable patterns of behaviors for normal people versus the people who tried to sneak in a cell phone and looked up information were clear, based on a report in the New York Times by Eisenberg (2013). The rejection of hypothesis $H3_0$ matches the report from employee who worked at proctoring company. Additionally, this study demonstrated that the most suspicious behavior was the head pose movement relative to the computer screen. Moreover, this feature could be automatically extracted by the recorded videos with CLM framework (Baltrusaitis et al., 2012). The final SVM with time delay and head pose

could give an average accuracy 70.24%, average AUC 73.79%, average false positive rates (8%) and average precision (68.38%) in 40-folds cross validation.

Hypothesis 4 was based on the importance of uncertainty associated with academic dishonesty found by Chuang et al. (2015) and the affective detection of confusion found by Craig et al. (2008). Chuang et al. (2015) found that the self-reported uncertainty had significantly positive relationship to cheating behavior and Craig et al. (2008) discovered that the mental state of uncertainty raised the physical exhibition of confusion, which could be detected based on the FACS of AU4 and AU7 (Ekman & Friesen, 1978). In this study, the confusion coded by AU4 and AU7 showed a positive association with cheating behavior. Additionally, confusion had stronger effect to indicating cheating in a shorter answering time.

Contributions

There are three major contributions from this dissertation. The first one is the introduction of a novel method to collect cheating incidents. The second one is proposing an innovative method of detection of academic dishonesty based on non-verbal cues recorded during online testing. The third one is validating that uncertainty and confusion during testing and demonstrating its significance to cheating behavior.

In previous research in academic dishonesty, most of the cheating incidents were based on participant's self-reports of potential intentions to cheat, not the actual behaviors (Crown & Spiller, 1998; Murdock et al., 2001; Şendağ, Duran, & Fraser, 2012; Smith et al., 2003). In this study, the drawback of the bias due to the use of a survey technique was overcome because cheating incidents were established by retrospective reports provided directly after online testing, rather than questionnaire survey.

It was found that a key process for induced cheating incidents within this study was making sure that test takers had an adequate understanding in the learning materials. In the pilot study, subjects were asked to watch the video lectures and to create the cheating sheets by themselves. After that, they took an online exam without any tutoring in between. It was found that most participants did not care about the online testing and guessed an answer even though forbidden resources were provided. In order to make sure that participants took the exam

seriously, experimenter went through the learning materials with test takers personally. Since test takers understood the learning materials and knew where to address the answer, they started cheating instead of guessing while hitting a question they have a problem to answer. Thus, it is recommended that future researchers should ensure that participants have adequate knowledge of the materials using processes like tutor to ensure there is a strong enough fidelity of implementation in the experimental design to induce cheating in the testing session.

The importance of tutoring during learning session can be explained by the transition of equilibrium to disequilibrium of the cognitive systems. Test takers cognitive systems are in equilibrium if they have high confidence of comprehension in the learning materials. Therefore, if they take an exam and hit an impasse, their cognitive systems are in disequilibrium, which therefore induce them to cheat to go through the impasse. Conversely, if test takers' do not have well comprehension in learning materials before taking an online exams, they will randomly guess an answer. It is because there is no transition of the cognitive systems from equilibrium to disequilibrium. As a result, they do not feel a need to go through the impasse by cheating in an exam.

Additionally, over the past few decades, the factors explored in the academic dishonesty research were primarily personal/situational constructs based on surveys of cheating intentions (D. F. Crown & Spiller, 1998; Murdock et al., 2001; Şendağ et al., 2012; Smith et al., 2003) However, the applications of previous findings are problematic. For example, in order to retrieve test taker's personal/situational constructs, it seems unlikely that educational institutions would conduct a mass profiling survey before an exam. The use of personal/situational factors and questionnaire survey to predict academic dishonesty, were diminished in the current study. Instead of using personal/situational factors, the detection of online academic dishonesty was based on the three dynamic features of test taker's behaviors: time delay, the changes of head poses relative to the monitor, and confusion exhibited on students' faces, all of which can be extracted during online testing.

Moreover, the fidelity of transferring the theory of an impasse from learning settings to testing settings is validated in this dissertation. It is well known that an impasse and confusion

are a key factor related to learning (Craig et al., 2004; Mello & Graesser, 2011; VanLehn et al., 2003). However, it was still unknown whether an impasse holds the same relationship to cheating intentions as to learning performance. In this study, the significant relationship between an impasse and cheating behavior has been validated not only by students' uncertainty ratings but also by the amount of confusion exhibited on test takers' faces. Therefore, similar to the use of confusion to support online learning (Craig et al., 2008; Mello & Graesser, 2011), the affect state, confusion, can also be used as an indication to cheating. For example, it is possible that the detection of confusion can help proctors intervene probable dishonest academic behavior in advance during remotely administrated exams.

Limitations and Future Research

There are several limitation and open questions left in this research. The first one is the relationship among time delay, certainty, and cheating. Theoretically, time delay and certainty are both influenced by the process of cheating. However, low certainty could be a prerequisite for cheating and time delay could be a consequence of the act of searching for the answer in the materials.

The second one is the validity of the research in the real world. It is possible that the research could not have the fidelity to transfer from the laboratory setting into a real world setting. Therefore, replication in the real world setting would be beneficial for understanding the generalizability of the finding. Moreover, based on the results in CHAPTER 4, CHAPTER 5, and CHAPTER 6, the most significant predictor is time delay on the question. However, in order to obtain an abnormal time delay that may indicate a suspicious behavior, such as inappropriately referencing materials, a supervised dataset should be trained ahead. It means that if a test bank is changed, the original model may not be valid. There is a need to improve the proposed model for more general usages.

The third one is that there are potential factors other than time delay, head pose, and confusion useful for detecting cheating behaviors during online exams. It is not recommended that the classification criteria of cheating behaviors are just based on the three proposed non-verbal cues. Proctors should combine more evidence, such as checking the recorded videos and

see if test takers actually access forbidden resources. The propose work only indicated that time delay and certainty rate are significant factors which may help proctors to improve the proctoring process in remotely administrated exams.

The fourth limitation is the assumption of the identical difficulty of a given question to all students. This simplification is questionable and should be improved as the future work. Given a question, there should be different response times for different students since each student has distinct problem solving ability. A better approach to model the response times can be a lognormal model proposed by van der Linden (2006). In his model, the response time was in lognormal distribution with mean distribution, $\beta_i - \tau$. The parameter, β_i , means time intensity, and it controls the time item i demands from the person. The larger β_i indicates that the larger the amount of time a person tends to spend on it. The parameter, τ , means the speed of the person. The larger τ indicates that the smaller the amount of time the person tends to spend on the item. As a result, the time delay on each question is a function of personal abilities and question difficulties.

The fifth limitation is the assumption of independent distribution of cheating probability for each question among the whole testing process. It is reasonable that the probability of cheating in each question is dependent to each other. If a test taker cheated in one question, it is reasonable to estimate a higher probability to cheat in other questions in the future. The model can be improved by Bayesian procedures proposed by van der Linden and Guo (2008). In their model, the test materials are adaptive exams, meaning that the probability of a correct response is modeled by the test taker's ability, the difficulty of the test item, the discrimination of the test item, and guessing probability. Given an adaptive test, the appropriate response time of each item is modeled by the test taker's ability and the observed response times on the current item by the test taker. The suspicious time delay patterns in the model could be: (1) an answer that is correct; (2) a large positive residual of time delay; (3) a high probability of success on the item (i.e., a high estimated ability relative to the difficulty of the item). Therefore, each item has different probability to elicit cheating for different participants based on test takers' abilities. Additionally, the

distribution of cheating probability is based on the previous testing item whether it indicates an aberrant response time.

Finally, proctoring has been shown to not only deter cheating in online assessments but also enhance learning performance in online courses. Wellman (2005) showed that online-module delivery paired with proctored quizzes was more effective in promoting learning when compared to un-proctored quizzes. The proctored group practiced more frequently than the un-proctored group, especially students in the bottom half of performers. In spite of the benefits, it can be impractical to supervise all quizzes in large online courses. Typically only high-stakes exams, such as midterms or final exams, are under surveillance (Luecht, 2006). The standard methods of proctoring and human surveillance are extremely resource intensive. This current work provides the first steps toward potential methods to automatically detect cheating during online assessments.

In the current study (see Chapter 6), the detection of confusion through manual coding, however, is not only human-labor intensive, but also time consuming. Instead, current software, Computer Expression Recognition Toolbox (CERT) (Littlewort et al., 2011) demonstrates an ability of measuring a wide array of facial action units in real-time and enables a possibility of automatic analyses of facial expressions. Grafsgaard et al. (2013) validated the performance of CERT among 650,000 frames of recorded tutoring videos, and showed that the average accuracy of AU4 and AU7 were 85% and 100% respectively. The successful detection rates in AU4 and AU7 point out a future direction for detection of uncertainty automatically.

Conclusion

This study investigated test takers' behaviors during online exams and found that time delay to answer the question, the variation of students' head poses relative to the computer screen, and confusion, had positive significant relationship for predicting cheating behaviors. Additionally, by use of time delay, $Stdev(P_Z)$, confusion, and the interaction term of confusion and time delay as predictors, a SVM was formulated. It was validated by leave-one-out cross-validation (40 folds), showing an average accuracy of 70.7% with 7.1% false alarm rate. The results of this current algorithm provide the possibility of building a proctoring system that could

flag suspicious students in remotely administered exams automatically. Additionally, it shows that confusion can be the first indicator to indicate when a proctor should intervene to prevent student's academic dishonesty during online exams. However, it should be noted that the proposed work does not indicate whether a student actually cheated. Instead, it shows whether a student is behaving in a manner consistent with cheating. If the observed effects hold, the current algorithm could reliably rule out a large portion of exam recordings as inconsistent with cheating using an automated method. This would thereby reduce the expected time to discover infractions on recorded exams for human proctors and provide a step toward solving the significantly large problem of student cheating during online exams.

REFERENCES

- Adkins, S. S. (2011). *The US Corporate Market for Self-Paced eLearning Products and Services: 2010-2015 Forecast*. Ambient Insight. Retrieved from <http://www.ambientinsight.com/Resources/Documents/Ambient-Insight-2010-2015-US-Corporate-eLearning-Market-Executive-Overview.pdf>
- Baltrusaitis, T., Robinson, P., & Morency, L. (2012). 3D constrained local model for rigid and non-rigid facial tracking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2610–2617). IEEE.
- Birenbaum, M. (2007). Assessment and instruction preferences and their relationship with test anxiety and learning strategies. *Higher Education*, 53(6), 749–768.
- Bronzaft, A. L., Stuart, I. R., & Blum, B. (1973). Test anxiety and cheating on college examinations. *Psychological Reports*, 32(1), 149–150.
- Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 27.
- Chapman, K., Davis, R., Toy, D., & Wright, L. (2004). Academic integrity in the business school environment: I'll get by with a little help from my friends, 26(3), 236–249.
- Chuang, C. Y., Craig, S. D., & Femiani, J. C. (2015). The Role of Certainty and Time Delay in Students' Cheating Decisions during Online Testing. In *37th Annual Cognitive Science Society*.
- Craig, S. D., Graesser, A., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3), 241–250.
- Craig, S. D., Mello, S. D', Witherspoon, A., & Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive–affective states during learning. *Cognition and Emotion*, 22(5), 777–788.
- Crown, D. F., & Spiller, M. S. (1998). Learning from the literature on collegiate cheating: A review of empirical research. *Journal of Business Ethics*, 17(6), 683–700.
- Crown, D., & Spiller, M. (1998). *Learning from the literature on collegiate cheating: A review of empirical research*. Journal of Business Ethics.
- Doherty-Sneddon, G., & Phelps, F. G. (2005). Gaze aversion: a response to cognitive or social difficulty? *Memory & Cognition*, 33(4), 727–733.
- Eisenberg, A. (2013, March 2). Keeping an Eye on Online Test-Takers. Retrieved from http://www.nytimes.com/2013/03/03/technology/new-technologies-aim-to-foil-online-course-cheating.html?_r=3&
- Ekman, P., & Friesen, W. V. (1978). *The facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.
- Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. *The Journal of Machine Learning Research*, 9, 1871–1874.

- Frank, A. (2010). Dependable Distributed Testing – Can the Online Proctor be Reliably Computerized? In *Proceedings of International Conference on e-Business (ICE-B 2010)* (pp. 22–31).
- Graesser, A. C., Lu, S., Olde, B. A., Cooper-Pye, E., & Whitten, S. (2005). Question asking and eye tracking during cognitive disequilibrium: Comprehending illustrated texts on devices when the devices break down. *Memory & Cognition*, 33(7), 1235–1247.
- Grafsgaard, J., Wiggins, J. B., Boyer, K. E., Wiebe, E. N., & Lester, J. (2013). Automatically recognizing facial expression: Predicting engagement and frustration. In *Educational Data Mining 2013*.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.
- Harmon, O. R., & Lambrinos, J. (2008). Are online exams an invitation to cheat? *The Journal of Economic Education*, 39(2), 116–125.
- Harmon, O. R., Lambrinos, J., & Buffolino, J. (2010). Assessment design and cheating risk in online instruction. *Online Journal of Distance Learning Administration*, 13(3).
- Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, 58(1), 47–77.
- Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.
- Izard, C. E. (1971). *The face of emotion*. East Norwalk, CT, US: Appleton-Century-Crofts.
- Izard, C. E. (1994). Innate and universal facial expressions: evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115, 288–299.
- Kennedy, K., Nowak, S., Raghuraman, R., Thomas, J., & Davis, S. (2000). Academic dishonesty and distance learning: Student and faculty views. *College Student Journal*, 34(2), 309–314.
- King, C., Guyette, R., & Piotrowski, C. (2009). *Online exams and cheating: An empirical analysis of business students' views*. The Journal of Educators Online.
- Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence* (pp. 115–141). Springer.
- Kryterion. (2014). Kryterion. Retrieved from <https://www.kryteriononline.com/>
- Lanier, M. M. (2006). Academic Integrity and Distance Learning*. *Journal of Criminal Justice Education*, 17(2), 244–261.
- Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011). The computer expression recognition toolbox (CERT). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on* (pp. 298–305). IEEE.
- Luecht, R. M. (2006). Operational issues in computer-based testing. *Computer-Based Testing and the Internet: Issues and Advances*, 91–114.

- McQuiggan, S. W., Robison, J. L., & Lester, J. C. (2010). Affective transitions in narrative-centered learning environments. *Journal of Educational Technology & Society*, 13(1), 40–53.
- Meijer, R. R., & Sijsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, 25(2), 107–135.
- Mello, S. D', & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, 25(7), 1299–1308.
- Mello, S. D', & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2), 145–157.
- Mello, S. D', Lehman, B., Pekrun, R., & Graesser, A. (2014). Confusion can be beneficial for learning. *Learning and Instruction*, 29, 153–170.
- Murdock, T. B., Hale, N. M., & Weber, M. J. (2001). Predictors of cheating among early adolescents: Academic and social motivations. *Contemporary Educational Psychology*, 26(1), 96–115.
- Nagin, D. S., & Pogarsky, G. (2003). An Experimental Investigation of Deterrence: Cheating, Self-Serving Bias, and Impulsivity. *Criminology*, 41, 167–194.
- National Science Board. (2012). *Science And Engineering Indicators*. National Science Foundation.
- Nelson, T., & Schaefer, N. (1986). Cheating among college students estimated with the randomized-response technique, 20(3), 321–325.
- Otero, J., & Graesser, A. C. (2001). PREG: Elements of a model of question asking. *Cognition and Instruction*, 19(2), 143–175.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37(2), 91–105.
- Petridou, A., & Williams, J. (2007). Accounting for aberrant test response patterns using multilevel models. *Journal of Educational Measurement*, 44(3), 227–247.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1988). Numerical recipes in C. *Cambridge University Press*, 1, 3.
- Prince, D., Fulton, R., & Garsombke, T. (2009). Comparisons Of Proctored Versus Non-Proctored Testing Strategies In Graduate Distance Education Curriculum. *Journal of College Teaching & Learning (TLC)*, 6(7), 51–62.
- ProctorU. (2014). Pupilcity ProctorU. Retrieved from <http://www.proctoru.com/>
- Rozin, P., & Cohen, A. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3, 68–75.
- Scheers, N., & Dayton, C. M. (1987). Improved estimation of academic cheating behavior using the randomized response technique, 26(1), 61–69.

- Securexam Remote Proctor. (2014). Securexam Remote Proctor. Retrieved from <http://www.softwaresecure.com/student-login/>
- Şendağ, S., Duran, M., & Fraser, M. R. (2012). Surveying the extent of involvement in online academic dishonesty (e-dishonesty) related practices among university students and the rationale students provide: One university's experience. *Computers in Human Behavior*, 28(3), 849–860.
- Sierra, J. J., & Hyman, M. R. (2008). Ethical antecedents of cheating intentions: Evidence of mediation. *Journal of Academic Ethics*, 6(1), 51–66.
- Smith, K. J., Davy, J. A., Rosenberg, D. L., & Haight, G. T. (2003). A structural modeling investigation of the influence of demographic and attitudinal factors and in-class deterrents on cheating behavior among accounting majors. *Journal of Accounting Education*, 20(1), 45–65.
- van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2), 181–204.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365–384.
- van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics*, 37(1), 180–199.
- VanLehn, K. (1998). Analogy events: How examples are used during problem solving. *Cognitive Science*, 22(3), 347–388.
- VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W. B. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3), 209–249.
- VProctor. (2014). VProctor. Retrieved from <http://vproctor.com/>
- Watson, G., & Sottile, J. (2010). Cheating in the digital age: Do students cheat more in online courses? *Online Journal of Distance Learning Administration*, 12(4).
- Wellman, G. S. (2005). Comparing learning style to performance in on-line teaching: Impact of proctored v. un-proctored testing. *Journal of Interactive Online Learning*, 4(1), 20–39.
- Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML* (Vol. 3, pp. 856–863).

APPENDIX A
EXEMPTION FROM IRB REVIEW

Type of Review:	Initial Study
Title:	Categorize Typical and Atypical Behaviors in online Assessments
Investigator:	Scotty Craig
IRB ID:	STUDY00000369
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	<ul style="list-style-type: none"> • HRP-502c - TEMPLATE CONSENT DOCUMENT -SHORT FORM.pdf, Category: Consent Form; • IRB for Lab Experiment.docx, Category: IRB Protocol; • Questions.docx, Category: Recruitment Materials; • CheatSheet.docx, Category: Recruitment Materials;

APPENDIX B
LERANING MATERIALS

Basic Python and Variables:

- In the terminal, you can go into the Python interpreter by just type “python”.
- In Python interpreter, you can execute your code without compiling your program.
- In Python, a variable can be initiated **without defining** the type of the object. The type of an object can be also redefined by just assigning a new value.
- A Python file can be executed by typing “python filename.py” in the terminal

<pre>>>>= a + b >>> a = 1 >>> b = 2 >>> c = a + b >>> c 3 >>> a = 'hi' >>> a 'hi' >>> b = 'student' >>> b 'student'</pre>	<pre>## add variable “a” and “b” and assign result to variable “c” ## entering an expression prints its value ## a can hold a string just as well ##b can hold a string just as well</pre>
--	--

Math and Numbers:

- In Python, a number with a decimal point is called “float”
- The modulo operator ‘%’ finds the remainder of division
- The exponent operator ‘**’ is a shorthand for repeated multiplication of the same thing by itself.

Division	Modulo	Exponent
<pre>>>> 20 / 9 2 >>> 20.0 / 9.0 2.2222222222222223 >>> 20.0 / 9 2.2222222222222223 >>> 20 / 9.0 2.2222222222222223 >>> 20 / 9. 2.2222222222222223 >>> 20. / 9 2.2222222222222223</pre>	<pre>>>> 9 % 4 1 >>> 9.0 % 4 1.0 >>> 8 % 4 0</pre>	<pre>>>> 10 ** 3 1000 >>> -10 ** 3 -1000 >>> (-10) ** 2 100 >>> 10 ** -2 0.01 >>> 1 / (10 ** 2) 0.01 >>> 2 ** -3 0.125 >>> 1 / (2 ** 3) 0.125</pre>

Comparison and concatenation:

- Operator “+” can be used as an addition when the types of the variables are numbers. When the types of the variables are strings, it can be used as concatenation of strings. However, it cannot combine two different types of variables. For example, string and integer.

<pre>>>> a = 'hi' >>> b = 'student' >>> c = 3 >>> a + b 'histudent'</pre> <p>TypeError: cannot concatenate 'str' and 'int' objects</p> <pre>>>> a == c False</pre> <pre>>>> a != c True</pre>	<pre>## concatenate string "a" and string "b" ## concatenate string and integer, report error! ## compare two different types of objects ## compare two different types of objects</pre>	<table><tr><th colspan="2">Comparison</th></tr><tr><td>Equality</td><td>==</td></tr><tr><td>Not Equal</td><td>!=</td></tr><tr><td>Bigger than</td><td>></td></tr><tr><td>Smaller than</td><td><</td></tr><tr><td>Bigger than or equal</td><td>>=</td></tr><tr><td>Smaller than or equal</td><td><=</td></tr></table>	Comparison		Equality	==	Not Equal	!=	Bigger than	>	Smaller than	<	Bigger than or equal	>=	Smaller than or equal	<=
Comparison																
Equality	==															
Not Equal	!=															
Bigger than	>															
Smaller than	<															
Bigger than or equal	>=															
Smaller than or equal	<=															

Type, length function:

- Length function can return the number of characters of a string object. An integer object cannot be passed as an argument.
- Type function can return the type of an object.

Length function	Type function
<pre> >>> a = 'hi' >>> b='student' >>> c = 3 >>> len(a) 2 >>> len(b) 7 >>> len(c) TypeError: object of type 'int' has no len() </pre>	<pre> >>> a = 'hi' >>> b = 1 >>> c = 2/5 >>> type(a) <type 'str'> >>> type(b) <type 'int'> >>> type(c) <type 'int'> </pre>

Exit Python:

- You can exit a Python interpreter by typing the following command: `quit()` and `exit()`. Or you can typing an end-of-file character (Control-D).

Define a Function in Python:

- A function cannot be defined after it is called.

<pre> 1 def compare(var1, var2): 2 if var1 == var2: 3 print 'Equal.' 4 else: 5 print 'Not equal' 6 a = 1 7 b = 1 8 compare(a, b) 9 b = 2 10 c = compare(a, b) </pre>	<pre> ## Run the compare function with arguments 1 and 1 ## Run the compare function with arguments 1 and 2 </pre>
--	---

Equal.

Not equal

<pre> 1 def f1(var1, var2): 2 return var1 + var2 3 def f2(var1, var2): 4 return var1 - var2 5 6 a = 1 7 print f1(a, a+1) 8 print f2(a, a+1) 9 print f1(a, a+1) * f2(a, a+1) </pre>	<pre> ## print the f1 function with arguments 1 and 2 (a = 1) ## print the f2 function with arguments 1 and 2 (a = 1) ## print the 3 * -1 = -3 </pre>
--	---

```

3
-1
-3

```

<pre> 1 a = 3 2 b = 'hi' 3 c = a == b 4 d = c == len(b) == a 5 print d </pre>	<pre> ## c holds value False because a and b are not equal ## the result of c == len(b) is False; the result of len(b) == a is False ## d holds False since it is the intersection between two False values </pre>
---	--

False

“c” holds False value because “a” and “b” are not the same types of objects. In line 4, the value in “d” is determined by the value of `c==len(b)` and `len(b) == a`. The value of “a” is 3 with an integer type. The value of `len(b)` is 2 with an integer type. The value of “c” is False with a Boolean type. Therefore, `c == len(b)` is False since they hold different types. `len(b) == a` is False since integer 2 is not equal to integer 3. The d value can be seen as intersection between two False values and it is False.

APPENDIX C
TESTING MATERIALS

Phase B

1. What is the value in the variable "c" by typing following codes in a Python interpreter?

```
>>> a = 6  
>>> b = 7  
>>> c = a == b
```

 - a) 6
 - b) 7
 - c) True
 - d) False
 - e) Error
2. Which one of the following commands can print the value of a variable in a Python interpreter?
 - a. >>> a
 - b. >>> print a
 - c. >>> print (a)
 - d. all of the above
 - e. None of the above
3. What are the print out values if you execute the following Python codes as a Python file?

```
a = 1  
b = 2  
c = a + b  
  
print (a+b+c)
```

 - a. 1
 - b. 2
 - c. 3
 - d. 6
 - e. a+b+c
4. What will be printed by typing the following command in a Python interpreter?

```
>>> print type(30)
```

 - a. <type 'number'>
 - b. <type 'float'>
 - c. <type 'double'>
 - d. <type 'int'>
 - e. <type 'tuple'>
5. What are the print out values if you execute the following Python codes as a Python file?

```
def function(v1, v2):  
    if v1 > v2:  
        return v1  
    else:
```

```

        return v2
a = 1.5
b = 1.2
print function(a, b + 1)

```

- a. 1.2
- b. 1.5
- c. 2.2
- d. 2.5
- e. 2.7

6. Which one of the following calculations will print a value smaller than zero in a Python interpreter?

- a. `>>> (-2) ** 2`
- b. `>>> -2 ** 2`
- c. `>>> 2 ** -2`
- d. `>>> 2 ** (-2)`
- e. None of the above has a value smaller than zero

7. Which one of the following statements is true in a Python interpreter?

- a. Two different types of numbers cannot be calculated together.
- b. Two different types of objects cannot be compared together.
- c. Two different types of objects cannot be printed together time.
- d. Two different types of objects cannot be concatenate together.
- e. None of the above is true.

8. What is the print value if you execute the following codes as a Python file?

```

x = 5
y = 10
x = y
y = x
print x, y

```

- a. 5 5
- b. 5 10
- c. 10 5
- d. 10 10
- e. None of the above

9. What will be printed by typing the following command in a Python interpreter?

```
>>> print 'Hello' + 'World' + '!'
```

- a. on one line the text:
Hello World !
- b. on one line the text:
HelloWorld!
- c. Hello, World, and exclamation are separated into different lines:
Hello
World
!
- d. Nothing will be printed

e. ERROR

10. Which one of the following commands will assign a value False to the variable test?

- a. test = 'hi'
- b. test = type('hi') != type('student')
- c. test = False == False
- d. test = len('test') == 4
- e. test = False != len('test')

Phase C

1. Which one of the following statements is true in a Python interpreter?

- a. A length function will return 5 for a variable with an integer value 5
- b. Your program needs to be compiled before executing in the Python interpreter.
- c. Operator plus can be used to add two integers or concatenate two strings
- d. A function can be defined after it is called without error.
- e. None of the above is true

2. Which one of the following commands cannot be used to leave a Python interpreter?

- a. exit()
- b. quit()
- c. leave()
- d. Control-D
- e. All of the above can be used to leave a Python interpreter

3. Which one of the following statements is false?

- a. In a command line, if you type "python", then you will go into a Python interpreter
- b. In a command line, if you type "python file.py" where file.py is a Python file, file.py will be executed
- c. In a Python file, a function will be only executed when it is called.
- d. In Python, double equal sign means mathematically equal while single equal sign means assign a value from the right to the left.
- e. In Python, the type of an object should be defined before being initiated.

4. What are the print out values if you execute the following Python codes as a Python file?

```
a = 3
```

```
b = a + 2
```

```
c = 'Value is'
```

```
print(c + b)
```

- a. Value is 5
- b. Value is a + 2
- c. Value is b
- d. c + b
- e. Error.

5. What are the print out values if you execute the following Python codes as a Python file?

```
def function1(v1, v2):
```

```

        return v1 - v2

def function2(v1, v2):

    return v1*v2

a = 2

b = 3

print function1(a+1,b) + function2(b+1,a)

```

- a. 6
- b. 7
- c. 8
- d. function1(a,b) + function2(b,a)
- e. ERROR

6. Which one of the following calculations will have the smallest number in a Python interpreter?

- a. `>>> 12 / 5`
- b. `>>> 12.0 / 5.0`
- c. `>>> 12 / 5.`
- d. `>>> 12. / 5`
- e. All of the above have the same values

7. Which one of the following calculations will have the largest number in a Python interpreter?

- a. `>>> -2 ** 3`
- b. `>>> 12 % 5`
- c. `>>> 12.0 / 5`
- d. `>>> 2 ** -3`
- e. `>>> 12.0 % 5`

8. What will be printed for the following command in a Python interpreter?

```
>>> print type(1/3)
```

- a. `<type 'number'>`
- b. `<type 'float'>`
- c. `<type 'double'>`
- d. `<type 'int'>`
- e. `<type 'tuple'>`

9. Which one of the following calculations will have the largest number in a Python interpreter?

- a. `>>> 2 * 2 * 2`

- b. `>>> 2.0 * 2.0 * 2.0`
- c. `>>> 2 ** 3`
- d. `>>> 2.0 ** 3`
- e. All of the above have the same value

10. What is the print value if you execute the following codes?

```
a = 4
```

```
b = 'test'
```

```
c = a == b
```

```
d = c != len(b) == a
```

```
print d
```

- a. True
- b. False
- c. 4
- d. 'test'
- e. ERROR

APPENDIX D
INTERVIEW MATERIALS

Behavior Questions

Phase C questions.

Q1

I. On a scale from 1 to 5, how much do you rate your certainty of question 1. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others _____

Q2

I. On a scale from 1 to 5, how much do you rate your certainty of question 2. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q3

I. On a scale from 1 to 5, how much do you rate your certainty of question 3. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

Part 2

II. Did you cheat on this question?

Yes

No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q4

I. On a scale from 1 to 5, how much do you rate your certainty of question 4. ?

Guess

1

2

3

4

5

I know the answer

II. Did you cheat on this question?

Yes

No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q5

I. On a scale from 1 to 5, how much do you rate your certainty of question 5. ?

Guess

1

2

3

4

5

I know the answer

II. Did you cheat on this question?

Yes

No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q6

I. On a scale from 1 to 5, how much do you rate your certainty of question 6. ?

Guess	1	2	3	4	5	I know the answer
-------	---	---	---	---	---	-------------------

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q7

I. On a scale from 1 to 5, how much do you rate your certainty of question 7. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others _____

Q8

I. On a scale from 1 to 5, how much do you rate your certainty of question 8. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others _____

Q9

I. On a scale from 1 to 5, how much do you rate your certainty of question 9. ?

Guess	1	2	3	4	5	I know the answer
-------	---	---	---	---	---	-------------------

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others_____

Q10

I. On a scale from 1 to 5, how much do you rate your certainty of question 10. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

II. Did you cheat on this question?

Yes No

III. If you cheated, how did you cheat?

Use cheat sheets.

Others _____

Phase B questions.

Q1

On a scale from 1 to 5, how much do you rate your certainty of question 1. ?

Guess	1	2	3	4	5	I know the answer
1						
2						
3						
4						
5						

Q2

On a scale from 1 to 5, how much do you rate your certainty of question 2. ?

Guess 1 2 3 4 5 I know the answer

Q3

On a scale from 1 to 5, how much do you rate your certainty of question 3. ?

Guess 1 2 3 4 5 I know the answer

Q4

On a scale from 1 to 5, how much do you rate your certainty of question 4. ?

Guess 1 2 3 4 5 I know the answer

Q5

On a scale from 1 to 5, how much do you rate your certainty of question 5. ?

Guess 1 2 3 4 5 I know the answer

Q6

On a scale from 1 to 5, how much do you rate your certainty of question 6. ?

Guess 1 2 3 4 5 I know the answer

Q7

On a scale from 1 to 5, how much do you rate your certainty of question 7. ?

Guess 1 2 3 4 5 I know the answer

Q8

On a scale from 1 to 5, how much do you rate your certainty of question 8. ?

Guess 1 2 3 4 5 I know the answer

Q9

On a scale from 1 to 5, how much do you rate your certainty of question 9. ?

Guess 1 2 3 4 5 I know the answer

Q10

On a scale from 1 to 5, how much do you rate your certainty of question 10. ?

Guess 1 2 3 4 5 I know the answer

Demographic Questions

1. What is your major? _____

2. On a scale from 0 to 5, how much do you rate your capability of computer programming before conducting the experiment?
 0. I get a no / low knowledge on the topic
 1. I get knowledge on the topic but that I have not applied
 2. I know the subject and have practiced it, but with no true step back on the results
 3. I have practiced this topic several times, with good results
 4. I know and apply the good practices on this topic (and I get good results)
 5. I master the topic and I am able to teach someone on this topic

2. What is your gender? _____

3. What is your age? _____

4. What is your year in school?
 - a. Freshman
 - b. Sophomore
 - c. Junior
 - d. Senior

If you want to cheat in an online exam, will you plan ahead before taking an exam or have an impulsive cheating while taking an exam?

APPENDIX E
INSTRUCTIONS OF THE EXPERIMENT PROCESS

Instruction

The goal of this experiment is to categorize non-verbal behaviors during online assessments and identify significant cues related to cheating. You will watch a 12 minutes video lecture first, take a 26 minutes computer-based exam, and finally have a 15 minutes feedback interview to report how you came about getting the answer (i.e. knew it, guessed, or cheated) and how you try to avoid being detected while cheating in an exam. Your keystrokes, mouse clicks, videos and audios, will be recorded. The whole experiment is separated into three parts. Part specific instructions will be presented to you before each of the three parts of the experiment. Do you have any questions?

Phase A

In this part of the experiment, you will watch a short video lecture along and review the material package on which the video lecture is based. All information in the video lecture can also be found in the material package. The video lecture will last 12 minutes and you have 3 minutes to review it. You are allowed to take a note while watching the video. Afterwards, you will be asked to complete a short multiple-choice test. Your engagement is important and please do not fall asleep or play a cell phone while watching the video. In the pilot study, most subjects can understand the content in video lecture without any problem. If you are still confused about the content, I will tutor you after the video lecture. So don't worry about it. Do you have any questions?

Phase B

You will now take a computer-based multiple choice exam. You will be given 13 minutes to take this exam. There will be total 10 multiple-choice questions and a timer will be showed up at the upper-right corner. At the fifth question, there will be a finish button shown as red. If you click it, the exam in this part will be finished. If time is up, all your answers will be automatically submitted and you will be direct to the finish page, too.

All of your behaviors will be recorded during taking the exam. This includes keystrokes, mouse clicks, audio, and video. All your material packages and notes will be taken away. You

should try to answer the questions by yourself as best as you can and you are not allowed to cheat. Do you have any questions?

Phase C

Right now, you will take another 13 minutes exam. There will be total 10 multiple-choice questions and the whole settings are the same as the previous one. It is assumed that you did not prepare well and therefore, you want to cheat in order to pass this exam. However, you do not want to be caught by the proctoring system, either, because you will fail in this exam if you are caught. All material packages, cheat sheets, and notes will be given back to you and you have two minutes to setup the environment to deceive the proctoring system successfully. The answers to the questions can be found in the cheating sheet Again, your behaviors including keystrokes, mouse clicks, audios and videos will be recorded. Please do not cover your faces or eyes while answering the questions. Do you have any questions?

Phase D

In this part of the experiment, the experimenter will conduct a brief interview with you. During this interview, we will go through your assessment video one question at a time. I will ask how you answered the question. Please answer truthfully, there will be no penalty toward your participation or your course grade based on your answers. The questions that we will cover for each question are below.

APPENDIX F
CONSENT INFORMATION

Categorize Typical and Atypical Behaviors in online assessments

I am a Ph.D student in the program of Simulation, Modeling and Applied Cognitive Science. In collaboration with my advisor Dr. Femiani, who is faculty in the Department of Computer Science and Engineering and, co-advisor, Dr. Craig, who is faculty in the Department of Human System Engineering, I am conducting a research study to evaluate and categorize the typical and atypical behaviors during online assessments.

I am inviting your participation, which will involve using the recorded data including keystrokes, mouse clicks, and video during an online testing. Your data will be used to develop an algorithm which can identify typical and atypical behaviors during online assessments. It will include 10 minutes lecture and a 20 minutes exam. There are two parts of exams: in the part one, you should try to answer the questions by yourself as best as you can and you are not allowed to cheat; in the part two, you are allowed to cheat in order to pass the exam. After the exam is finished, we will ask you to give your opinions about when you cheated and how you tried to avoid being detected while cheating during online assessments.

Your participation is voluntary, and you have the right to reject the participation of this survey. If you choose not to participate in the study, there will be no penalty; your grade in any of the courses will not be affected. To be eligible to participate you must be 18 or older and enrolled in PSY 101 during the Fall semester 2014.

The experiment will take about one hour. You will receive one hour of course credit for your participation in this experiment toward your PSY 101 research requirement. You may not gain any direct benefits for participating in this study; however, the results from this study will be used to improve future offering of online courses and assessments. There are also no foreseeable risks or discomforts to your participation.

Your name and any identifying information will be stripped from the data you submit. You will be assigned a unique identifier code (UIC), which will be used to identify your data. The UIC will be connected to your identity by way of a single separate locked file that will only be accessible to the research team. The UIC will minimize opportunities for your identifiable information to become public. All data will be destroyed after 6 month of the end of courses. The results of this study may be used in reports, presentations, or publications but your name or references to your identity will not be made.

We would like to record your testing behaviors including keystrokes, mouse clicks, videos and audios during the computer-based exam. The exam will not be recorded without your permission. Please let me know if you do not want the exam to be recorded; you also can change your mind after the exam starts, just let us know.

If you have any questions concerning the research study, please contact research team at : (Mr. Chia-Yuan Chuang – cchuang6@asu.edu; Dr. John Feminai – john.femiani@asu.edu; and Dr. Scotty Craig – scotty.craig@asu.edu). If you have any questions about your rights as a subject/participant in this research, or if you feel you have been placed at risk, you can contact the Chair of the Human Subjects Institutional Review Board, through the ASU Office of Research Integrity and Assurance, at (480) 965-6788.

By signing below you are agreeing to be part of the study.

Name: _____

Signature: _____ Date: _____