Getting to grips with exam fraud:

A qualitative study towards developing an evidence based educational data forensics protocol.

Researcher :
Christiaan Jan van Ommering
s1754262
c.vommering@explain.nl

Supervisor(s) :
Prof. Dr. Bernard Veldkamp
b.p.veldkamp@utwente.nl

Prof. Dr. Theo Eggen
Theo.eggen@cito.nl

External supervisor and organization:
Dr. Sebastiaan de Klerk
s.dklerk@explain.nl
eX:plain
Disketteweg 6,
3821 AR Amersfoort

August, 2018.

Acknowledgements

The process of writing my *Educational Science and Technology* master thesis lasted seven challenging months. I can sincerely say that I am proud of this thesis, the process I have been through, and the final protocol that was developed as a result. However, I certainly did not do it all on my own. Therefore, I would like to take this opportunity to thank those who supported me and guided me through this process.

First, I would like to thank both my supervisors; Prof. dr. Bernard Veldkamp for providing me with good feedback, new insights and regular communication about the thesis, and my second supervisor Prof. dr. Theo Eggen. Also, a special word of thanks is in order regarding their flexibility during the final phase of my graduation. Secondly, I would like to thank my colleagues at eX:plain for their support and for providing a very pleasant working atmosphere during this period. Special thanks go to dr. Sebastiaan de Klerk and Kees Boonman for the confidence that they placed in me at the start of this project, and for their support along the way.

I would also like to express my gratitude to my Isabella for being patient with me the last couple of months, and for always encouraging me to follow my dreams. Finally, I would like to thank my family and close friends for the motivational talks and support.

Thank you.

Christiaan Jan van Ommering.
July 2018

Getting to grips with exam fraud:

A qualitative study towards developing an evidence based educational data forensics protocol.

Keywords

Test Security, Data Forensics, Exam Fraud, Cheating, Misconduct.

Abstract

Combating exam fraud before it occurs is often described as ideological and unattainable. On the other hand, it is described as a vital step to preserve the integrity of the exam. Hence, the question among practitioners is: 'How do we get to grips with exam fraud'. For this reason, eX:plain, a Dutch testing agency, developed a data forensics monitor. This monitor analyses response data using multiple fraud indices. Subsequently, practitioners must be able to act on indications of exam fraud.

This design research was focused on developing standards covering the entire process of examination to limit the chances of security risks (e.g., the prevention of exam fraud as much as possible, and detection by means of data forensics), together these standards form the Educational Data Forensics Protocol. Two research questions guided this study. The first question, which standards regarding preventing and detecting fraud in the process of examination need to be included into the EDF protocol? In addition, practitioners must be able to act on indications of exam fraud based on these standards. Therefore, a second research question was formulated, namely which conditions must be considered during development of the EDF protocol to support practitioners in detecting possible gaps in the security of their examination process?

The EDF protocol was developed and validated in five consecutive steps. This thesis analyses on the theoretical base of developing the EDF protocol (*Step 1*) and the considerations for developing a prototype (*Step 2*). The prototype was being validated (e.g., establishing correctness of the content) through seven semi-structured interviews with content experts in the field of either test security or data forensics (*Step 3*). Statements from these interviews were used to adjust in the prototype to finalize the EDF protocol (*Step 4*). Finally, to determine the practical value, the final version of the EDF protocol was used to flag gaps in the security of the exam process and determine possible security risks for one of eX:plain' s exam programs (*Step 5*).

Content

List of Figures and Tables

Figures

Tables

Introduction

In education, performance is measured mainly by using grades. These grades appear to have a major impact on student lives by means of pressure to perform well, and being concerned about failing (McCabe, Butterfield & Trevino, 2006). Hence, cheating reflects the need to get passing grades, especially considering high-stakes testing. Over the last two decades, interest in exploring ways of detecting and preventing cheating, fraud, or (test) misconduct in education has been growing. Literature provides separate definitions, due to this diversity it can be difficult to make a clear distinction between these terms. In all cases (i.e. cheating, fraud or misconduct) they refer to the intention of deliberately influence (parts of) the examination process with the aim of obtaining a different result on the exam or for personal gain. Rather than try to define these terms separately, the definition applies when these terms are used alternately in this study.

McCabe (2005) reported that 26% of the students admitted to cheat during test taking back in 1961, this percentage increased to 52% in 1991. In a 1999 study 75% of students admitted to cheat during tests. In similar fashion, based on a longitudinal research, Murdock, Hale and Weber (2001) reported an increase in cheating over the last decades, while they also found that the severity of individuals perceived dishonest behaviour has decreased. Although this research is outdated, it does show a certain trend in behaviour. In fact, in a more recent study, both Novotney (2011) and Witherspoon, Maldonado and Lacey (2012) state that students whom cheat during college are also more likely to partake in other unethical behaviour, for example cheat on their spouse, and cheat at work. Although prevalence numbers on cheating differ amongst research findings, ranging from 50% to 95% (McCabe, 2005; McCabe et al., 2006; Oleck, 2008; Witherspoon et al., 2012; Yee & MacKown, 2010), there seems to be consensus in the fact that cheating is a growing problem in contemporary education.

In similar fashion Computer-Based Testing (CBT) is becoming a more popular administration mode for examination. CBT is particularly popular with standardized testing, because of its operational advantages (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014). Although CBT provides many advantages in terms of testing and analysing the test data, it also comes with new security risks, for example the ability to obtain and share test information (Impara, Kingbury, Maynes & Fitzgerald, 2005; Marianti et al., 2014). Fortunately, next to the technological progress that brought CBT, also various methods are the topic of research in detecting cheating, for example (educational) data forensics (Impara et al. 2005; de Klerk, 2017; Plackner & Primoli, 2012).

Educational Data Forensics (EDF) offers a promising opportunity to detect cheating behaviour, for example by looking at aberrant response patterns, response time, and suspicious test results on individual levels as well as group level by comparing test results to prior examinations. Despite these advantages on the levels of securing the examination program, analysing examinees' behaviour and the exam results, it is important to proceed with caution when using data forensics. Mainly because few studies report on the practical use and the reliability of data forensic methods (van Noord, 2018). Therefore, caution is advised in terms of decision making, for example. After all, data forensic indices provide indications of potential fraud rather than detect actual fraud. Therefore, the results of data forensics analysis should lead to further investigations (e.g., discus possible irregularities during examination), rather than sanctioning an examinee based on the results (de Klerk, 2017). The number of high-stakes testing programs that use data forensics is growing, due to the increasingly popular idea that it is essential to act on evidence of test misbehaviour to protect the validity of test programs (Fremer, 2011). These aberrant patterns may not always be caused by exam fraud, which stress the demand for reliable and accurate data forensics to detect aberrant patterns.

Two questions rise from the promising prospect of using data forensics to detect potential misconduct during the process of examination; can we be sure that indications of cheating are based on flawless and validated data forensics, and can we be sure students are not innocently accused of fraud? These questions support the vital need for an evidence based protocol when it comes to detecting and indicating misconduct after analysing response data. Detecting fraud remains an extremely difficult endeavour (Impara et al., 2005). While some students who get accused of cheating acknowledge it, many do not (Howell, Sorensen & Tippets, 2016). Therefore, it is the responsibility of testing agencies, such as eX:plain, to stay aware of the latest developments in terms of how test takers cheat, and can react on that with fitting responses, to preserve the integrity of examination standards (Howell et al., 2016).

To that aim, eX:plain started a data forensics project (Xquiry) involving both fixed and randomized exams. In fixed exams, all examinees are presented with the same set of questions. For

randomized CBT, the questions are randomly drawn from an item bank, and the answers are presented in randomized order. In practice this means that a group of examinees answer completely different sets of items. In collaboration with the University of Twente, eX:plain developed a data forensics monitor (DFM). The DFM is an online web application that can analyse large amounts of data using multiple data forensic indices. Van Noord (2018) reported, by means of an experimental study, on the validity of the data forensics monitor. The DFM flagged 38% of potential cheaters with a reliability of 97%. As far as the usability of the DFM is concerned, these findings are promising. However, what remains unanswered is the question if we, to some extent, can prevent cheating from happening, and what the follow-up steps should be when cheating is detected by means of the DFM.

Therefore, the current study is considered a follow-up on the experimental study by Van Noord (2018). The aim of this study is to provide evidence based security standards, in the form of an EDF protocol for preventing, detecting and acting on indications of exam fraud. In practice, this protocol should be able to be implemented, both with and without the application of the data forensics monitor. The educational data forensics protocol can be regarded as an audit on the safety and fraud resistance of the exam and/or exam process. This audit shows possible security gaps and provides the user with practical guidelines to act, hence preventing misconduct in the future. To determine the practical value, the EDF protocol will be validated through interviews with experts from the field and subsequently implemented at one of eX:plains exam programs.

Conceptual framework

To clarify the context of this study, relevant concepts are discussed in this section. First, literature on data forensics will be discussed. Secondly, this study is part of the educational data forensics cycle developed by eX:plain, therefore an extensive discussion of this cycle is also provided.

Educational data forensics

In practice a variety of statistical methods are used for the detection of aberrant behaviour in response data. According to Impara et al. (2005) aberrance in response data is observed when data provided by an examinee is inconsistent with his or her expected performance (i.e. knowledge and behaviour). These methods offer the opportunity to look at response times, unusual response patterns, and computing indices of collusion on the individual, group and school level (Impara et al., 2005; Plackner & Primoli, 2012). On the level of the individual test taker, aberrant patterns in response times could indicate fraudulent behaviour (Marianti et al., 2014). This collective of statistical methods used for detecting exam fraud is referred to as (educational) data forensics (Clark & Kingston, 2014).

Provided test data is available over multiple test occasions, data forensics allows to analyse individual responses on every item of the test on every test occasion (Fremer, 2011). These results can be used to create models, which indicate a normal response to test items, and detect test-taking irregularities. Item Response Theory models can be used to objectively determine if a set of responses is in accordance with the probability model, which accounts for multiple levels (Impara et al., 2005). Plackner and Primoli (2012) for example, did an exploratory study on ten different data forensic methods (e.g., erasure analysis, scale score changes, performance level changes) to report if these models were accounting for variation in response irregularities. Although they point to some prudent considerations for future research they conclude that, to a certain degree, they all do account for irregularities in the test scores. Impara et al. (2005) reported on a study in which (among other things) aberrance, in a computerized adaptive testing environment could indicate cheating behaviour.

An example of aberrant behaviour can be that a student answers multiple consecutive questions correct within a very short time. This can be suspicious, particularly considering that the examinee would not have had time to read multiple questions and think about what the answers should be (Musthaler, 2008). Although this example seems to indicate that a student is cheating, caution is still advised in terms drawing conclusions, because aberrant response patterns are not always caused by misconduct. In this example poor preparations on the test and guessing could also have influenced the test response model. Accordingly, if there is a suspicion of fraud based on the data forensics, additional inquiry and evidence is required (Ferrara, 2017).

These findings emphasize the added value of using data forensics to preserve the integrity of the exam. However, the exam process consists of more steps than just 'examination' and 'result analysis'. Moreover, owing the Ferrara's (2017) earlier statement, practitioners should be provided with a clear set of guidelines or a 'tool-kit' to protect the integrity of the exam (e.g., flag potential security risks within the exam process). As a result, eX:plain started a project, called Xquiry, in which an educational data forensics improvement cycle was designed. The implications of this improvement cycle are explained in the next section.

The educational data forensics improvement cycle

The Xquiry educational data forensics improvement cycle consists of the EDF Monitor, the EDF Protocol, and the EDF Qualifier (Figure 1). The purpose of the cycle is to reduce the frequency in which exam fraud occurs. Although there is no set order for applying this cycle (e.g., both the monitor or the protocol can be the starting point). First the monitor will be discussed, because it has already been developed and validated through experimental research (van Noord, 2018). This way the improvement cycle can be described in order of development, since the development of the EDF protocol is the basis of this design study.

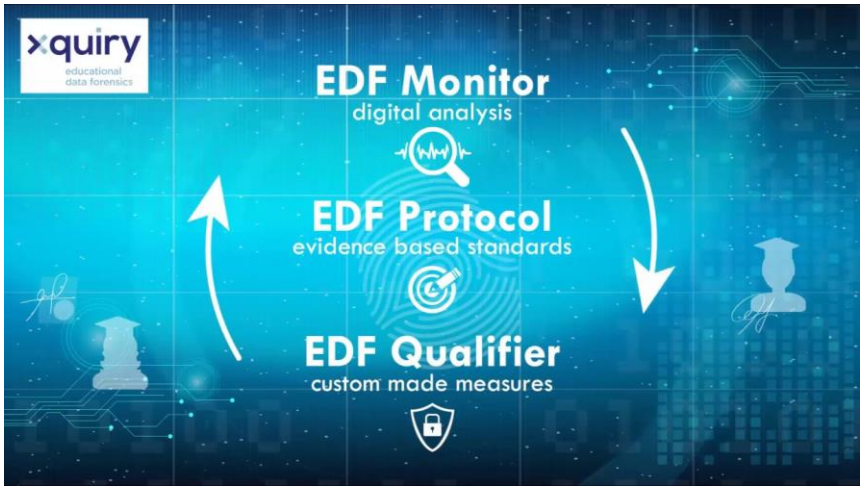

Figure 1. The educational data forensics improvement cycle (Xquiry, 2017).

The EDF Monitor

The EDF monitor can be considered an entry point of the improvement cycle, because this is a separate service offered by Xquiry. The monitor can be used to detect aberrant patterns in the exam data. These aberrant patterns are detected by means of a web based data forensics monitor (DFM) in which fraud indices are programmed to search the data for suspicious results. eX:plain developed the data forensics monitor in collaboration with the University of Twente. In the DFM there are several fraud indices included that can indicate possible fraud. These indices were the topic of a recent study to determine their validity and effectiveness (van Noord, 2018). Van Noord conducted an experimental study in which examinees were assigned to specific 'cheating behaviour' groups in order to determine if these kandidates would be flagged through DFM analyses. The results showed that the data forensics monitor can flag 37% of fraudulent behaviour, with a reliability of 97%.

The monitor is not only useful for detecting possible fraud among individual candidates, groups of candidates, exam officers, but also very useful for determining how often exam fraud generally occurs within the exam. For example, it is possible to check whether exam fraud occurs more often in certain item banks or certain exams. Provided the data is available regarding exam officers, locations, etc. Any suspicious patterns are automatically reported in the software interface, which uses available exam data up to 2 years to calculate (Xquiry, 2017).

However, the usability of this service is limited to analysing response data, with the aim of fraud detection, whereas fraud can occur at various steps within the process of examination. Therefore, Xquiry also aims to offer client a tool-kit to prevent exam fraud by means of the EDF protocol.

The EDF Protocol

The EDF protocol, which has yet to be developed by means of this design study, is aimed to be a quality assurance system, validated by content experts, which is aimed at prevention (i.e., the prevention of exam fraud as much as possible in advance) and detection of exam fraud. Although exam fraud can never be fully banned, the protocol will provide security standards to limit the chance of exam fraud. These standards can be used to determine whether there is a security risk (e.g., a high, medium, or low security risk). That is why the interaction with the EDF monitor is of the utmost importance, because it can highlight fraud trends and possible security gaps. By using the EDF protocol, based on scientifically based standards, gaps in the exam security can be detected (Xquiry, 2017).

The standards in the EDF protocol that relate to the prevention of exam fraud can be divided into two categories, namely physical and technological standards. Physical standards relate to how candidates enter the examination room, what kind of materials they can bring, and what precautions are taken against (technological) aids etc. The technological standards have an ICT technical impact. This relates to test construction (test items), the way items are manufactured and exchanged between test experts, and how an item bank is secured etc. In terms of detecting exam fraud, the EDF protocol standards can also be divided into physical and technological standards. Surveillance during exams, and the detections of unwanted objects in the examination room are examples concerning physical standards. The technological standards are mainly based on the EDF monitor. Applying an exam fraud detection system is of course an important standard for detecting fraud (Xquiry, 2017).

The EDF Qualifier

The third part of the cycle is the EDF qualifier. The EDF qualifier is an organisation which applies the improvement cycle. This means that the monitor is used repeatedly to measure the effect of the protocol. Also, based on the outcomes of the protocol application, the next step is to implement tailor-made measures. These measures are aimed to make exam fraud more difficult to achieve, and to lower exam fraud grades as determined by the EDF monitor or protocol. Hence, this makes the improvement cycle complete. In this way the EDF continuous improvement cycle offers a good working system for fraud prevention and detection, which is indispensable to guarantee the quality of the exam (Xquiry, 2017).

Research question and model

This research is focused on developing standards covering the entire process of examination in order to limit the chances of security risks (e.g., the prevention of exam fraud as much as possible, and detection by means of data forensics). Accordingly, the corresponding research question is:

1. Which standards regarding preventing and detecting exam fraud in the process of examination need to be included into the EDF protocol?

In addition, practitioners should be able to act on indications of exam fraud based on these standards, this study will therefor also answer a second research question:

2. Which conditions must be considered during development of the EDF protocol to support practitioners in detecting possible gaps in the security of their examination process?

Scientific and practical relevance

The aim for this study is to develop an evidence based EDF protocol. Data forensics is not a new field of research, computer based testing is still growing in popularity, however also not new. What makes writing an EDF protocol a challenging endeavour lies in the fact that this is unchartered territory. So far data forensics had only been used in analysing fixed examinations. Xquiry however, like mentioned before, only just determined the validity of their fraud indices by means of an experimental study on randomized exams. Exam fraud is a serious threat to the validity of the exam. Most examination organizations put a lot of time, money and effort into developing reliable exams. However, the time, money and effort invested into fraud prevention and especially detection is often very little. Even though this is a vital part of ensuring exam validity. Therefore, a practical and evidence based protocol for fraud prevention and detection is indispensable for practitioners to guarantee the quality of the exam. This study aims to provide an in-depth understanding of factors to prevent, detect, and act on indications of exam fraud and thereby add to the examination practice.

Method

To develop a set of standards on securing the process of examination this study is based on a qualitative design research method. Hence, a naturalistic approach is used to understand a context specific phenomenon (Golafshani, 2003), namely misconduct during examination (i.e. cheating, fraud).
The EDF protocol for preventing and detecting exam misconduct was constructed and validated in five consecutive steps: [1] a literature search relating relevant standards and criteria on security of the examination process, and also prevention and detection of exam misconduct; [2] development of the prototype; [3] validation of the prototype standards and criteria through semi-structured interviews with content experts; [4] adjustment of the prototype towards a final version of the EDF protocol; and [5] empirical testing of the protocol by putting the protocol to practice.

Step 1- Literature search

For the first step the PRISMA framework described by Moher, Liberati, Tetzlaff and Altman (2009) was used for conducting the literature review. To compile an evidence base for the development of the EDF protocol, three major databases were searched: Scopus, Web of Science, and Google Scholar. Foremost, using multiple databases enabled a more diverse collection of literature, but it also compensated for subjectivity by the researcher and uncontrolled validity threats (Rousseau et al., 2008).

For the main topic of the study several search terms were used (see table 1). Boolean search operators were also used during this step (e.g., AND, OR, NOT, and *). The initial search findings were thinned through excluding duplicates. Hereafter, the articles were first screened on title, and secondly the abstract. Articles were included in the study if the main topic of the paper or chapter related to security of examination, or if the paper or chapter provided a structured set of guidelines or standards on security of examination. This method not only summarized existing literature, but also aimed to generalize and transfer findings for policy making and practice (Cassell et al., 2006). Prior to the development of the EDF prototype, an overview was made of the most important findings from the literature review. These insights were used to develop an EDF prototype.

Table 1
Search terms used in the literature search

Keywords	Related/more specific/broader
Test Security	Educat*, Prevention, Detection, Standards, Fraud, Cheating
Data Forensics	Educat*, Fraud, Cheating

Step 2- Developing an EDF-Protocol prototype

Based on the findings of step 1, a consultation with the scientific advisor and manager of Xquiry was organised in order to ensure that development was in line with their goals and expectations. The insights gathered in the consultation were used in the development of the first set of standards of the prototype (Part A), as well as a corresponding grading system. This initial prototype was hereafter discussed during consultation with the Xquiry team. The intention was to make the standards (concerning prevention of misconduct during the process of examination) as complete as possible before starting the interviews. Owing to this, subsequent feedback loops emphasized the content of the final version of this part of the prototype.

The development of part B (i.e. the standards and criteria for detection of misconduct by means of using data forensics) took considerably more time and effort. Although there is a considerable amount of scientific literature on the possibilities of using data forensics, research is mostly focused on case- or compare studies, and thus often lacking proper directions for practical implementation. For this reason, these standards have been less elaborated and to a lesser extend discussed with the Xquiry team compared to the content of part A. The intention with this part of the prototype was therefore to enter the interviews more open minded, hence gain insight on what the content experts deem to be included or excluded in terms of data forensic standards.

During this step a deliberate choice was made for a distinction between a set of standards for prevention as well as a set of standards for detection (by means of data forensics) because these actions not always coincide in practice.

Step 3- Validating the EDF-Protocol standards

Ethics and Participants

Before establishing the correctness of the prototype standards, ethical approval for conducting the semi-structured interviews was requested and granted by the Behavioural, Management and Societal Sciences Ethics Committee of the University of Twente (Appendix B). The prototype was validated by means of seven semi-structured interviews. All approached experts have practical and theoretical experience on the subject. These interviews were held with content experts from different backgrounds, amongst them psychometricians, policy makers and practitioners in the field of test security or education. Multiple experts who focus their work on (parts of) test security have been approached with the question to partake in this design research to gain an insight on their opinions and experiences concerning standards for securing the process of examination. To keep development of the prototype and validation of the content separate steps, the participating experts were not involved during the development of the prototype.

Procedure and Materials

For the purpose of developing a protocol which can be use in practice, the current study, will make use of experts interviews. Systematic expert interviews offer the possibility of identifying strengths and weaknesses in the content (McKenney & Reeves, 2012; Piercy, 2004). This method is a valuable source of data collection, particularly when establishing the correctness (e.g., validating) content of a product (Wools et al. 2011). The interview format consists of four categories; category one focused on general questions concerning the protocol (n=7), category two focused on questions concerning the protocol content (n=4), category three related to the grading of the protocol (n=5), and category four focused on the data forensic standards (n=5). An example question of a question would be: "The goal of the protocol is to provide a good check whether the process of examination is secure. Do you think this is feasible in the current form?"

All potential respondents were first contacted through e-mail or Linked-In. This e-mail contained a brief introduction of the researcher, a short explanation of the context of the design research (e.g., the duration of the interview, and method of the interview), and finally the request for an interview (Appendix C). Initially eight requests have been sent to potential respondents. However, due to a possible conflict of interest, one candidate preferred not to take part in an interview. A second e-mail was send to the seven remaining respondents. In this e-mail a more thorough explanation of the protocol's goal was provided, the respondents were asked to read the prototype and the interview questions (Appendix D) in preparations of the interview, finally an appointment was set for the interview. This was either face-to-face (n=2), through phone (n=2), or through Skype(n=3). The choice for the interview method was in consultation with the respondents, and mainly based on convenience for both parties.

At the start of the interview, each respondent was asked for consent verbally. This means that they were asked whether the interview could be recorded and whether the input from the interview could be used to validate the content of the prototype. It was also agreed in advance with the participants that they would receive the transcript of the interview, to be completely transparent about the input that was collected. The semi-structured interviews were conducted between April 17th and June 1st of 2018. Table 2 provides an overview of the interviews, including the focus of the interview and background information on the participants.

After the interviews, all the recordings have been converted to verbatim transcripts to keep statements in their proper context (Appendix E). Cues and codes were written in the margin of the transcript to indicate a reference to a specific question or part of the prototype. Subsequently, text fragments were summarized based on the interview categories (n=4). The selection of usable statements was done on an individual basis by the author.

Table 2
Overview of the interview dates, focus of the interview and respondents in randomised order

Part A/B	Name	Company	Background
Part A	-	eX:plain	-
Part A	-	VU Amsterdam	-
Part A	-	CvTE	-
Part A	-	Bureau ICE	-
Part B	-	Bureau ICE	-
Part B/A	-	CITO	-
Part B/A	-	Tilburg University	-

Step 4- Adjustment of the prototype and final EDF protocol

In the fourth step, the statements from the experts were used to transform the prototype into a final version of the EDF protocol. In the result section a comprehensive overview is provided on all changes made based on statements from the interviews. These statements provided several new insights, especially in terms of usability and assessment. This emphasizes the significant impact of the interviews on the validation process.

Step 5- Implementation of the EDF protocol

The first four steps mainly focused on validating the design, purpose and content of the protocol. In order to be able to determine the actual value for practice it had to be used in a real situation. In the fifth step of this design research the final EDF protocol was used to determine if there was a possible security risk within the process of the --- exam. In the scope of the current study, this step has been taken to determine the actual practical value of the protocol.

Results

In this section the results for each consecutive research step is described: [1] findings of the literature search; [2] development of the prototype; [3] validation of the prototype standards and criteria through seven semi-structured interviews with content experts; [4] adjustment of the prototype and final version of the EDF protocol based on the input of the interviews; and [5] empirical testing of the protocol by putting the protocol to practice.

Step 1- Literature search

The literature search was split into two main topics. Firstly, the search for literature on 'Test Security', and secondly the search for 'Data Forensics' related literature. The literature found is up to June 2018. As was described in the method section the PRISMA framework was used in this step (Moher et al., 2009).

The first major topic was 'Test Security'. The key search term was based on the research question, namely test security. To broaden or specify the search, the following search terms were also used: prevention, detection, standards, fraud and cheating. Not all search terms provided usable information. Figure 2 shows the steps of the search process. An overview of the articles and handbooks, that were included in the current study, is provided in table 3.

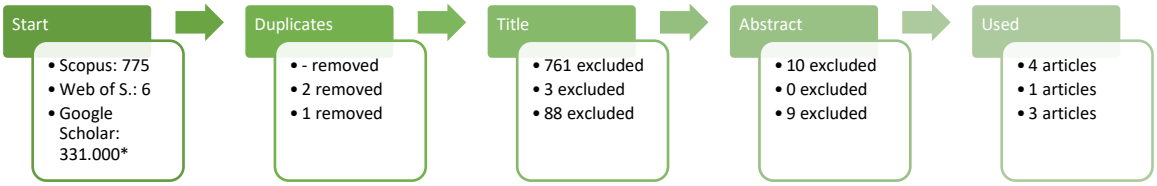

Figure 2. PRISMA flow chart showing the search process in the query of Test Security.
Note: *only the first 10 pages were scanned (n=100 hits).

Table 3
Overview of literature results based on the search terms 'test security', 'standards', and 'fraud' (n=8).

Document title	Authors	Year	Database / Source
Detecting Fraudulent Erasures at an Aggregate Level	Sinharay	2018	Web of Science / Journal of Educational and Behavioural Statistics 2018, Vol. 43, No. 3, pp. 286–315
A Framework for Policies and Practices to Improve Test Security Programs: Prevention, Detection, Investigation, and Resolution (PDIR	Ferrara	2017	Scopus / Educational Measurement: Issues and Practice Fall 2017, Vol.36, No.3, pp.5–23
Which Statistic Should Be Used to Detect Item Pre-Knowledge When the Set of Compromised Items Is Known?	Sinharay	2017	Scopus / Applied Psychological Measurement 2017, Vol. 41(6) 403–421
Detection of Item Pre-Knowledge Using Likelihood Ratio Test and Score Test	Sinharay	2017	Scopus / Journal of Educational and Behavioural Statistics 2017, Vol. 42, No. 1, pp. 46–68
Robust Detection of Examinees With Aberrant Answer Changes.	Belov	2015	*Journal of Educational Measurement*, 52 (4), pp. 437-456.
Item and test development strategies to minimize test fraud	Impara & Foster	2006	Google Scholar / Handbook of test development
Selected-response item formats in test development	Downing	2006	Google Scholar / Handbook of test development
The security risk assessments handbook	Landoll	2006	Google Scholar / The security risk assessments handbook

The second major topic of this step was focused on gathering literature on data forensics. For this topic, the main keyword, data forensics, directly relates to the main research question. Again, to broaden or specify the search at certain points, the following search terms were also used: educat* standards, fraud and cheating. Also on this account not all search terms provided usable information. Figure 3 shows the steps of the search process. An overview of the articles and handbooks, that were included in the current study, is provided in table 4.

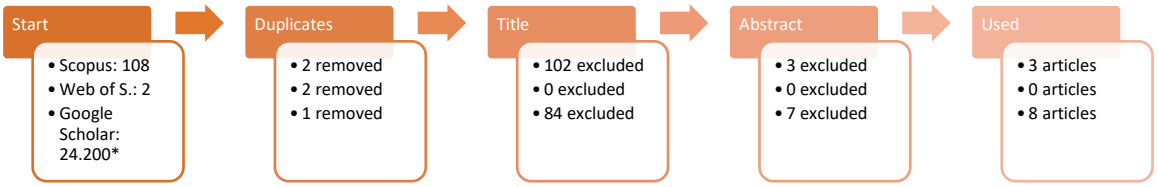

Figure 3. PRISMA flow chart showing the search process articles in the query of Data Forensics. Note: *only the first 10 pages were scanned (n=100 hits).

Table 4
Overview of literature results based on the search terms 'data forensics, and 'fraud' or 'cheating' (n=11).

Document title	Authors	Year	Database / Source
Exploring the impact of organizational investment on occupational fraud: Mediating effects of ethical culture and monitoring control	Suh, Shim & Button	2017	Scopus / International Journal of Law, Crime and Justice 53 (2018) 46–5
A New Statistic for Detection of Aberrant Answer Changes	Sinharay, Duong & Wood	2017	Scopus / Journal of Educational Measurement Summer 2017, Vol. 54, No. 2, pp. 200–217
Three New Methods for Analysis of Answer Changes	Sinharay & Johnson	2016	Scopus / Educational and Psychological Measurement 2016, Vol. 77(1) 54–81
Detecting cheating in computer adaptive tests using data forensics	Impara, Kingsbury, Maynes & Fitzgerald	2005	Google Scholar
Detecting potential collusion among individual examinees using similarity analysis	Maynes	2017	Google Scholar
Handbook of Quantitative Methods for Detecting cheating on Tests	Cizek & Wollack	2016	Google Scholar
Test Fraud: Statistical Detection and Methodology	Kingston & Clark	2014	Google Scholar
Educator cheating and the statistical detection of group-based test security threats	Maynes	2013	Google Scholar
Data forensics: A compare and contrast analysis of multiple methods	Plackner & Primoli	2012	Google Scholar
The magnitude and extent of cheating and response distortion effects on unproctored internet-based tests of cognitive ability and personality	Arthur, Glaze, Villado & Taylor	2012	Google Scholar
The new (and old) news about cheating for distance educators	Howell, Sorensen & Tippets	2016	Google Scholar

Because the literature search did not yield the desired results a snowballing approach, presented by Wohlin (2014) was used to find more relevant literature on the topic of test security. As a result of this method, scanning the reference lists of the articles and handbooks that were found during the initial literature search provided new information on studies in the field of data forensics and test security. Some of these are highlighted here, because they have proved to be very valuable for this study. For example, the handbook of test security (Wollack & Fremer, 2013) provided several directions for prototype content in terms of prevention criteria as well as input for the data forensics standards. Secondly, the Handbook of Quantitative Methods for Detecting Cheating on Tests (Cizek and Wollack, 2017) provided the ground work for the data forensics standards in the EDF prototype through offering multiple methodologies for identifying cheating of tests.
The third handbook, Test Fraud, by Kingston and Clark (2014) provided a solid summary of statistical detection methods. In terms of justification, these handbooks were very valuable since they summarize

lessons learned from practice and involve numerous content experts in writing these handbooks. A full overview of the additional findings through the snowballing method is shown in table 5.

Table 5
Overview of additional literature findings based on the snowball method (n=20)

Document title	Authors	Year	Database / Source
Utilization of Response Time in Data Forensics of K-12 Computer-Based Assessment	Liu, Primoli &Plackner	2013	Google
Combating academic fraud: are students reticent about uncovering the covert	Malgwi & Rakowski	2009	Google
Educator cheating and the statistical detection of group-based test security threats.	Maynes	2013	Google
Fraude onder studenten	Scholten	2013	Essay.utwente.nl
The role and responsibility of auditors in Prevention and Detection of Fraudulent Financial Reporting	Zager, Malis & Novak	2015	Google
Detecting and Preventing Cheating During Exams	Yee & MacKown	2010	Google
Fraud and plagiarism in school and career	Agud	2014	Google
Testing for Aberrant Behavior in Response Time Modeling	Marianti, Fox, Avetisyan, Veldkamp & Tijmstra	2014	Google
Comparing the Performance of Eight Item Preknowledge Detection Statistic	Belov	2015	Google
Detecting Test Tampering Using Item Response Theory	Wollack, Cohen & Eckerly	2015	Google
Detecting test tampering at the group level	Wollack & Eckerly	2017	Google
An Investigation of Answer Changing on Large-Scale Computer-Based Educational Assessment	Tiemann	2015	Google
Beat the Cheat	Novotney	2011	Google
Cheaters should never win: Eliminating the benefits of cheating	Fendler & Godbey	2015	Google
Observing and Deterring Social Cheating on College Exams	Fendler, Yates & Godbey	2018	Google
Countering Fraud for Competitive Advantage	Button & Gee	2013	Google
Detecting fraud: the role of the anonymous reporting channel	Johansson & Carey	2016	Google
Using response times to detect aberrant responses in computerized adaptive testing	van der Linden & van Krimpen-Stoop	2003	Google
Using response time to detect item pre-knowledge in computer-based licensure examinations	Qian, Staniewska, Reckase & Woo	2016	Google
A bivariate lognormal response-time model for the detection of collusion between test takers.	Van der Linden	2009	Google

When looking for guidelines on securing the process of examination, a number of documents are available in the literature. For example, the Dutch Association for Examination (NVE), in collaboration with Caveon Test Security, has drawn up a document containing 15 guidelines which aim to secure the content of examinations (Caveon & NVE, 2015). These guidelines describe numerous criteria which need to be considered to secure the exam process. Another guideline is provided by SURF (2014), which contain guidelines for the security of digital testing. This document provides practical tips, and examples, however this document is limited to the part of actual examination and result analysis. Also, SURF firmly states that the user of these guidelines must continue to examine to what extent the provided examples are applicable, because the usability can be context bound. On a more international scope, the International Test Commission (ITC) released their International Guidelines on the security of Tests, Examinations, and Other Assessments (2014). In similar fashion, the Association of Educational Assessment – Europe (AEA Europe, APA & NCME 2014) released a framework for practitioners which can be used to compare and evaluate practices in the stages of development, scoring and reporting of the exam. A full overview of guidelines and frameworks that were found during the literature search in provided in the table below (Table 6).

Table 6
Overview of openly available guidelines (n=8).

Document title	Authors	Year	Database / Source
Framework of Standards for Educational Assessment	AERE, APA & NCME	2014	Google
Richtlijnen voor het beschermen van de inhoud van examens	NVE & Caveon	2016	Google
Test Fraud Threats	Caveon	2016	Google
Security Planning Rubric	CoSN	2017	Google
Guidelines of the Security of Tests, Examinations, and Other Assessments	ITC	2014	Google
Richtsnoer Veilige digitale toetsafname	SURF	2014	Google
COTAN Beoordelingssysteem voor de kwaliteit van tests	COTAN	2010	Google
RCEC Beoordelingssysteem voor de kwaliteit van studietoetsen en examens	RCEC	2015	Google

In summary, there is a multitude of frameworks and handbooks available in literature on security guidelines. However, to a certain degree these guidelines lack direction in terms of a concrete model for assessing a possible security risk. Most guidelines are very precise in stating what to consider, yet a distinction between a high, medium or low security risks are often absent. On this note, the COTAN rating system for test quality (2010), and the RCEC rating system for quality test and exams (2015) both provide valuable input. All literature described in this section was used to support and justify the content of the prototype. For this reason, the content of these articles will be described during step 2.

Step 2- Developing an EDF protocol prototype

Based on the literature review and reading through similar protocols, manuals and handbooks, two main areas for development were identified. First, an area concerning standards and criteria with a focus on preventing misconduct during the process of examination (Part A). Second, an area with a set of standards concerning the detection of misconduct after examination by means of data forensics (Part B). The EDF prototype' s body of content is presented in figure 4. The full EDF prototype is included in appendix A. The prototype standards each relate to the most commonly used and accepted guidelines on test security: The Security Guidelines by NVE and Caveon (2016), the guidelines by SURF (2014), and the Standards for Educational and Psychological Testing (AERE et al. 2014) as well as other literature to support the inclusion of these criteria in the prototype.

An important note for the development of the EDF protocol must be made beforehand. Namely, internally within eX:plain a protocol has been developed (Boonman, 2016), for which the NVE & Caveon (2016) guidelines have been adopted. To continue this path, the same guidelines also form an important base in the development of the current prototype. If there is no additional argumentation or reference to literature considering a certain standard or criterion, then the inclusion is considered valid because it already reflects the policy of eX:plain.

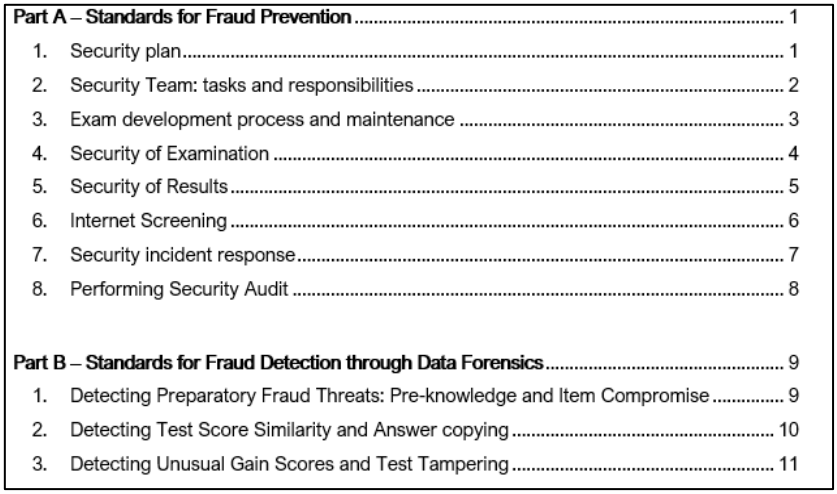

Figure 4. Overview of the EDF prototype content.

Content of Part A

The first part of the prototype involves eight standards. For part A, the aim was to describe a set of standards that cover all parts of the exam process, which were evidence-based, and are measurable.

The first standard is '*Security Plan'*. To check whether current practice surrounding the exam process is secure, a security plan that aims to minimize the chance that the content of the exam will become known to unauthorised persons and parties must be present within the organization (NVE & Caveon, 2016). According to Ferrara (2017), such a security plan should include rules, guidelines, requirements and procedures. This led to five criteria, which together form this standard, namely, (1) security plan, (2) security goals, (3) security policy, (4) actuality, and (5) financial resources. A plan containing goals for each process step as well as policies are vital to determine a starting point for secure practice. When a security plan exists, it can function as an instrument for comparing and evaluating practices in all stages of the examination process (AEA Europe, APA & NCME 2014).

Table 7
Criteria and Evidence base concerning standard 1

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Security plan	Guideline 1	Guideline 1.5.1	Section 2
Security goals	Guideline 1	Guideline 1.5.1	Section 2
Security policy	Guideline 1/4		Section 2
Actuality	Guideline 1/3		
Financial resources	Guideline 3		

The second standard is *'Security team: tasks and responsibilities'.* Both the tasks and responsibilities of all persons involved in the security of the exam content should be well-defined (NVE & Caveon, 2016). In this way the organization prevents weak points in the overall security of the exam process. Moreover, the organization must ensure that all parties involved recognize the value of security and that they carry out the relevant procedures correctly and carefully. Four criteria together form this standard, namely, (1) Security officer, (2) Security team, (3) Team responsibilities, and (4) Team competency. Zager, Malis and Novak (2015) state that there should be a clear distinction between roles and responsibilities of key stakeholders within an organization. This does not mean that tasks or responsibilities cannot overlap, but it emphasizes the importance of be being able to hold someone accountable for the responsibilities that they have. If these criteria are made explicit within practice (e.g., in the security plan) it increases awareness. In similar fashion, it is important that everyone involved, knows their place in the examination process, therefore ethics, fairness, and rights should be emphasized within security policies (AEA Europe et al., 2014). Part of these policies should be the training of personnel to protect and enforce security (Ferrara, 2017). In like manner, Ferrara (2017) states that a culture of professional ethics will stimulate both examinees and personnel to behave more professional in terms of reporting irregularities, and collaborating with investigations on such irregularities.

Table 8
Criteria and Evidence base concerning standard 2

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Security officer	Guideline 2		Section 4
Security team	Guideline 2		Section 1/3/4
Team responsibilities	Guideline 2/13		Section 2/3/4
Team competency	Guideline 2/13	Guideline 1.5.2	Section 2/4

Standard three is *'Exam development process and Maintenance'.* The organization is responsible for protecting the content of the exam during the development process through formal agreements and concrete security procedures (NVE & Caveon, 2016). These procedures ensure that the chance of fraud is minimized, which extends the usability of the exam content. A total of five criteria form this standard, namely, (1) content development, (2) exam construction, (3) items, (4) disclosure,

and (5) storage. Various statements from literature support the inclusion of these criteria (Table 9). Test developers can prepare multiple forms of an assessment to deter examinees from copying from their peers (Clark & Kingston, 2014). Moreover, they state a more extensive item bank increases the number of potential items an examinee might encounter, also making it more challenging for an examinee to memorize the items should they be disclosed.

Table 9
Criteria and Evidence base concerning standard 3

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Content development	Guideline 5	Guideline 1.5.2	Section 5
Exam construction	Guideline 6	Guideline 1.5.2&3	Section 3/4
Items	Guideline 5	Guideline 1.5.2	Section 2/3/4
Disclosure	Guideline 5/7/10	Guideline 1.5.2&3	Section 5
Storage	Guideline 6/7/10	Guideline 1.5.2&3	Section 5

Standard four is '*Security of Examination'.* Examination should be done in a secure manner (NVE & Caveon, 2014). Cheating can also be prevented by providing clear documentation regarding what is and is not allowed during examination, by having a clear policy and informing test takers of the policy in advance, examinees cannot claim ignorance of the boundaries of acceptable behaviour (Clark & Kingston 2014). Four criteria together form this standard, namely, (1) Proctoring, (2) Examination, (3) Planning and Acting, and (4) Use of materials. Proctors are considered a good method for preventing misconduct during examination (Bellezza & Bellezza, 1989; Cizek, 2001). However, to be effective, proctors should be trained on what to look for and how to handle suspicious behaviour in an appropriate manner (Cizek, 2001). Disclosing to examinees that statistical methods (e.g., data forensics) will be used to examine the exam results on their fairness can also deter exam misconduct (Bellezza & Bellezza, 1989; Cizek, 2001; Sinharay, 2017a, 2017b). When examinees have knowledge of forensics being used to analyse for indications of cheating, they are more likely to change their behaviour considering the negative consequences (Ferrara, 2017; Kranacher et al., 2010). Also, Clark & Kingston (2014) state that verifying the identities of examinees prior to examination ensures that individuals do not attempt to take the exam for another person. Something that is not directly part of these criteria is the fact that assigned seats significantly decline cheating (Fendler, Yates & Godbey, 2018), this could be a strong example of wat can be considered a good security measure.

Table 10
Criteria and Evidence base concerning standard 4

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Proctoring	Guideline 8/10	Guideline 1.5.2	Section 2/4/5/6/7/8
Examination	Guideline 8	Guideline 1.5.2	Section 2/4/5/6/7/8
Planning and Acting	Guideline 8/11	Guideline 1.5.2	Section 2/4/5/6/7/8
Use of materials	Guideline 8/11		Section 2/4/5/6/7/8

Standard five is '*Security of Results*'. Considering the high stakes of many exams it is necessary to verify that exam scores have been achieved in a correct and unsuspected manner. This makes it clear that they are suitable for decision making (NVE & Caveon, 2014). A total of five criteria form this standard, namely, (1) Plan, (2) Screening, (3) Transfer, (4) Data forensics, and (5) Sharing results. The majority of literature on validity of results is based on survey data. This is indeed a means to verify possible misconduct, however there is reason to believe that survey data is not sufficient. For example, some examinees might brag about cheating and another might be afraid to admit it. Therefore, one should ask himself how appropriate it would be to ask a potential cheater to honestly tell if they were dishonest during examination. Instead, Fendler et al. (2018) propose data forensics techniques to empirically observe actual cheating behaviour. A benefit of this method compared to surveys is that examinees do not exactly know what is being measured or observed.

In terms of both planning and acting, and sharing results, the right of examinees should not be overlooked in case of an incident during examination. Ethical considerations should be part of this standard (AEA Europe et al., 2014).

Table 11
Criteria and Evidence base concerning standard 5

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Plan	Guideline 9	Guideline 1.5.3	Section 2
Screening	Guideline 9		
Transfer	Guideline 9/11		Section 9
Data Forensics	Guideline 9/11	Guideline 1.5.5	
Sharing results	Guideline 11	Guideline 1.5.5&6	Section 9

Standard 6 is '*Internet Screening'*. Not only internal criteria need to be considered in securing the examination. Organizations also should check the internet for possible disclosure of exam content (NVE & Caveon, 2014). Not many literature touches this subject, however this is also a vital part of the exam process. Exposed exam content can cause serious validity threats (COTAN, 2010; RCEC, 2015) Also these criteria are already included in the current eX:plain policy. Therefore, four criteria form this standard, namely, (1) Monitoring, (2) Reporting, (3) Evaluation, and (4) Actioning.

Table 12
Criteria and Evidence base concerning standard 6

Criteria	**Evidence base**
	NVE & Caveon
Monitoring	Guideline 12
Reporting	Guideline 12
Evaluating	Guideline 12
Actioning	Guideline 11

Standard seven, *'Security incident response'*. Literature does not provide clear standards or criteria on decision making or sanctioning in terms of security incidents. In fact, Ferrara (2017) states that decisions in this area are often best accepted by public if they are based on clearly communicated policies, guidelines and practices. Also, the organizations should take responsibility for ensuring that all personnel involved recognizes the value of proper incident response and carry out these procedures correctly and carefully (NVE & Caveon, 2016; Harris & Huang, 2017; SURF, 2014). This again relates back to a well-defined security plan. Therefore, the following four criteria are included in this standard; (1) Incident response, (2) Incident management, (3) Sanctioning, and (4) Sanctioning responsibility.

Combating fraud is primarily seen from the perspective of the organization, and although students may not be expected to actively participate in the fight against fraud, students' opinions on how to combat fraud could provide new insight in developing criteria (Malgwi & Rakovski, 2009).

Table 13
Criteria and Evidence base concerning standard 7

Criteria	**Evidence base**		
	NVE & Caveon	**AEA Europe et al.**	**SURF**
Incident response	Guideline 14/11	Guideline 1.5.3	Section 2/4
Incident management	Guideline 14	Guideline 1.5.3	Section 2/4
Sanctioning	Guideline 14	Guideline 1.5.1&2	Section 4
Sanctioning responsibility	Guideline 14	Guideline 1.5.1	Section 4

The final standard of part A is standard eight, *'Performing security audit'.* Organizations must ensure that certain criteria are set, and procedures are described for conducting a security audit (NVE & Caveon, 2014). According to Ferrara (2017), performing security audits is a vital part of a security plan. Audits should be performed to confirm if current practice corresponds with the goals and policies included in the security plan (AEA Europe, APA & NCME 2014). Several studies reported on the effectiveness of performing audits on preventing and detecting fraud (Button & Gee, 2013; Johansson & Carey, 2016; Suh, Shim & Button, 2017).

Table 14
Criteria and Evidence base concerning standard 8

Criteria	**Evidence base**	
	NVE & Caveon	**AEA Europe et al.**
Responsibility	Guideline 15	Guideline 1.5.4&7
Archiving	Guideline 15	Guideline 1.5.4&7
Security audit	Guideline 15	Guideline 1.5.4&7
Updating security plan	Guideline 15	Guideline 1.5.4&7

Grading of Part A

The multitude of guidelines and handbooks on security guidelines lack direction in determining the degree of possible security risks. In all consulted documents a fitting grading system for determining the security risk was absent. For this reason, also guidelines outside of the test security domain were consulted in order to develop a grading system for the current prototype. On this note, the COTAN rating system for test quality (2010), and the RCEC rating system for quality tests and exams (2015) provided reasonably validated and justified examples of grading systems. For example, both systems make use of a three-level classification system, namely 'insufficient', 'sufficient', and 'good' for assessing certain criteria. This system is also adopted for developing the prototype. Each of the eight standards are designed by means of a classification (e.g., rubric) including a description of what can be considered 'insufficient', 'sufficient' or 'good'. In similar fashion, a score is assigned to these classifications, respectively a score of '0' (zero), '1' (one), or '2' (two). So, a total score can be calculated for each standard, which in the end determines the level of the possible security risk.

The COTAN (2010) and RCEC (2015) models both distinguish between the value of certain criteria. This means that for each standard one or more 'base-questions' should be assessed before assessing the full standard. In case this 'base question' is assessed 'insufficient', the user automatically receives an 'insufficient' score for the entire standard. In comparison, this is not the case in grading the EDF protocol. In order to ensure valid decisions, scoring and grading must be separate steps (AEA Europe et al., 2014). The prototype should be able to evaluate the quality of current practice before determining the security risk.

The labels for determining the possible security risk are; 'low', 'medium' and 'high'. The aforementioned total scores will determine the security risk. These labels were determined during consultation with the scientific advisor and manager of Xquiry. This decision was made to determine a starting point for the protocol. The feasibility and applicability of the grading system as well as these security labels had to be determined by the content experts through the interviews. An example of the grading system is provided in table 15.

Table 15
Example of the prototype's grading system

Determining security risk for Standard 8	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Content of Part B

Within the Standards for Educational and Psychological Testing (AERE, APA & NCME, 2014), a certain standard recommends active efforts to prevent, detect and correct scores which are caused by fraudulent behaviour to ensure the integrity of the test. This supports the demand for adequate standards and criteria around detection by means of data forensics. Owing to this, the second part of the prototype involves three standards. For part B the aim was to provide an informing checklist on possibilities of data forensics use in the examinations process. The evidence base is mainly formed by the Handbook of Quantitative Methods for Detecting Cheating on Tests (Cizek & Wollack, 2017), and the Handbook of Test Security (Wollack & Fremer, 2013). Most literature around data forensics report on case studies in which certain indices are applied. Hence, the conclusions drawn there are often not applicable in the current design context, which has led to more general information from the aforementioned handbooks for the development of these standards.

The first standard is '*Detecting pre-knowledge and item compromise'*. Three criteria form this standard, namely, (1) pre-knowledge, (2) compromised items and/or people, and (3) obtaining exam content from an inside source. According to Eckerly (2017) and Sinharay (2017a), benefitting from pre-knowledge is a form of fraudulent behaviour. Pre-knowledge is considered a form of fraud, although everyone is expected to have some degree of pre-knowledge when entering an exam. It is considered fraud when test takers study compromised test items, either received or bought from peers or obtained through the internet illegally (Ferrara, 2017). The direction of the standard as well as the criteria were adopted from section IIb of Cizek and Wollack's handbook (2017), furthermore these criteria are supported by literature (e.g., Eckerly, 2017; Sinharay, 2017a; 2017b). Item compromise occurs when the performance of an item changes over time (Zara, 2006), this could be the result of pre-knowledge. Eckerly (2017), states that examinees can gain pre-knowledge from a variety of different sources of item compromise. According to Sinharay (2017a), benefitting from pre-knowledge during educational assessments is a major type of fraudulent behaviour. Several methods are proposed in literature to detect pre-knowledge. Belov (2016) for example suggested the 'posterior shift statistic'. By the same token, Sinharay (2017a) suggested a 'likelihood based test' statistic for detecting pre-knowledge.

The second standard is '*Detecting test score similarity and answer copying'.* Here, also three criteria determine this standard, (1) response similarity (2) answer copying, and (3) colluding with others. The direction of the standard as well as the criteria were adopted from section IIa of Cizek and Wollack's handbook (2017). Unusual similarities in responses between examinees or aberrant response patterns, which can be seen in test data, are types of irregularities that can indicate potential fraud (Zopluoglu, 2016). Collusion among individual examinees can also be potentially detected by using similarity statistics (Maynes, 2017), or by analysing response times (van der Linden, 2009). Although, similarity statistics and answer copying statistics are related, similarity statistics provide a means of detecting general forms of collusion that are not as easily detected by answer-copying statistics (Maynes, 2017). For this reason, both statistics are adopted in this standard. However, the applicability of these statistics can be limited depending on the type of exam. For example, with a computerised adaptive test, in which exam items are selected based on the users' skill level, or with linear-on-the-fly testing in which items are randomly assigned, these statistics connot be applied.

The third standard is *'Detecting unusual gain scores and test tampering'*. In the prototype five criteria were included, namely; (1) high response time, (2) answer changing behaviour, (3) harvesting, (4) group success rate, and (5) individual success rate. Input for the development of the prototype criteria was adopted from section IIc of Cizek and Wollack's handbook (2017), and was supported by additional

literature. Several studies report on the possibilities of using response time to detect aberrant behaviour, for indicating possible fraudulent behaviour (e.g., Impara et al., 2005; van der Linden & van Krimpen-Stoop, 2003; Marianti et al., 2014; Plackner & Primoli, 2012; Qian, Staniewska, Reckase & Woo, 2016).

On a different note, erasure analysis is getting more attention in the literature (Sinharay, 2018; Sinharay & Johnson, 2016; Wollack, Cohen, & Eckerly, 2015; Wollack & Eckerly, 2017). Erasures refer to the possibility of changing you answer on a test for example. In a CBT environment the possibility of changing answers may support fraudulent answer changing. Research on detecting changing behaviour is mainly focused on a group level (Maynes, 2013). Analysing patterns of erasures can lead to the possible detection of test tampering (Sinharay, 2018; Sinharay, Duong & Wood, 2017). A separate section of the Standards for Educational and Psychological Testing is focused on erasure analysis. This section includes recommendations on analysis of erasure pattern to detect possible irregularities (AERE et al., 2014). Test tampering can be analysed on both the individual as well as the group level. The 'Erasure Detection Index' (EDI) suggested by Wollack et al. (2015), is used on individual examinees to detect fraudulent erasures, and is based on item response theory. In similar fashion, to detect erasures at the group level (e.g., group, location, proctor) Wollack and Eckerly (2017) extended the EDI.

Step 3- Validating the EDF protocol standards

The content of the prototype was validated by means of seven semi-structured expert interviews. The interview is divided into four categories; category one focused on general questions concerning the protocol (n=7), category two focused on questions concerning the protocol content (n=4), category three related to the grading of the protocol (n=5), and category four focused on the data forensic standards (n=6). Below, the results are discussed per category together with an overview of corresponding statements from the interviews. In some cases, multiple statements were included from a single expert on a topic. In Some cased the statements made by the content experts were shortened because of readability. The full statements, both in English and Dutch, as well as the full interview transcripts are included in the appendixes (Appendix F, and E respectively).

Table 16

Overview of statements per interview category and question

Question	Category	N
What were your first impressions after reading through the EDF-protocol?	General: first impression	6
What is your opinion on the design of the protocol?	General: design opinion	4
In the current form, do you think the goal of securing the process of examination by using this protocol is feasible?	General: feasibility	5
Currently there are eight standards. Are these standards sufficient to describe the process of examination?	General: Current content	6
Do you know of any comparable guidelines or manuals, which have the same goal?	General: Comparison	4
Ideally, what should this protocol (aimed at prevention and detection of exam fraud) be able to do?	General: Ideal protocol	5
Would you like to use this protocol yourself or recommend it to colleagues?	General: Recommendation	6
For each standard, can you explain if and why it makes sense that this standard is included in the protocol?	Content: Standards	58
For each standard, are all underlying criteria clear?	Content: Current criteria	2
For each standard, are all underlying criteria complete?	Content: Criteria Missing	3
Are all underlying criteria equivalent?	Content: Equivalency	3
Currently, there is an insufficient-sufficient-good grading system. What is your opinion on this?	Grading: Grading system	5
How exhaustive should these rubrics be according to you?	Grading: Concreteness	5
The scoring of the rubric is currently 0-1-2, what is your opinion on this?	Grading:	4

Are low-medium-high security risks realistic labels?	Rubrics Grading: Security labels	4
Where lie the boundaries between these labels?	Grading: Security assessment	5
In the current form, do you think it is feasible to provide an informing checklist on possibilities of data forensics use in the examinations process?	Data Forensics: Feasibility	2
Can you explain for each standard if and why it makes sense for you that this standard is included in the protocol?	Data Forensics: Current standards	2
[Currently there are 3 standards describing data forensics]. Are these standards sufficient to describe the possibilities?	Data Forensics: Completeness	2
Are there fraud types missing?	Data Forensics: Types of fraud	3
Are there data forensics indices missing?	Data Forensics: Missing indices	3

General questions of the protocol

In the first category, seven questions were asked. An overview of the questions is shown in Table 16. The statements from the experts are categorized per question if this was possible. Table 17 provides an overview of the statements on the first category of the interview. In case statements showed a high degree of similarity, they were merged in the overview to present a clear structure of the statements.

Table 17

Overview of expert statements on the general category

General question	**Statements:**	**N**
First Impression	The protocol looks clear and applicable	4x
	The protocol looks well-thought-out and manageable	3x
Design	The design of the protocol looks clear / intuitive	4x
	The design of the protocol is currently not very sexy	2x
Feasibility	Securing the process of examination by using this protocol is feasible	5x
	There will always be a chance of fraud, but minimalizing the chance of fraud is feasible	1x
Current content	The current standards seem to be sufficient	6x
	Currently criteria on responsibility and integrity are missing	1x
	I would have the assessor as a separate standard	2x
Comparison	Not to my knowledge. But we have our own internal documents regarding security	3x
	I see duplications with what we have in our manuals	1x
Ideal protocol	Ideally, this protocol should initiate awareness	4x
	Ideally, this protocol provides users with insight into possible security gaps	2x
Usage	I would use the protocol myself or recommend it to colleagues	5x
	I would use the protocol, provided that the protocol would be made more explicit	1x

First, the experts were asked about their first impression of the protocol. Four experts indicated that their first impression was that the protocol looked clear. They gave the following arguments; "*I recognize a lot of things from which I think it is super valuable",* and "*In practice you can see a lot of protocols which are often too difficult so they won't get accepted".* Also, one of the experts already performed an audit this the content of the prototype (e.g., "*Through the protocol I have received confirmation that our current practice is good"*).

Only four experts were asked about their opinion of the prototypes' design. A striking similarity in the answers was that two out of these four experts indicated that they would like the protocol to be 'more sexy'. To illustrate with a quote, "*Functional, but it may be a bit more sexy in terms of design"*. Other statements contained terms like 'easy to scan', 'it looks well-cared for', and 'it looks clear'.

The third question in this category referred to the feasibility of securing the process of examination by applying the protocol. All five experts whom were asked this question, indicated that the current protocol could be a very useful tool for this purpose. To illustrate, "*Yes, because every process step has also been mentioned, for each step it is possible to describe if it can be scored sufficient. So that seems fine to me".* An important remark that was made, was that there will always be a chance of fraud. This remark was also already found in the literature (SURF, 2014), and therefore highlights an important consideration for the final protocol.

The fourth question in this category refers to the current standards. All experts (n=6) consider the current standards of part A to be sufficient and useful in terms of adequately securing the exam process. To illustrate with a statement, "*They are about what is needed for fraud prevention and detection. They are relevant*". Yet, some experts emphasized that additions could be made to the current standards. "*I would just look at the exam process, if you follow those steps I would have the assessor as a separate standard."*, and "*I am still missing the role of an assessor*".

For the fifth general question, the experts were asked if they knew of any comparable guidelines or manuals which also aim to provide security standards. This question was asked to four experts, of which three indicated that they could not think of any (e.g., *"I honestly cannot think of that", "Not to my knowledge").* However, they indicated that there are some internal documents in terms of security of parts of the exam process. The fourth expert stated that everything was already well described in internal handbooks, stating *"we have everything in manuals which we keep up-to-date. So, if you look at the exam development process, there are a lot of these steps that relate to safe storage, so I see duplications with what we have in our manuals and in our own quality management system".*

Several experts were asked what the protocol should be able to do ideally. Four out of five experts indicated that the goal for the protocol should be to initiate awareness through application. To illustrate, *"The most important function that the protocol has for me, is awareness."*. One expert stated that the focus of the protocol should be to flag possible security gaps (e.g., "*this protocol should be able to, in my view, provide the users of the protocol with insight into possible security gaps".*

Finally, the experts were asked if they would like to use the EDF protocol themselves or either recommend it to colleagues. Five out of six experts indicated that they would like to use it themselves or recommend it to colleagues. One expert indicated that the protocol should embed more concrete examples instead of just directions before it can be applied. These assenting statements were supported by arguments such as; "this protocol is much more workable compared to other available guidelines, because it nicely divided into parts", and "I think in general the risks are underestimated". Other statements referred to the possibility to make people aware of security risks.

Standards & Underlying criteria

The second category of the interview involved four questions, aimed to discuss the content of the prototype. The list of questions is shown in Table 16 or appendix D. An overview of the most important statements on the content is provided in Table 18. These are then further explained in text.

Table 18
Overview of expert statements on the content category

General question		Statements:	N
Standards:	Std. 1	The inclusion of this standard makes sense	4x
	Std. 1	Include more detailed information	1x
	Std. 2	The inclusion of this standard makes sense	5x
	Std. 2	Awareness should be part of this standard	1x
	Std. 3	The inclusion of this standard makes sense	4x
	Std. 3	Currently it is mainly operational, add a certain level of awareness	1x
	Std. 4	The inclusion of this standard makes sense	5x
	Std. 5	The inclusion of this standard makes sense	2x
	Std. 6	The inclusion of this standard makes sense	4x
	Std. 6	Maybe this should not be a separate standard	1x
	Std. 6	Hacking could be an addition to this standard	1x
	Std. 7	The inclusion of this standard makes sense	5x
	Std. 8	The inclusion of this standard makes sense	4x
	Std. 8	Maybe this should not be a separate standard	2x
Current criteria		They seem to be clear	5x
		Some criteria are very similar (std. 4)	2x
Criteria Missing		They seem to be complete and clear	2x
		Examples could be included	2x
Equivalency		Most criteria are equality important	3x
		Impact should weight more than equality of the criteria	4x

As was already shown in Table 16, the first question of category two, about the current standards of the prototype, produced widely useful statements (n=30). Next, the discussion of these eight standards

is illustrated by describing the results including some statements made by the experts. What must be noted in advance, is that not every expert commented on every standard. In other words, only if there was reason to discuss the usability or the content of the standard, input was given.

All four experts whom were asked if the security plan standard had to be included in the protocol, indicated that the inclusion of a standard around a security plan was necessary (e.g., "*security plan, that is of course something you'll start with. So, this would provide a good starting point", "When it comes to high-stakes exams, and about taking assignments, safety is an indispensable factor. So, a security plan is a fundamental point").* Out of the other experts, no one had questioned the value of this standard. Although, it was suggested to include more detailed information, because the current description was broad.

Five experts commented on the second standard, the security team. Remarkably, they all give a different interpretation to the term security team. One stressed the important of appointing a security officer (e.g., "*I think that assigning these tasks to people in your organization in incredibly important, so that it is continually part of practice and it no longer only starts based on an incident")*, while another state that one person would suffice (e.g., *You just need an officer, but a whole team sounds big and often not feasible).* However, they agree that it should be clearly described how employees should be held accountable for their responsibility. Particularly, it was mentioned that this standard lacked certain direction when it comes to this point, for example "*This touches on the point that I just mentioned, about awareness. So here I would like to include the integrity scan and integrity awareness into the protocol".*

The third standard that was discussed was about exam development and maintenance. On this account, all experts agreed that it is rightfully included in the protocol. One expert called this standard useful. Another supported the inclusion by giving the following statement: "*Exam development process and maintenance, you see that parties that do not work with a well-thought-out plan sometimes hear something but have no idea how they should act. This standard is needed".* Considering the underlying criteria a few comments were made. For example, the description of the 'item bank' criterium was found to be too specific (n=1). Another expert stressed the fact that also in this standard awareness should be included in the criteria, because the current criteria were found to be focused too much on an operational level.

Standard four concerns the actual taking of the exam, 'security of examination'. A discussion of the results on this standard can be brief, as all experts indicated that this part is a vital piece of the process. To illustrate with a statement: "*Exam security, of course, we do not have to say much about it. I think that makes sense. If the examination is not secure, you do not have to organize the rest*". In terms of the criteria, an expert indicated the following, "*I cannot think of any indicators that can be added, period".* However, two experts indicated that the description of two criteria within this standard was similar. This is discussed further in question two of this category.

Three experts commented on the inclusion of the 'security of results' standard. Two statements indicate the value of the inclusion of the standard; "*Security of results, I think it's good, because exam program also has a threat even though the exam is completely reliable*", "*The nice thing about this is that with a number of very crisply formulated indicators you seem to be covering the whole aspect, in this case results"*. Also, one expert suggested to make it clear in the descripting who would be responsible for these actions. During the interviews no indications were discussed of this standard being not meaningful.

Discussing standard six, 'internet screening' provided some interesting insights. First, an expert questioned if this should be a separate standard, stating "*I only wonder if you should do internet screening in a separate topic. You could also include this in security of examination*". Similar to earlier statements surrounding 'impact', also for this this standard an expert stressed the fact that this should be considered when including internet screening as a separate standard. To illustrate, "*As if you can read my comment. Because I have indeed written down 'depends on stake and secrecy'. In our organizations, for some products we do this regularly, but for other products we do not do this.* There was also a suggestion of including a new criterium, namely 'hacking' to test security. This idea came from the following statement: "*Is hacking in here? Would that be part of this. In other words, you also actively look for possible gaps in your own systems".*

'Incident response' was the seventh standard of the prototype that was discussed. Several suggestions for adjustments came from the interviews. For example, one expert suggested that responsibility should be included in the description (i.e. "*This is very procedural, and that is important.*

Although I would suggest including the responsibility. Procedures must be in order, including accountability"). In similar fashion, the description of the criteria should include direction on when this applies (e.g., before, during or after specific part of the process). On a more general level, the inclusion of this standard made sense according to all experts. To illustrate with a statement, "*Incident response is, I think, good to name separately. Because that is often the question 'yes, what happens now?' If you then discover something and then? You see that a lot now that everything is being discovered but then no idea what to do with it*". The value of including this standard is also well reflected in the following statement: "*Yes, this is generally necessary, especially in view of what we've discussed so far. Detection is one thing, but that you are also going to act on it is very important to me. You just should have your processes ready for how you deal with this. This is very important because if you do not pay attention to it, the rest does not make sense. And I think you have set the criteria for that*".

Finally, standard eight was discussed. This standard consists of performing a security audit. Here opinions were somewhat divided. Not in terms of including this step. Inclusion made sense according to the experts, to illustrate: "*Yes, I also see this as maintaining all agreements. So that you can discuss together how you get to a higher level. So, this is important to me. I had no further comments on this, except that it is very important*". However, not all experts were convinced that this should be a separate standard. For example, one expert suggested the following "*The security audit could be included within the security team"*. Several experts stated that they had no comments on this standard.

Clearly, the discussion of the prototype standard offered some interesting insights. However, more questions were asked considering the content. Secondly, the question was asked if the underlying criteria were clear. The experts (n=5) in their own words indicated that de majority of criteria were clear (e.g., "*I tend to start looking for what I am still missing. But that is why I cannot think of what would be missing").* Two experts noted that there were two criteria with a rather similar description. To illustrate, "*what is not entirely clear to me was the use of 'unauthorized materials'. This comes back twice. I did not quite understand that. At proctoring and use of materials",* and *"As far as I am concerned, there is a doubling in 'use of materials'. I did not know what the difference was in that".*

Third, the experts were asked if the set of criteria was complete. Several experts indicated that this was the case in their opinion (e.g., *"in my view they are complete").* Also, one of the experts commented of the functionality of the protocol in terms of the number of criteria by stating "*What repeatedly strikes me is that the number of indicators is limited. And that's nice. I have also come across procedures where you had to go through pages with all sorts of indicators. The nice thing about this is that with several very crisply formulated indicators you seem to be covering the whole aspect, in this case results. That is the power of this model, which makes it very manageable".* To describe the experience of working with the protocol by practitioners, this statement is valuable. However, on a more critical note another expert expressed the preference of including examples into the criteria. This comes from the following statement: "*What I think of is practical examples, no data sharing, no information on a stick that kind of things come to mind. That is currently implicit, I would make it explicit. I think people are very sensitive to examples".*

The final question in this category related to the 'weight' of the criteria within each standard. To give an example concerning standard one; is having a security plan, with goals (criterium 2) equally important as evaluating it on a yearly basis (criterium 4). An important remark here is that in most interviews, it appeared that the criteria were not looked at, at a very deep level of detail. Several experts indicated that most of the criteria were equally 'heavy', however an important condition was that the impact should be considered. This might result in inequivalence. To illustrate with a statement from one of the interviews "*You should be able to deviate consciously on parts, I think, and it gives you a picture of what you should focus on when completing. So, I do not expect points. It is not a hard science, so you cannot say that you score enough points. If you want to improve somewhere, you can also measure improvement with this. Giving weight to all parts would not be a goal for me for this list (Protocol). It is about becoming aware, so if you go too deep into a weighting, I think you'll miss your goal".*

Grading of the protocol

The third category of the interview contained questions about the grading system of the protocol. Five questions have been asked (see Table 16 or appendix D). Like the previous categories, the statements are categorized per question. Table 19 provides an overview of these results and are subsequently discussed.

Table 19
Overview of expert statements on the grading category

General question	Statements:	N
Grading system	The current grading system is relevant and good	4x
	Add the option 'not applicable'	1x
	An insufficient or sufficient score would already be sufficient	1x
Concreteness of the rubric	Being more concrete may lower the usability of the protocol	5x
Rubric scores	The current scores have added value	2x
	Scoring should be interactive when 'not applicable' is included	1x
	An insufficient or sufficient score would already suffice	1x
Security labels	These terms fit, and are realistic	4x
Security assessment	Determining the security risk depends on the impact of an insufficient score	4x
	When a criterium is scored insufficient, a high risk is fine	1x

First, the experts were asked about their opinion on the current grading system (e.g., insufficient, sufficient, and good). Five experts answered this question, in which four indicated that they find this system relevant or even good. Examples of statements were "*For the purpose of checking whether a part works, those three categories are good",* and *"That seems relevant to me. This gives you the opportunity to show growth*". One of these experts also suggested to add the option 'not applicable', stating that in some cases a criterium may not be relevant, while this is not yet reflected in the current system. The remaining expert expressed doubt whether a 'good' level was needed. Providing the following argument, *"Above all, I must be able to say that the matter is being done correct. So, I can imagine that a 0 or 1 score would suffice*.

Secondly, the question was asked how exhaustive the rubrics of the protocol should be. The tone of the experts was clear and unanimous in stating that being more concrete will have a negative impact on the usability of the protocol. The rubrics should also focus on impact when talking about possible security risks. This question yielded several very useful responses. To illustrate, "*I also work with rubrics. It can be tough. There is no correct answer to this question, so you fall back to 'keep it simple'*", and "*If you are going to make the criteria more concrete you will lose in terms of usability*. *I would say, go out there and use the protocol, see if it works. Afterwards you would still be able to do some finetuning".*

When asked for the opinion of the experts considering the scoring of the rubric, similar responses were given as to question one of this category. For example, they said that insufficient scores would now negatively impact the scoring (e.g., *"if several criteria turn out to be 'not applicable', in the current form I will get a low score. So, then it would be nice if you can make the scoring interactive, so if I would score a not applicable to a criterium you would automatically get a different total score"*; n=1). Also, there was doubt on the added value of a good score option (n=1). Like question one, the other experts agreed with the current scores.

Next the experts were asked whether they found the low, medium, and high security risk labels applicable (n=4). Again, they unanimously replied that the correct labels are realistic and applicable. Among others, some statements were: *"I think it covers the content", "I am very much used to work with those terms. It simply fits the usual terms of risk management"*, and "*Yes, I think so, as long as you leave 'not applicable' criteria out of the grading".*

Finally, it was asked where the boundaries lie between the security risk labels. Four experts stressed the fact the boundaries between these labels is determined by the impact of an insufficient score on the security of the exam. To illustrate, "*Let me put it this way, some criteria contain a higher security risk impact than the other. And I think it is important to include that in you weighing", "chance versus effect, so how big is the chance the content could become known".*

Standards on data forensics

The final category of the interview involved five questions on the topic of data forensics (see Table 16 or appendix D). The psychometricians among the expert group were asked to give input on this topic (n=2). In some cases, other experts commented of these questions as well. An overview of the most important statements on this category is provided in Table 20. These are then further explained.

Table 20
Overview of expert statements on the data forensics category

General question	Statements:	N
Feasibility	Yes. In the sense that it is aimed to provide information	1x
	Currently, this part is quite compelling	1x
Current standards	Currently, the classification is not identical	2x
Completeness	The question is whether it should be standards	1x
	The question is whether this is logical	1x
Types of fraud	No real fraud types are missing	1x
	Collusion is broad	1x
	Correcting by an assessor is not described within de fraud types	1x
Missing indices	No, obvious analyses are missing	3x

First the experts were asked about their impression of 'part B' of the protocol. Particularly, about the feasibility in terms of providing an informing checklist on possibilities of data forensics use in the examinations process. The statements made here point to important considerations for adjustments. To illustrate with a quote, *"In the sense that it is aimed to provide information. And with detailed examples of cases it can be just that. The sequence of fraud detection starts with a signal after examination. This offers you the opportunity to look at the data. I think this is essential, if you would go at it the other way around, it is probably a difficult issue. Can we simply look at the data without signals? The latter may require a different protocol".* Another statement showed that the current form is rather compelling.

Then the experts were asked if the current three standards made sense in their opinion. Both stated that this was not the case. They gave the following arguments; *"They are very similar, so the question is whether the distinction is because of the behaviour, or because of what the data can tell you. Currently I have the idea that it is a combination of the two",* and "*well the third I did not quite understand. The first one is about pre-knowledge, so everything is about unauthorized knowledge. I understand that, so that makes sense. Especially with adaptive tests with larger item banks. Although this is not necessarily fraud".*

Subsequently, the question was asked if the current standards were sufficient in describing the data forensics possibilities. Based on the obtained statements, it cannot be said with certainty that this question has been answered. However, there is room for improvement, based on the following statements; *These are things you can do, but whether that really should be standards, that is the question. For example, what I would worry about is that the statistics do not have a lot of power",* and "*I understand what you're saying, but I don't know if I find it logical, because you want to set a standard for how to detect, then I can imagine that you would want to detect pre-knowledge, but the third standard is essentially different, because then you say something about the effects. I think that is a difficult one, because the first is a kind of fraud, the second one is in fact a kind of fraud, but the third is a consequence of the type of fraud".* Despite not really answering the question, both statements were included because it clearly indicates that the current standards raise question for the readers.

The fourth question in this category refers to any missing fraud types. Several comments were made on this topic. However, no statements were made considering missing fraud types that can be directly flagged by means of data forensics that weren't already included. Yet, some experts emphasized that additions could be made: "*What I saw myself, correcting by an assessor. What I hear more often, identity fraud. This is often not the case when the lecturer is responsible, but it still is a type of fraud. I did not see these things"*, and "*Well, colluding with others is broad. That can mean anything, but what it not says now is looking directly at the work of someone else*".

The final question of this category was if there were data forensics indices missing. One of the experts simple stated *"no",* two other indicated that all the 'mainstream' analysis are currently included. However, several interesting considerations emerged. For example, "*These should be in your security plan, but you also really have to act on these things. So, the way you act should also be part of the cycle,* and "*I would not screen everyone with these kind of indices*".

To conclude, the third step of the research yielded several valuable statements made by the content experts, which have been incorporated into the final version of the EDF protocol. The way these statements are embedded in the protocol and the arguments for inclusion are described in step four.

Step 4- Adjustment of the prototype and final EDF protocol

After carefully reading the interview transcripts, an overview was made of the most important statements from the interviews. These insights were used to adjust the prototype to finalize the EDF protocol. The considerations if to make some adjustments, were based on the number of statements from the interviews that focused on the same standard or criteria of the prototype. In some cases, the content experts preferred certain changes, however these preferences were sometimes not widely supported (e.g., by other expert statements), or the suggestions were not in line with the goal or functioning of the protocol. Therefore, not all statements resulted in adjustments to the final protocol.

The interview statements were summarised into three categories. The first category describes adjustments based on statements referring to the protocol in general (e.g., "include possible evidence in the protocol"). The second category include adjustments referring to the content (e.g., "include awareness in the protocol"). The third category include grading adjustments (e.g., "add the option 'not applicable"). The EDF protocols' body of content is shown in figure 5. The final version of the full EDF protocol is included in appendix G.

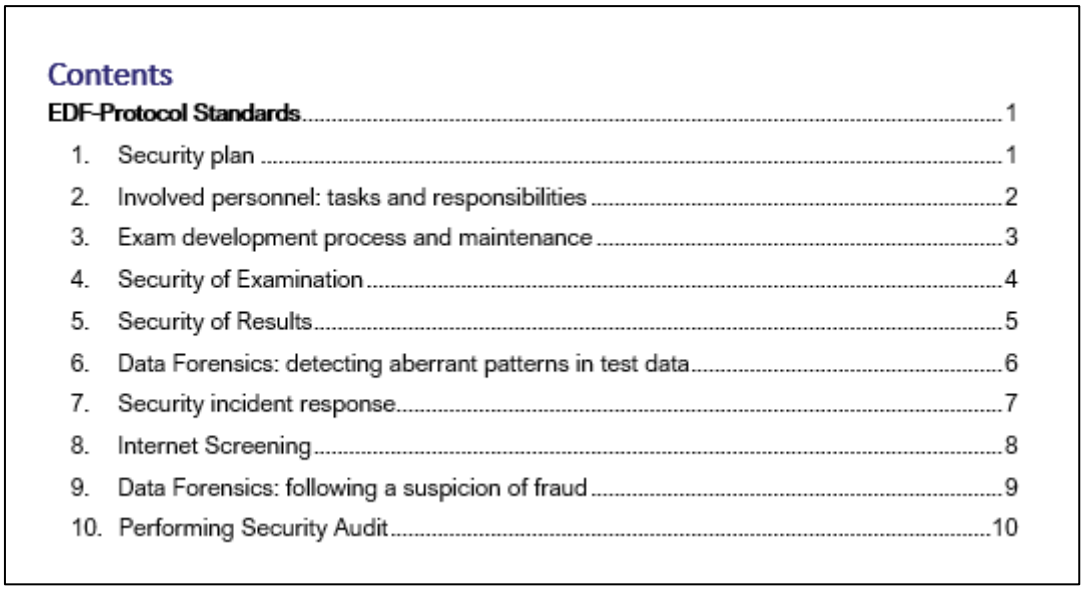
Contents

Figure 5. Overview of the EDF protocol content.

General protocol adjustments

The first adjustment made, was that there no longer is a distinction between part A and part B. After statements from several content experts, the three data forensics standards have been revised into two standards, and hereafter included within part A. Thus, resulting in a set of ten standards concerning security of the examination process. The first data forensics standard (standard 6), describes several criteria around detecting aberrant patterns in test data. The second data forensics standard (standard 9) include criteria aimed for handling a suspicion of fraud or misconduct. Subsequently, these two data forensics standards now have the same grading system as the other standards. These adjustments have been made to make the EDF protocol more fluid in general and the content more consistent.

The second adjustment, was the introduction of an evidence table for each standard (figure 6). This adjustment was based on two categories of statements. First, this table offers the opportunity to gather concrete insights per standard on how each criterion is currently dealt with. Secondly, the provided evidence gives the opportunity to enter a discussion. For example, to determine potential security risks, and decision making in terms of change management. The third general adjustment, was a change in the order of the standards. They have been adjusted to make the standards more logically reflect the process of examination in a chronological way.

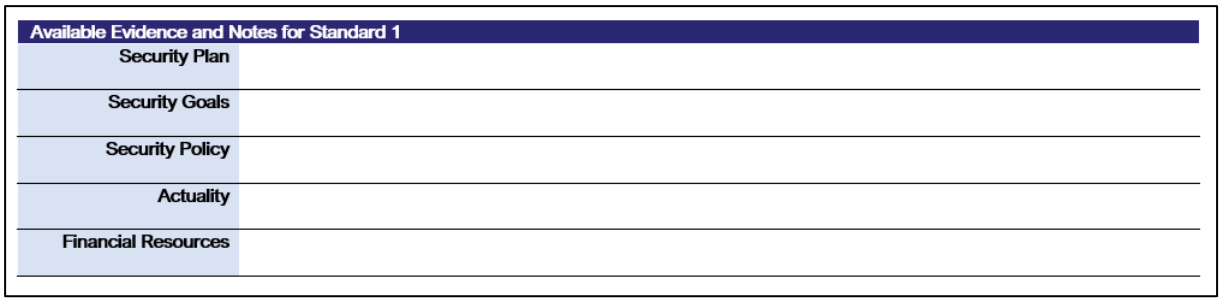

Figure 6. Example of the evidence table in the final protocol.

Content adjustments

Standard two has been revised based on several expert statements. Firstly, the name 'Security team' raised questions, and was considered too big or too vague. The image created with this standard was that a separate team should be responsible for securing the exam process. However, this was not intended with this standard. This idea was also caused, because 'human actions' were already included in other standards. However, the aim for this standard was to support awareness and to offer guidance in assessing the responsibility and integrity of all involved personnel within the process of examination. Accordingly, the name of standard two was revised into 'Involved personnel: tasks and responsibilities'. Also, the description of the four criteria have been revised to support security awareness.

Another clearly voiced point of feedback in some interviews was the lack of a standard concerning the assessor of exams or tests. The significance of including this in the protocol was made very clear, however instead of devoting an entire standard to the assessor, several criteria have been revised, and new criteria were developed to meet the statements made in this area (e.g., standard 2: criteria 2, 3 and 4, standard 4: criteria 5, and standard 5: criteria 4). An argument for doing so was that the integrity of all personnel involved was already included in the revised second standard.

Finally, several adjustments have been made in terms of naming the criteria. Reason for these adjustments were not always found in the interview transcripts, but were for example based on the fact that the original naming of some criteria did not fully represent what a criterion aimed for. Therefore, adjustments were in some cases necessary to better indicate the direction of these criteria. In one case, however, two content experts rightly pointed to the fact that criteria one (Proctoring) and four (Use of materials) of standard four, of the prototype, aimed to measure the same. Namely, the use of unauthorized materials. As a result, the name and description of the latter was revised. An overview of the revised criteria is presented in table 21.

Table 21
Overview of revised criteria.

Final protocol content	Prototype criteria	Final Protocol criteria
Standard 2- criterion 2	Security Team	Exam process member
Standard 2- criterion 3	Team responsibility	Responsibility
Standard 2- criterion 4	Team competency	Competency
Standard 4- criterion 2	Examination	Identification
Standard 4- criterion 3	-	Instruction
Standard 4- criterion 4	Planning and Acting	Plan and Act
Standard 4- criterion 5	Use of materials	Reporting
Standard 5- criterion 4	Data forensics	Assessor
Standard 8- criterion 4	Actioning	Act

Grading adjustments

In all interviews, on various topics, statements were made about the risk of drawing conclusions by means of the rubrics could be risky, especially considering the impact these conclusions might have. In the prototype the impact of the assessment was not clearly reflected in the criteria when considering assessing a diversity of exam programs. Therefore, several adjustments have been made to make the protocol even more manageable in terms of grading. First the rubrics have been revised. In the prototype all levels of grading (e.g., insufficient, sufficient and good) had a description. To focus on what is sufficient, only a clear description of the 'sufficient' level was now included in the rubric. The

descriptions of the other levels have become fixed, namely: (0) Insufficient: the described criteria are not met; (2) Good: the criteria are amply met/demonstrates hoe this is acted upon. Because they now have a fixed character they are excluded from the rubrics and included as a note under each standard, see figure 7.

Secondly, a new grading option was introduced, the option 'Not applicable' has been included. This adjustment is based on comments from experts whom stated, 'I understand that you've included this criterion, but for me this would not apply'. In the prototype, there was no way of indicating applicability of certain criteria. Thirdly, a minor change was made in terms of usability. In the prototype the awarding of a score was open. This could be done, for example, by filling in an 'X' by hand. In the final version blocks have been added, when clicking a particular block an 'X' will automatically be applied. This makes the protocol slightly more user-friendly and more intuitive.

Criteria	Description	n.a	0	1	2
Security Plan	Security plan; [1]exists as an internal document approved by the management, and [2]is made available to all personnel involved in the process of examination	☐	☐	☐	☐
Security Goals	A mission statement on security goals is present *Goals include at least:* [1]an aim towards preventing disclosure of exam content as much as possible, and [2]a statement of how security is integrated with practice.	☐	☐	☐	☐

Scoring:
(n.a)Not applicable = The described criteria does not apply to this exam
(0) Insufficient = The described criteria are not met
(1) Sufficient = The described criteria are met
(2) Good = The criteria are amply met / demonstrates how this is acted upon

Figure 7. Excerpt of the content and grading in the final version of the protocol (standard 1).

The final adjustment in the grading category refers to the tables which help determining the security risk for each standard. In the prototype these tables included three levels of risks (e.g., low, medium or high security risk). However, based on the statements concerning the impact for the exam, these levels have been revised. The table for determining the security risk now describes two levels instead of three. The new description also considers any 'not applicable' scores, see figure 8.

Determining security risk for Standard 1	
The total score on this standard is '5' or 'higher', without an 'insufficient' score (a 'not applicable' score lowers the total possible score)	→ Low security risk
Depending on the impact for the exam, one or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score.	→ Medium / High security risk

Figure 8. Example of the security risk table in the final protocol.

Step 5- Implementation of the EDF protocol

During the fifth and final step, the EDF protocol was used to evaluate and measure possible security risks within one of eX:plain' s exam programs (---). In the scope of the current study, this step has been taken to determine the actual practical value of the protocol. A consultation with the manager of the exam program was organised to validate the EDF protocol, Xquiry's scientific advisor was also involved during the consultation. The application of the protocol in the exam program was the final validation strategy for the content of the protocol. In doing so, the application of the protocol has demonstrated that it is functioning as intended, and therefore this step confirmed its added value for practice. The effectiveness of the protocol can best be described by presenting the results, hence the validation process will be discussed together with the findings and recommendations.

Standard 1: Security plan

Discussing the criteria of the first standards resulted in a '*medium/high security risk'* assessment. Currently, there is a 'Work Plan' for this exam, also adjustments are made to this document if there seems to be a reason to do so. However, this does not happen according to a fixed schedule. In addition, no checks are carried out on compliance with this work plan. Based on this approach, it is concluded that ad hoc action is taken in terms of security. For --- there appears to be no manual or guideline that focuses

on the security of the exam process. This means that there is no ‘tool-kit’ offered to all employees to be able to work 'safely' and 'responsibly'. An important remark on this finding is that this does not mean that current practice is wrong, however current practice could lead to security risks. Examples of possible risks can be: (1) that employees themselves give an interpretation of what is responsible working, or (2) the approach to responsible work is made more difficult because there is no guideline.

The findings on this standard led to the following recommendation. This standard would be eligible for a low security risk assessment when a security plan is drawn up and made available to all those involved in the exam process. For this purpose, this audit and the standards of the protocol can be taken as a starting point, so that the scope of the security plan can be limited to 3 to 5 pages.

Standard 2: Tasks and responsibilities

Evaluation of standard two shows that several issues have been adequately covered in terms of security. For example, when external people are recruited for parts of the exam process, they receive a training. In addition, all personnel (e.g., internal or external) sign a confidentiality agreement upon entering employment. For all internal personnel, there is insight into the tasks and responsibilities that they have. It was noted, however, that there is not necessarily a separation between these responsibilities. For example, it appears that employees can also carry out ‘analysis’ although they are responsible for ‘item-development’. This ‘task-flexibility’, is a risk in terms of accountability when incidents occur. In addition, currently nobody is responsible for raising awareness of secure working. The following statement was made by the manager on this topic: “*We have never looked at it like this before actually*”. These findings resulted in a ‘*medium/high security risk’* assessment.

Accordingly, two recommendations originate from these findings. First, it is advisable to use a system by which you can determine who is responsible for the execution of certain tasks. Second, because of the diversity in tasks, it is recommended to appoint a single person who is responsible for entering the conversation with all involved personnel about security awareness and integrity.

Standard 3: Exam development & maintenance

Evaluating the criteria of standard three resulted is a ‘*medium/high security risk’* assessment, even though a number of things are in order in this area. For example, the final version of the exam is stored securely. Secondly, as far as the item bank is concerned, exam items are randomly drawn and there are about 4 items per question number. This results in a 'good' score on this criterion. Thirdly, during an audit, it appears that it has occasionally happened that items have become known. These items were taken out of production, modified and then re-included in the item bank. Although this process is in order, nothing is described about this in a ‘security plan’, because of which an insufficient score is scored on this criterion. The development of exam items by constructors generally happens at home. This is very risky because there is no insight into the way in which the items are created, who gets to see these items, the way of storing on PC, and the security of the PC etc. It is also currently the case that all employees of eX:plain have access to the content of ---'s practical exams. This is not stored in a secure environment. This could pose a severe risk, since everyone can access this content without any form of liability.

These findings yielded a recommendation for the short term. Namely, data relating to the practical exams must be stored in a secure environment. For example, in Microsoft Teams, this ensures that these materials are only available to authorized personnel.

Standard 4: Security of Examination

This standard was assessed with a ‘*low security risk’*. This means that the measures taken in this area and the procedures are sufficient. Because everything was in order, no recommendations were made.

Standard 5: Security of Results

Standard five was assessed with a ‘*medium/high security risk’*, despite most of the criteria were sufficiently met. For example, both in the aforementioned work plan (see standard 1) and in the regulations for the examination committee, it is described that exam results will be analysed. Standard analyses and data forensics are also carried out, and these findings are shared with all parties involved. However, the last two criteria were assessed insufficient, because these criteria are not described in a

security plan, and were being addressed ad hoc. Notably, the '*assessor´* criterion was assessed insufficient, which was included after the validation of step three.

The recommendation for this standard is to describe these criteria in the security plan. Provide clear steps for addressing these criteria to qualify for a low security risk assessment.

Standard 6: Data Forensics I

Like standard four, this standard was also assessed with a '*low security risk'*. This means that the measures taken on the account of standard data forensics, and the procedures in this area are sufficient. Because everything was in order, no recommendations were made.

Standard 7: Incident response

In case of security incidents, there appears to be no manual or guideline that focuses on incident management. The responsibility for the handling of incidents during the examination is transferred to the exam committee. Any objections from the candidates will also be dealt with through this route. This part is sufficient. For any incidents during the other phases of the exam process (e.g., during item development), no clear procedures or guidelines have been established. Because of this, standard seven was assessed with a '*medium/high security risk'.*

As was indicated in previous recommendations, the description of the responsibilities and procedures in case of security incidents should be described in the Security Plan.

Standard 8: Internet Screening

The internet screening standard is assessed with a '*medium/high security risk'.* During the audit, it became clear that attention is paid to the criteria of this standard. However, the extent to which this is done is unknown. There is no fixed structure or procedure in this area. In addition, formal reports have not yet been delivered that contain findings. However, the manager stated that in the case of signals there is immediate action. For example, if an instructor says he can offer exam items for exercise. In response to this, this material is requested by --- in order to be able to determine the extent to which the examination content is compromised.

The recommendation for this standard was that this standard would result in a *'low security risk'* if an annual cycle would be developed in which research in this area is done, the findings are shared and action is taken where necessary. The reports should also be stored for reference.

Standard 9: Data Forensics II

The third standard with a '*low security risk'* assessment is standard 9, all criteria (if applicable) were assessed with a 'good' score. Again, this means that the measures and procedures on the account of data forensics following a suspicion of fraud are sufficient. Therefore, no recommendations were made for this standard.

Standard 10: Security Audit

The final standard was assessed with a '*low security risk'*, but a remark needs to be made on this account. A dedicated employee conducts annual audits for ---. The checklist used as a starting point for this audit is based on experiences from the year before. The results of this audit are stored internally for reference. These steps all seem to be in order, and they are. However, this procedure is not part of a security plan. Therefore, the remarks, and recommendation is to make this procedure part of the security plan. The current course of action is not described in a security plan; therefore, the liability cannot be checked.

In summary

To summarize, 6 out of 10 standards were assessed with a '*medium/high security risk'.* Although this is not an ideal score for the exam program, it does show that the protocol can flag security gaps in the examination process and due to the open nature of the criteria it was also possible to provide several concrete recommendations in order to limit the chances of security risks in the future. In addition, the remaining 4 out of 10 standards were assessed with a *'low security risk'*. This indicated that the standards were developed in such a way that proper security measures also get rewarded by the protocol.

During the consultation it was asked whether the manager thought the standards and the criteria were fit to evaluate the exam program. The manager replied with an affirmative answer. The second

question asked, was if the manager could think of standards or criteria that were missing. This was not the case according to the manager, no parts were missing. To be sure, the question was rephrased, the manager was also asked if other measures or steps were taken considering the security of the exam program, but this was also not the case. Although exam fraud can never be fully banned, these findings advocate the current content of the protocol, since it seemingly provides standards covering the entire process of examination.

Conclusion

This design research started on the premise of developing a set of standards, enabling practitioners to prevent and detect possible misconduct during the process of examination. In the end, the research provided a set of standards aimed at achieving a well-secured exam process as well as increasing awareness in doing so. This thesis reported on the theoretical base of developing the EDF protocol and the development considerations along the way. The prototype was being validated through seven semi structured interviews with content experts in the field of either test security or data forensics. Statements from these interviews were used to adjust in the prototype to finalize the EDF protocol. Finally, in order to determine the practical value, the final version of the EDF protocol was used to determine possible risks for the --- exam.

By means of the five design steps carried out in this study, the main research question is unambiguously answered by stating that the current set of ten standards, within the EDF protocol, provide sufficient direction and guidance in securing the entire process of examination. To summarize these standards: (1) Security plan, (2) Tasks and responsibilities, (3) Exam development and maintenance, (4) Security of examination, (5) Security of results, (6) Data forensics I, (7) Incident response, (8) Internet screening, (9) Data forensics II, (10) Security audit. Continuous application of the protocol in the future must determine whether the current set of standards and underlying criteria is sufficient. To illustrate, within this study the protocol was used for an exam program that did not have a security plan (including goals and procedures in terms of fraud prevention). Although this was well illustrated by applying the protocol, which emphasizes the usability of the protocol, we do not yet know how the protocol responds to a well secured exam program in terms of evaluating and measuring the possible security risks.

To answer the second research question, during development, several conditions have been considered to provided practitioners with the ability to act on indications of exam fraud based on these standards. By adding an ‘evidence-table’ for each standard, organizations are given the opportunity to provide concrete insights per standard on how each criterion is currently dealt with, meaning they can now include their own practice in the protocol. Secondly, it provides the foundation for an internal discussion. By doing so, security awareness is being encouraged on a personal level, and at a policy level, again, the foundation is laid for a well secure exam program. Also, the implementation of the protocol results in a ‘protocol report’, including findings for each standard as well tailor- made recommendation (e.g., short term or long term). A deliberate choice was made not to include a set of fixed recommendations into the protocol, on the contrary, these recommendations are now the result of implementation. In doing so the protocol can be used more widely in various exam programs, without compromising or limiting the quality of implementing the EDF protocol for individual exam programs.

Discussion

Establishing a tailormade protocol

The starting point for the EDF protocol was to develop a set of universal standards and underlying criteria for securing the process of examination. This would allow the protocol to be deployed more widely within different organizations or on a diverse set of exam programs. A remark on this idea however, is that although the standards describe the entire process properly, the underlying criteria often lacked a certain amount of concrete examples or conditions. This remark was also clearly voiced during the validation of the prototype (step 3). During one of the interviews it was suggested to make use of a focus group in order to establish a tailormade protocol, fit for a specific exam program. Although this would be a good suggestion if only one exam program would by audited by use of the protocol, this does not seem to be the solution for sustainable deployment of the EDF protocol within a diversity of organizations.

That is why it is recommended to first go through the entire protocol and provide evidence and arguments for the entire set of standards. Doing this, you will create a baseline, or a frame of reference for each exam program or organization in which the protocol is used. From this point, it can then be determined whether improvement is needed in current practice. The added value of this method is that each exam program can determine when conditions are truly sufficient, also considering the impact for their exam. In this way, the universal character of the EDF protocol is retained, but at the same time the various examination programs are given the tools to use the protocol in a both sustainable and concrete way, hence securing the process of examination through empirical implementation.

The practical value of the protocol

Besides providing practitioners with a set of standards and criteria for securing the examination process, the added value this protocol offers in comparison with other available guidelines lies in the possibility of assessing potential security risks based on the users' current practice. Despite this promising potential, an assessment model generally also has a downside. Namely, when using this protocol in an audit some standards or criteria may not appear to be 'fit' or even suitable because of their broad description. For this reason, it is important to recall the proper value of these standards. The EDF Protocol provides standards, underlying criteria, and the possibility to provide evidence if these criteria are sufficiently met. However, those who will apply this protocol must continue to thoughtfully examine to what extent the provided criteria are applicable, because the applicability can be bound to the context of the exam or exam process.

To illustrate, standard 3, criterion '*items*'. The description of what is considered sufficient is that the item bank should be large enough to offer multiple equivalent exams. However, determining the security risk, when this criterion is assessed with an insufficient score, is bound to the context of the exam. For example, having 120 items for a 40 items university exam is often considered sufficient. In case items would become known they can simply be replaced. Whereas the impact of compromised items would be higher in case it would involve a high-stakes exam for branch certificates. Even when it involves an item bank of 500 items. These factors should be carefully considered by the auditor when working with the EDF protocol to determine security risks.

Proper use of data forensics

An important consideration for policy making is how to proceed if there are statistical indications of misconduct based on the data forensic on the examinee, proctor or location level. When using the data forensic results, caution is strongly advised. Unless the data can be supported by more direct evidence, such as reported irregularities by a proctor, punitive actions are not directly advised. Ideally, communicating in advance that data forensics are conducted would be sufficient deterrent, however in a high stakes exam setting the risk taking might be high as well. In this respect the use of multiple indices is especially important. Also, it is important that the use of data forensics operate as deterrents and detectors rather than judge and jury.

In addition to the mentioned consideration for policymaking on indications of misconduct, also decisions on the most effective way of communicating these indications must be considered. Before examination, the communications should emphasize a positive message mentioning the use of

sophisticated data forensics to ensure that everyone is treated equally and fair during examination. While brought in a positive manner, the underlying message is that test tampering or misconduct will be detected and acted upon. In closing, care must be taken that the rights and privacy of the involved individuals is being protected.

Limitations of the study and future recommendations

To determine the scientific and practical value of this study, the limitations must also be considered. The prototype's content was discussed with a rather small sample of content experts (N=7). Also, these experts were not randomly selected. The majority of interviewees were approached because they were in the professional network of Xquiry team members. Despite the purposeful sampling and the size, an attempt has been made to approach as diverse a group as possible within this field. Reason to do so was to give experts in different area's of examination a voice in developing the EDF protocol. After the first seven interviews a deliberate choice was made to stop conducting interviews, due to overlapping expert input. There was mainly diversity in personal preferences among the experts, if these did not relate to the practical use of the protocol they were left out of consideration as much as possible. Reason to do so is because it is difficult to generalize findings which include personal experiences or opinions.

Furthermore, to make the interviews fluent and natural, not all pre-established questions have been asked during each interview. In some interviews certain input led to new questions and directions which turned out to be valuable. Another reason is that it became clear in some interviews that the experts prepared for the interview by reading the prototype's standards and grading system in general, however not reading all the criteria in detail. As a result, the criteria have not been discussed in detail in some interviews. This is an important note for the furture. When the protocol is applied in practice, it may be that the criteria prove to be incomplete. This however, was not the case during the implementation phase of the study.

There were also some risks involved in terms of the method that was used in this study. Using a semi-structured interview is a subjective method of data collection in which the researcher can steer the direction of the interview fairly easy. For this reason it would be interesting to keep discussing the content of the protocol with experts and practitioners in the field. For example by asking if the content is still up-to-date, or if new needs arise from practice which should be implemented in the protocol.

Final words by the author

In closing, the EDF protocol is a quality assurance system, aimed at the prevention (i.e., the prevention of exam fraud as much as possible in advance) and detection (i.e. by means of data forensics after examination) of misconduct in the exam process. Although exam fraud can never be fully banned, the protocol provides standards covering the entire process of examination in order to limit the chances of security risks. That is why the interaction with the EDF monitor as described earlier, is vital, because together they can flag possible misconduct and potential security gaps.

Reference list

Agud, J.L. (2014). Fraud and Plagiarism in School and Career. *Revisita Clinica Espanola. 2014,* (214) 410-414.

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.

Belov, D.I. (2015a). Comparing the Performance of Eight Item Pre-knowledge Detection Statistic. *Applied Psychological Measurement* 1–15. DOI: 10.1177/0146621615603327

Belov, D.I. (2015b). Robust Detection of Examinees With Aberrant Answer Changes. *Journal of Educational Measurement*, 52 (4), 437-456. http://www.blackwellpublishing.com/journal.asp?ref=0022-0655&site=1 doi: 10.1111/jedm.12094

Button, M., & Gee, J. (2013). *Countering Fraud for Competitive Advantage*. John Wiley & Sons, Chichester, UK.

Cassell, C., Denyer, D., & Tranfield, D. (2006). Using qualitative research synthesis to build an actionable knowledge base. *Management Descision, 44*(2), 213-227.

Clark, A.K., & Kingston, N.M. (2014). A brief history of research on test fraud detection and prevention. In N. M. Kingston & A. K. Clark (Eds.), Test fraud: Statistical detection and methodology (pp. 4-7). New York, NY: Routledge.

Cizek, G. J. (1999). Cheating of Tests, How to Do It, Detect It, and Prevent It. Mahwah, NJ: Lawrence Erlbaum Associates

Cizek, G. (Ed.). (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Publishers.

Eckerly, C. A. (2017). Detecting item pre-knowledge and item compromise: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of detecting cheating on tests. Washington, DC: Routledge

Fendler, R.J., & Godbey, J. M. (2015). Cheaters should never win: Eliminating the benefits of cheating. *Journal of Academic Ethics*, 1-15

Fendler, R.J., Yates, M.C., & Godbey, J.M. (2018). Observing and Deterring Social Cheating on College Exams. *International Journal for the Scholarship of Teaching and Learning: Vol.12:* No. 1, Article 4. https://doi.org/10.20429/ijsotl.2018.120104

Ferrara, S. (2017). A framework for policies and practices to improve test security programs: Prevention, detection, investigation, and resolution (PDIR). *Educational Measurement: Issues and Practice, 36*(3), 5-23. doi:10.1111/emip.1215

Fremer, J. (2011). Data Forensics. [Blogpost] https://www.fsbpt.org/FreeResources/NPTEArticles/articleType/ArticleView/articleId/40/Data-Forensics.aspx

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative Report, 8*(4), 597-606. Retrieved [22-01-2018], from http://www.nova.edu/ssss/QR/QR8-4/golafshani.pdf

Harris, D.J. & Huang, C-Y. (2017). *Establishing Baseline Data for Incidents of Misconduct in the Nextgen Assessment Environment.* In: Handbook of Quantitative Methods for Detecting cheating on Tests (309-322). Routledge: London.

Harzing, A. (2017, September 14). Journal Quality List 60th. Retrieved from http://harzing.com/resources/journal-quality-list

Howell, S.L., Sorensen, D. & Tippets H.R. (2016). The New (and Old) News about Cheating for Distance Educators. https://www.researchgate.net/publication/268359500

Impara, J.C., Kingbury, G., Maynes, D., & Fitzgerald, C. (2005). Detecting Cheating in Computer adaptive Test Using Data Forensics. [Paper]. National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.

Johansson, E., & Carey, P. (2016). Detecting fraud: the role of the anonymous reporting channel. *Journal of Business Ethics* 139 (2), 391–409.

Kingston, N. M. & Clark, A. K. (2014). Test fraud: Statistical detection and methodology New York, NY: Routledge.

Kranacher, M-J., Riley, R., & Wells J.T. (2010). *Forensic Accounting and Fraud Examination.* 1st Edition. Hoboken, NY:John Wiley & Sons
van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. Journal of Educational and Behavioral Statistics, 34, 378–394.
van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant responses in computerized adaptive testing. Psychometrika, 68, 251–265.
Malgwi, C. & Rakovski, C. (2009). Combating academic fraud: are students reticent about uncovering the covert? *Journal of academic ethics, 7* (3):207-221. DOI: 10.1007/s10805-009-9081-4
Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics, 39*(6), 426-451. doi:10.3102/1076998614559412
Maynes, D. D. (2017). *Detecting Potential collusion among Individual Examinees using Similarity Analysis* from: Handbook of Quantitative Methods for Detecting cheating on Tests. Routledge: London.
Maynes, D.D. (2013). Educator cheating and the statistical detection of group-based test security threats. In J. A. Wollack & J. J. Fremer (Eds.), Handbook of test security (pp. 173-199). New York, NY: Routledge
McCabe, D.L. (2005). CAI Research. Center for Academic Integrity. http://www.academicintegrity.org/cai_research.asp.
McCabe, D.L., Butterfield, K.D., & Trevino, L.K. (2006). Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action. Academy of Management Learning & Education, 5(3), 294-305.
McCracken, G. (1988). *The long interview.* Newbury Park, CA: Sage Publication.
McKenney, S., & Reeves, T. C. (2012). Conducting educational design research. New York, NY: Routledge Education.
Moher, D., Liberati, A., Tetzlaff, J., & Altman D.G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(6): e1000097. Doi:10.1371/journal.pmed1000097
Murdock, T.B., Hale, N.M., & Weber, M.J. (2001). Predictors of cheating among early adolescents: Academic and social motivations. *Contemporary Educational Psychology*, *26*, 96–115.
Musthaler, L. (2008, September 29). How data forensics help root out certification cheaters. NetworkWorld. Retrieved from http://www.networkworld.com/newsletters/techexec/2008/092908techexec1. html
Novotney, A. (2011). Beat the Cheat. *American Psychology Association, 42*(6), 54.
Oleck, J. (2008, March 10). Most High School Students Admit to Cheating. School Library Journal. Retrieved from http://www.schoollibraryjournal.com/article/CA6539855.html
Piercy, K.W. (2004). Analysis of semi-structured interview data. Retrieved from https://www.scribd.com/document/249952429/Piercy-Analysis-of-Semi-structured-Interview-Data
Plackner, C. & Primoli, V. (2012). Data Forensics: A compare and contrast analysis of multiple methods. [Paper] Presented at the 2012 conference on statistical detection of potential exam fraud in Lawrence, KS.
Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using response time to detect item preknowledge in computer-based licensure examinations. Educational Measurement: Issues and Practice, 35, 38–47.
Rousseau, D., Manning, J., & Denyer, D. (2008). 11 Evidence in management and organizational science: Assembling the field's full weight of scientific knowledge through syntheses. *The academy of management annuals, 2*(1), 475-515.
Scholten, D. (2013). Fraude onder studenten: De invloed van risicopercepties, machiavellisme en onzekerheidsvermijding op studiefraude. *Masterthesis.* Retrieved from essay.utwente.nl.
Sinharay, S. (2018). Detecting Fraudulent Erasures at an Aggregate Level. Journal of Educational and Behavioral Statistics 2018, Vol. 43, No. 3, pp. 286–315 DOI: 10.3102/1076998617739626
Sinharay, S. (2017a). Detection of item pre-knowledge using likelihood ratio test and score test. Journal of Educational and Behavioral Statistics, 42, 46–68. doi:10.3102/107699861667387

Sinharay, S. (2017b). Which Statistic Should Be Used to Detect Item Pre-knowledge When the Set of Compromised Items Is Known? Applied Psychological Measurement 2017, Vol. 41(6) 403–421. DOI: 10.1177/0146621617698453

Sinharay, S., Duong, M. Q., & Wood, S. W. (2017). A new statistic for detection of aberrant answer changes. Journal of Educational Measurement, 54, 200–217.

Sinharay, S., & Johnson, M. S. (2017). Three new methods for analysis of answer changes. Educational and Psychological Measurement, 77, 54–81.

Suh, J.B, Shim, H.B., & Button, M. (2017). Exploring the impact of organizational investment on occupational fraud: Mediating effects of ethical culture and monitoring control. *International Journal of Law 53* (2018) 46-55.

Tiemann, G. (2015). *An Investigation of Answer Changing on Large-Scale Computer-Based Educational Assessment. (*Master thesis, University of Kansas, USA).

Van Noord, S. (2018). *Detecting cheating in computer based multiple-choice testing.* (Master thesis, University of Twente, The Netherlands)

Witherspoon, M., Maldonado, N. & Lacey, C.H. (2012). Undergraduates and academic dishonesty. *International Journal of Business and Social Science. Vol. 3 No. 1.*

Wohlin, C. (2014). Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. http://dx.doi.org/10.1145/2601248.2601268.

Wollack, J. A., Cohen, A. S., & Eckerly, C. A. (2015). Detecting test tampering using item response theory. Educational and Psychological Measurement, 75, 931–953.

Wollack, J. A., & Eckerly, C. (2017). Detecting test tampering at the group level. In G. J. Cizek & J. A. Wollack (Eds.), Handbook of detecting cheating on tests (pp. 214–231). Washington, DC: Routledge.

Wools, S., Sanders, P. F., Eggen, T. J. H. M., Baartman, L. K. J., & Roelofs, E. C. (2011). Evaluatie van een beoordelingssysteem voor de kwaliteit van competentie-assessments [Testing an evaluation system for performance tests]. Pedagogische Studie¨n, 88, 23–40

Xquiry (2017). *XQUIRY Whitepaper 1.0*. Xquiry: Amersfoort.

Yee, K. & MacKown, P. (2010). Detecting and Preventing Cheating During Exams. Center for Academic Integrity, Rutland Institute for Ethics, Clemson University. Available at http://www.academicintegrity.org.

Zager, L., Malis, S.S., & Novak, A. (2015). The role and responsibility of auditors in prevention and detection of fraudulent financial reporting. *Procedia Economics and Finance, 39* (2016) 693–700. doi: 10.1016/S2212-5671(16)30291-X

Zara, A. (2006). *Defining item compromise*. Paper presented at the annual meeting of the national council on measurement in education, San Francisco, CA.

Zopluoglu, C. (2016). Similarity, Answer copying, and aberrance: understanding the status Quo. In Handbook of quantitative Methods for Detecting Cheating on tests. (2017). Routledge. New York.

Appendix A – The EDF prototype

EDF Prototype

Standard 1: Security plan

	Insufficient (0)	Sufficient (1)	Good (2)	Score
Security Plan	Security practices exist without a formal security plan	Security plan exists as an internal document, approved by the management	The security plan is available to all involved personnel	
Security Goals	Provides minimal direction and oversight on security issues	A mission statement on security goals is present Goals include at least: an aim towards preventing disclosure of exam content as much as possible	A clear mission statement on security is present and integrated with practice	
Security Policy	Policy governing security efforts is limited to general statements that may be challenging to translate into measures	Policy governing security efforts provide adequate directions for security measures Policy includes at least: Everyone who has access to the content of the exam signed an agreement which prohibits the disclosure of exam content	Policy governing security efforts provide effective directions with sufficient clarity to ensure appropriate implementation	
Actuality	The security plan has not been reviewed / revised within the past 24 months	The security plan is reviewed/revised within the past 12 months	The security plan is reviewed/revised within the past 12 months, and is discussed internally in the past 12 months	
Financial Resources	There is no sufficient budget to be able to implement the security plan and/or to solve security incidents	The security costs are included in the budget for maintenance and development of the exam. Budget are in accordance with the security plan	The budget is checked according to a yearly set timetable and adjusted if necessary	
	Total score on standard:			

Determining security risk for Standard 1	
The total score on this standard is '10'	Low security risk
The total score on this standard is '5' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 2: Security Team: tasks and responsibilities

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Security Officer	No one specifically assigned to attend to security	Security team shares security responsibility	A chief security officer is appointed	
Security Team	No formal security team exists	A security team exists, and members are authorized by the management	Security team members are authorized to develop/revise a security plan and oversee its implementation	
Team Responsibilities	No clear division of responsibilities, with clear assignments, tasks and roles present	Clear division of responsibilities, with clear assignments, tasks and roles	Team responsibilities are established by management and reviewed yearly	
Team Competency	Team members are insufficiently trained to perform security audits	Members receive appropriate training in parts of the security plan and associated security policies and procedures that are relevant to their tasks and responsibilities	Additionally, team members are cross-trained to provide backup support	
	Total score on standard:			

Determining security risk for Standard 2	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score **Advise**: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria **Advise**: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 3: Exam development process and maintenance

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Content Development	The development process takes place in a unsecured environment	All activities in the area of exam development take place in a secure (online) environment	During development of the exam, computers are used that are protected and do not have direct (open) contact with the internet.	
Exam Construction	There is no clear distinction between the various stages of exam development, intermediate products, and the final exam	There is a clear distinction between the various stages of exam development (intermediate products, and the final exam)	An administrator archives intermediates and copies during the development and ensures that they are not made available for use	
Items	The bank of exam content (item bank) offers **(…)**	The item bank is large enough to offer multiple equivalent exams, at least **(?)**	In case of incidents or disclosure, an update policy is linked to the item bank	
Disclosure	No measures are taken to minimizes disclosure of exam components during the design phase	The development process is designed in such a way that it is possible to monitor and control possible disclosure of exam content	Also, in the event of a security incident a replacement process will take effect immediately	
Storage	Final exams are stored without certain security measures	The final exam (all exam content) is stored in a secure location	Intermediates and copies during the development are securely archived in a separate location	
	Total score on standard:			

Determining security risk for Standard 3	
The total score on this standard is '10'	Low security risk
The total score on this standard is '5' or 'higher', without an 'insufficient' score **Advise**: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria **Advise**: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 4: Security of Examination

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Proctoring	Insufficient if one or more criteria under 'Sufficient' is not met	All criteria are met: - during examination control takes place on use of unauthorized materials - proctors observe candidates during examination directly (e.g., cameras or other tools)	There is demonstrably more done to keep the security risk as low as possible [?]	
Examination	Insufficient if one or more criteria under 'Sufficient' is not met	All criteria are met: - A correct and conclusive identification procedures of examinees prior to examination take place - A list is kept with individuals who are excluded from examination - The proctor informs examinees of the fact that security measures take place (e.g., data forensics, observations) - There is a procedure for reporting deviations in exam management - There is a contact option for examinees to report suspicious activities before, during or after the examination	There is demonstrably more done to keep the security risk as low as possible, like: - Proctors and other officers involved in the examination took part in the education trajectory - The exam is administered where the education trajectory takes place	
Planning and Acting	During examination security practices exist without a formal security plan or not at all	Appropriate security interventions are described and come into effect if the situation gives reason to do so	Proctors and other officers involved in the examination process are trained for implementing the security plan if the situation gives reason to do so	
Use of Materials	No formal control takes place around the use of materials during examination	During examination control takes place on use of unauthorized materials	Both before and after examination the exam materials will be checked to ensure that they are used in accordance with the applicable agreement	
	Total score on standard:			

Determining security risk for Standard 4	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 5: Security of Results

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Plan	No plan for investigating deviations and errors in exam results exist	Developed an action plan for screening and investigating deviations and errors in exam results	Also, describing procedures for detecting fraud and for subsequent sanctioning if fraud is detected (Also see part B)	
Screening	Periodic screening of the exam results does not take place (after every examination)	Periodic screening of the exam results for possible security incidents and their effect on exam results take place (after every examination)	Actively using procedures for discovering and evaluating suspicious exam results, unusual performance, and changes is the exam *(Data Forensics)*	
Transfer	Insufficient if criteria under 'Sufficient' is not met	All relevant data (results, reports with deviations and/or suspicious activities are sent to responsible parties immediately after examination	Also, all exam material not used is safely stored and destroyed when it is no longer needed	
Data Forensics	No one specifically assigned to attend to data forensics (internal or external)	At least one security team member is assigned to analyse exam results (internal or external)	This member ensures the periodic preparation of reports with findings and recommendations (after every examination)	
Sharing Results	Sharing information regarding exam results and candidates takes place without considering policy and procedures	Sharing information regarding exam results and candidates takes place according to the established policy and procedures	Also sharing takes place in such a way that the identity of the individuals involved and (if applicable) organization is not known	
	Total score on standard:			

Determining security risk for Standard 5	
The total score on this standard is '10'	Low security risk
The total score on this standard is '5' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 6: Internet Screening

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Monitoring	Internet screening practices exist without a formal plan/ or not at all	A formal screening plan ensures regular monitoring (within the last 12 months) of the internet and other media for activities that indicate the possible disclosure of exam components or the sharing of copyright information	A formal screening plan ensures regular monitoring (within the last 6 months) of the internet and other media for activities that indicate the possible disclosure of exam components or the sharing of copyright information	
Reporting	No periodic reporting is planned	Ensures the periodic preparation of reports with findings and recommendations based on the screening	Also, reports are shared with all security team members	
Evaluation	No periodic evaluations is planned	Ensures periodic evaluation of the activities in this context (within the last 12 months)	Ensures periodic evaluation of the activities in this context (within the last 6 months)	
Actioning	No action plan exists for dealing with alleged or actual theft through the internet	An action plan exists for dealing with alleged or actual theft through the internet of exam content	Also involves steps for removing webpages or websites that reveal (part of) the content of the exam	
	Total score on standard:			

Determining security risk for Standard 6	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 7: Security incident response

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Incident Response	No clearly defined procedures in place for incidence response	Clearly defined procedures are in place that include how to report a security incident	Clearly documented procedures that include how to report and document security issues, steps for response and follow up	
Incident management	There are no incident guidelines or policies	There are guidelines and policies in place that are shared with all relevant personnel	There are guidelines and policies in place that are shared with all relevant personnel on a regular basis (at least once a year)	
Sanctioning	Suitable penalties, actions and/or consequences for each type of security incident have not been specified	Decision-making criteria, procedures and requirements have been specified which regulate sanctioning	Also, legal advice has been sought regarding the scope and fairness of the decision-making process in this context	
Sanctioning Responsibility	No person or committee has been appointed and authorized to assess accusations of exam fraud and to impose sanctions.	A person or committee has been appointed and authorized to assess accusations of exam fraud and to impose sanctions.	Also, the disclosure of imposed sanctions takes place in such a way that the identity of the individuals involved and (if applicable) organization is not known	
	Total score on standard:			

Determining security risk for Standard 7	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 8: Performing Security Audit

	Insufficient (0)	*Sufficient (1)*	*Good (2)*	*Score*
Responsibility	Responsibility for managing the security audit process is not defined	Responsibility for managing the security audit process is clearly defined	Also, a standardised procedure ensures the consistent execution of the audit procedures	
Archiving	No archive is kept of security audits	An archive is kept of security audits and the results of these audits	Also, evidence has been collected in such a way that this is legally usable	
Security Audit	No security audit for vulnerabilities/ no review of security policies completed within the past 24 months	Security audit completed within de last 24 months, based on all EDF protocol security standards	Security audit completed within de last 12 months, based on all EDF protocol security standards	
Updating Security Plan	In case of security risks no actions are taken	In case of security risks, the security plan is reviewed/revised within 12 months	In case of security risks, the security plan is reviewed/revised within 3 months	
	Total score on standard:			

Determining security risk for Standard 8	
The total score on this standard is '8'	Low security risk
The total score on this standard is '4' or 'higher', without an 'insufficient' score ***Advise***: Although all criteria score at least 'sufficient', it is advised to improve your security measures to meet 'Good' rubric descriptions, to reduce the security risk	Medium security risk
One or more 'Insufficient' score(s) on one of the criteria ***Advise***: Direct your resources towards the criteria with the 'Insufficient' score, as it forms a high security risk for your exam	High security risk

Standard 1: Detecting Preparatory Fraud Threats: Pre-knowledge and Item Compromise

Type of fraud	*Type of fraud explained*	*Detection implication*	*Data Forensics (Indices)*
Pre-Knowledge	The examinee obtains the full or a part of the exam prior to examination (e.g., exam questions and or answers).	Especially the impact for the exam is high because of the damage to the exam.	1. Gutman score + distance 2. Response time 3. Differential Person Functioning [DPF] 4. Differential Item Functioning [DIF]
Compromised items and/or people	The examinee, for example answers several questions faster than average, indicating pre-knowledge.	This easy-to-understand index is capable of detecting first signs of fraud, and therefore be of good use from a descriptive or investigative perspective.	1. Response time 2. Log-normal model for response time 3. Response Time Effort 4. DPF 5. DIF
Obtaining exam content from an inside source	The examinee obtains the full or a partial exam prior to examination (e.g., exam questions and or answers from an inside source).	This is a rare type of fraud. However, like pre-knowledge the impact for the exam is high because of the damage to the exam.	See Pre-Knowledge, with a specific aim towards proctor and/or location data

	Available Evidence and Notes for Standard 1
Pre-Knowledge	
Compromised items and/or people	
Obtaining exam content from an inside source	

Standard 2: Detecting Test Score Similarity and Answer copying

Type of fraud/behaviour	*Type of fraud explained*	*Detection implication*	*Data Forensics (Indices)*
Response Similarity	High answer similarity between examinees.	Multiple high scores within a group are expected. Influenced by type of testing.	1. Percentage of same results
Answer Copying	Copying and/or sharing the answers from another test taker. This is also possible for computerised tests.	When using randomized questions and/or answers this cheating method carries a low risk. The risk is higher with fixed exams.	1. Number of identical correct responses 2. Number of identical incorrect responses 3. Person-Fit Indices
Colluding with Others (individual or group)	The examinee is requesting or getting help from someone during examination (e.g., the proctor provides the examinee with the correct answer)	This type of fraud is not easily detected because it can occur in different ways. This is a simple index which shows the deviation from the success rate in a group compared to the long-term average. The success percentage can be analyzed per group.	1. Group success rate

	Available Evidence and Notes for Standard 2
Response Similarity	
Answer Copying	
Colluding with Others (individual or group)	

Standard 3: Detecting Unusual Gain Scores and Test Tampering

Type of fraud	*Type of fraud explained*	*Detection implication*	*Data Forensics (Indices)*
(unusual) High Response Time	The reaction time average and standard deviation per item is calculated (e.g., the examinee answers several questions slower than average, indicating help)	This easy-to-understand index is capable of detecting first signs of fraud, and therefore be of good use from a descriptive or investigative perspective.	1. Absolute response time 2. Log-normal model for response time 3. Person-Fit model 4. Response Time Effort
Answer Changing Behaviour	The examinee changes several answers during the exam. This could indicate help from others / the use of unauthorized materials	In particular, the number of WTR erasures is often regarded as indicative of fraud. These candidates have changed extremely many answers from wrong to right.	1. WTR (wrong to right) 2. RTW (right to wrong) 3. WTW (wrong to wrong)
Harvesting	The examinee is purposefully trying to memorize/record exam content during the exam to share (with others) after examination. This may involve recording devices however this increases the possibility of detection.	This type of fraud results in a high response time because of trying to memorize and hereafter failing the test in order to re-exam.	1. Score/Time Ratio 2. Percentage same response
(unusual) Group Success Rate	The success percentage can be analyzed per group. This is a simple index which shows the deviation from the success rate in a group compared to the long-term average.	The expectation is that the success rate within the group does not differ (significantly) from the long term average. If a (significant) difference is found, this may indicate two forms of possible fraud; the candidates help each other/ a proctor helps the examinees	1. Group Score + Time interactions
(unusual) Individual Success Rate	The success percentage can be analyzed per examinee. This is a simple index which shows the deviation from the success rate compared to the long-term average.	The expectation is that the success rate does not differ (significantly) from the long term average. If a (significant) difference is found, this may indicate two forms of possible fraud; the candidates help each other/ a proctor helps the examinees	1. Score + Time interactions

	Available Evidence and Notes for Standard 3
(unusual) High Response Time	
Answer Changing Behaviour	
Harvesting	
(unusual) Group Success Rate	
(unusual) Individual Success Rate	

Appendix B – Ethical Approval

COMMISSIE ETHIEK (CE) FACULTEIT GEDRAGSWETENSCHAPPEN
AANVRAAGFORMULIER BEOORDELING VOORGENOMEN ONDERZOEK DOOR CE, VERSIE 2

1. Achtergrond proefpersonen

1. Betreft het een medisch-wetenschappelijk onderzoek? NB: Medisch-wetenschappelijk onderzoek wordt in deze context gedefinieerd als 'onderzoek dat als doel heeft het beantwoorden van een vraag op het gebied van ziekte en gezondheid (etiologie, pathogenese, verschijnselen/symptomen, diagnose, preventie, uitkomst of behandeling van ziekte), door het op systematische wijze vergaren en bestuderen van gegevens. Het onderzoek beoogt bij te dragen aan medische kennis die ook geldend is voor populaties buiten de directe onderzoekspopulatie.'

Nee

2. Titel

2b. Datum van de aanvraag 27-03-2018
2a. Wat is de titel van het onderzoek (max. 50 tekens)? LET OP: Als u van het SONA systeem gebruik gaat maken, moet hier dezelfde titel worden vermeld als de titel die in SONA zal worden gebruikt. Deze titel zal ook zichtbaar zijn voor de proefpersonen (bij gebruik SONA).

Exam Fraud: A Qualitative study towards Develop...

3. Contactgegevens onderzoekers/uitvoerders

3a. Voorletters C.J.
3b. Achternaam van Ommering
3c. Vakgroep (indien van toepassing) 0
3d. Studentnummer 1754262
3e. E-mailadres c.j.vanommering@student.utwente.nl
3f. Telefoonnummer (tijdens het onderzoek): 0611354429
3g. Indien er meer dan één uitvoerder is, dan graag in het onderstaande invulblok de gegevens (voorletters/achternaam/emailadres/telefoonnummers) van alle uitvoerders van het onderzoek invullen.

nvt

4. Contactgegevens hoofdonderzoeker/begeleidend docent

LET OP: De eerst verantwoordelijke onderzoeker/begeleidend docent is verantwoordelijk voor de bij deze aanvraag verstrekte gegevens en het onderzoek als geheel en verleent (indien van toepassing) met de aanvraag in dit formulier toestemming aan ANDERE PERSO(O)N(EN) (zie vraag 3) om voornoemde onderzoek met proefpersonen uit te voeren.
Deze eerst verantwoordelijke onderzoeker is een gepromoveerde onderzoeker.
4a. Voorletters B.P.
4b. Achternaam Veldkamp
4c. Vakgroep OMD
4d. E-mailadres b.p.veldkamp@utwente.nl
4e. Telefoonnummer tijdens het onderzoek +31534893653

5. Beoogde begin- en einddatum onderzoek

5a. Wat is de beoogde begindatum van het onderzoek? 19-12-2017
5b. Wat is de beoogde einddatum van het onderzoek? 26-06-2018

6. Doel en vraagstelling onderzoek

Geef een duidelijke en voldoende uitgebreide omschrijving van het onderzoek, waarmee een voldoende ethische beoordeling mogelijk is.
6a. Wat is het doel van het onderzoek? The aim for this study is to develop an evidence based educational data forensics [EDF] protocol.
Exam fraud is a serious threat to the validity of the exam. Most examination organizations put a lot of time, money and effort into developing reliable exams. However, the time, money and effort invested into fraud prevention and especially detection is often very little. Even though this is a vital part of ensuring exam validity and security. Therefore, a practical and evidence based protocol for fraud prevention and detection is indispensable for practitioners to guarantee the quality of the exam.

This study aims to provide an in-depth understanding of standards and criteria to prevent, detect, and also act on indications of exam fraud and thereby add to the examination practice.
In writing the protocol I want to involve practitioners and content experts.
6b. Wat is de vraagstelling van het onderzoek? This research will focus on determining the requirements for developing an EDF protocol in terms of prevention and detection and will therefore answer the following research question: 1.What standards regarding preventing and detecting exam fraud need to be included into the EDF protocol?

7. Binnen welk kader wordt het onderzoek uitgevoerd?
7. Het onderzoek wordt uitgevoerd in het kader van een studie. Het gaat specifiek om een: Masterthese

8. Aard van het onderzoek
8. Wat is de aard van het onderzoek? Onderzoek d.m.v. interviews

9. Gebruik Proefpersonen uit SONA
9. Wilt u voor uw onderzoek met proefpersonen gebruik maken van SONA? Nee

10. Omvang aantal sessies
Probeer een zo goed mogelijke schatting te geven van de benodigde duur van het onderzoek.
LET OP: Het onderzoek moet worden aangevraagd in eenheden van 15 minuten. Proefpersooncredits worden toegekend per standaard eenheid van 15 minuten.
10a. Zal een proefpersoon zijn/haar deelname afronden in één of meerdere sessie(s)? In één sessie (vragen 10b en 10c zijn niet van toepassing)
10d. Wat is de totale duur van de sessie(s) in minuten? 45 a 60 minuten

11. Beoogde aantal proefpersonen, verdeling, inclusie en exclusie criteria
11a. Wat is het beoogde aantal proefpersonen?
6 (max 10)
11b. Wat is de beoogde verdeling man/vrouw onder de proefpersonen? op dit moemtn 4/2
11c. Wat zijn de beoogde inclusiecriteria? expertise op het gebied van fraude bij examinering
11d. Wat zijn de beoogde exclusiecriteria? geen expertise op het gebied van fraude bij examinering

12. Procedure van het onderzoek
12. Wat moet een proefpersoon die aan dit onderzoek deelneemt doen? Een duidelijke beschrijving van de procedure van het onderzoek (instructies aan de proefpersonen, te meten variabelen, condities, manipulaties, meetinstrumenten) is vereist.
Fysiek of via Skype aanwezig zijn op nog nader te bepalen plaats en tijden. Vooraf krijgen ze uitleg over het interview en daarin het doel, de vragenlijst en het EDF protocol om zich voor te bereiden. Ze zullen onderdeel zijn van een semigestructureerd interview. de vragen zijn gericht op het bepalen van de juiste inhoud, afstemming van criteria en de beoordeling van de standaarden en criteria.

13. Is een van de onderstaande situaties van toepassing?
n.v.t.

14. Mogelijke gevolgen van het onderzoek voor de proefpersonen.
14a. Kan het onderzoek mogelijk ongemak en/of risico's opleveren voor de proefpersonen? Nee
14b. Toelichting Indien Nee: Graag toelichten. Indien Ja: Leg uit op welke wijze het ongemak en/of de risico's voor de deelnemende proefpersonen gerechtvaardigd worden in het licht van mogelijke opbrengsten van het onderzoek (voor de proefpersonen en/of andere groepen). Leg ook uit welke maatregelen worden getroffen om ongemak en risico's zoveel mogelijk op te vangen of te beperken.
Er is geen sprake van een experiment of dat de deelnemers bepaalde handelingen moeten verrichten.

15. Wilsbekwaamheid proefpersonen
Wilsbekwaamheid houdt in dat de proefpersonen beschikken over het individuele vermogen om zelfstandig beslissingen te nemen.
Proefpersonen zijn wilsbekwaam als zij: •18 jaar of ouder (meerderjarig) zijn, en •ieder voor zich in staat zijn tot een redelijke beoordeling van het eigen belang ter zake. Volwassenen die daartoe niet in staat zijn, zijn wilsonbekwaam.(zie ook www.ccmo.nl/nl/onderzoek-bijwilsonbekwame-volwassenen)
15a. Zijn de proefpersonen wilsbekwaam? Ja

16. Leeftijdscategorie
16. In welke leeftijdscategorie vallen de proefpersonen? Meerderjarig: 18 jaar en ouder (alleen toestemming proefpersoon nodig)

17. Volledige voorlichting vooraf
17a. Worden proefpersonen (en/of ouders/verzorgers) alvorens zij meedoen aan het onderzoek volledig over doel en inhoud van het onderzoek voorgelicht, bijvoorbeeld door middel van een brochure? Ja
17b. Toelichting Indien Ja: op welke wijze? Indien Nee: waarom niet?
Via mail of telefonisch.
17c. Welke informatie ontvangen proefpersonen (en/of ouders/verzorgers) vooraf over het doel en de inhoud van het onderzoek? Uitleg over het interview De lijst met interview vragen Het prototype EDF protocol, waarop de vragen zijn gericht

18. Informed Consent
18a. Verlenen proefpersonen (en in geval van niet-wilsbekwame proefpersonen: de voogd of ouders/verzorgers) vooraf schriftelijk toestemming voor het onderzoek door middel van een 'Informed Consent' formulier met daarin informatie over doel, aard en duur, risico's en bezwaren? Het gebruik van een Informed Consent formulier heeft sterk de voorkeur! Een standaard Informed Consent formulier is te vinden op de website van de Commissie Ethiek.
Ja

19. Volledige voorlichting achteraf
19. Op welke manier vindt de debriefing plaats? Kunnen proefpersonen (en/of hun ouders/verzorgers) bijvoorbeeld naderhand nog in contact treden met de onderzoeker over het onderzoek?
Indien Ja: op welke wijze? Indien Nee: waarom niet?
Jazeker, telefonisch en via mail. Daarnaast geef ik ze aan het eind van het interview de keuze om op de hoogte te blijven van de uitkomsten van het onderzoek.

20. Afhankelijkheid proefpersonen
20a. Beschrijf de relatie tussen de hoofdonderzoeker/onderzoekers enerzijds en de proefpersonen anderzijds. Het zijn experts/ deskundigen op het gebied van fraude bij examinering / beveiliging van examinering en hierdoor de eventuele eindgebruiker van het product. Verder is er geen afhankelijkheid.
20b. Zijn de proefpersonen, buiten de context van het onderzoek, in een afhankelijke of ondergeschikte positie t.o.v. de onderzoeker? Nee
20c. Toelichting Indien Ja: op welke wijze?

21. Duidelijkheid t.a.v. terugtrekken
21a. Wordt proefpersonen duidelijk gemaakt dat zij zich te allen tijde zonder verklaring/rechtvaardiging kunnen terugtrekken? Ja

22. Beloning proefpersonen
LET OP: Alleen voor onderzoek waarbij alleen proefpersoon credits worden gegeven, kan gebruik gemaakt worden maken van het SONA systeem.
22. Welke beloning(en) kunnen proefpersonen ontvangen voor hun deelname aan het onderzoek.
Geen

23. Opslag en verwerking gegevens
23a. Worden gegevens van het onderzoek vertrouwelijk behandeld en anoniem opgeslagen en verwerkt? Ja

24. Inzage gegevens
24a. Hebben proefpersonen achteraf inzage in hun eigen gegevens?
Ja

Opmerkingen
n.v.t.

Appendix C – Interview request

Dear Sir/Madam,

My name is Christiaan van Ommering. I am an Educational Science & Technology master student at the University of Twente. For my thesis I am doing a design research on the topic of preventing and detecting forms of cheating in the exam process. The goal is to develop a set of both practical and evidence based standards.

In order to validate the theoretical and practical value of these standards, I am looking for experts with whom I can discuss the prototype in a semi-structured interview. Due to your experience and your contributions to the field of test security, your feedback would be a valuable contribution to my research. My question is therefore whether you are willing to be part of my research.

When you agree to my request for an interview, I will send you the current version of the protocol together with my interview questions and we can then schedule an appointment for the interview. The interview would take approximately 45 minutes to one hour.

I look forward to your response.

Kind regards,

Christiaan van Ommering
Student Educational Science & Technology

Appendix D – Format Interview

Introduction:

- Explanation of the research (theme Prevention and Detection of Exam Fraud)
- Research Phase (Phase 2: validating to content of the protocol)
- Nature of the interview: 45 minutes/1 hour, semi-structured, privacy and use of data

General questions of the protocol (1)	
1	What were your first impressions after reading through the EDF-protocol?
2	What is your opinion on the design of the protocol?
3	In the current form, do you think the goal of securing the process of examination by using this protocol is feasible?
4	[Currently there are 8 standards describing the exam process]. Are these standards sufficient to describe the process of examination?
5	Do you know of any comparable guidelines or manuals, which have the same goal?
6	Ideally, what should this protocol (aimed at prevention and detection of exam fraud) be able to do?
7	Would you like to use this protocol yourself or recommend it to colleagues?
Standards & Underlying Criteria: (2) *Security Plan* *Security Team [tasks en responsibilities]* *Exam development process and maintenance* *Security of examination* *Security of results* *Internet screening* *Security incident response* *Performing security audit*	
1	Can you explain for each standard if and why it makes sense for you that this standard is included in the protocol?
2	Are the underlying criteria clear? (in terms of what they should be able to do)
3	For each standard, is the set of underlying criteria complete?
4	Are all underlying criteria equivalent (equally heavy in terms of prevention)
Grading (3)	
1	Currently, there is an insufficient-sufficient-good grading system. What is your opinion on this?
2	How exhaustive should these rubrics be according to you? (be more concrete at the expense of usability)
3	The scoring of the rubrics is currently 0-1-2. What is your opinion of this?
4	Are low-medium-high security risks realistic labels?
5	Where lie the boundaries between these labels? (in terms of when do we grade a 'high Risk')
Standards on data forensics use: (4) *Item Compromise & Pre-Knowledge* *Answer Copying & Score Similarity* *Gain scores and Test Tempering*	
1	[The goal of this part of the protocol is to provide an informing checklist on possibilities of data forensics use in the examinations process]. In the current form, do you think this is feasible?
2	Can you explain for each standard if and why it makes sense for you that this standard is included in the protocol?
3	[Currently there are 3 standards describing data forensics]. Are these standards sufficient to describe the possibilities?
4	Are there fraud types missing?
5	Are there data forensics indices missing?

Appendix E – Interview transcripts

Appendix F – Translated statements of the interviews

Overview of the interview input on category 1: General questions of the protocol

Theme	Dutch quote	Translated quote
First impression	1: ik vind het heel overzichtelijk en toepasbaar. In de praktijk zie je veel protocollen en die zijn vaak te moeilijk waardoor het niet geaccepteerd gaat worden. Maar dat vind ik in dit geval wel, ja. 2: ik denk dat dit er vrij doorwrocht eruit ziet, vrij compleet. 3: Ik heb in het protocol bevestiging gezocht of ik de vragen die gesteld worden kon beantwoorden met voldoende of goed. En ik heb ook met een schuin oog gekeken of er onderdelen onvoldoende zijn, dit is niet het geval. Ik denk hierin nog een keer de bevestiging te hebben gekregen dat het deugt. 5: Nou mijn beeld is dat het zo een goed beeld kan geven. Een goed handvat, het is redelijk compact en hanteerbaar. Wel mis ik soms informatie. Als je zo'n lijst invult dan kan het soms zijn dat het vrij te interpreteren is. Maar het lijkt me een hele goede basis. 6: Ik herken een hele hoop zaken zeg maar waarvan ik denk ja dat is super waardevol daar moet je goed naar kijken. 7: Met name deel A denk ik dat dat heel nuttig is.	1: I find it very clear and applicable. In practice you can see a lot of protocols which are often too difficult so then won't get accepted. However, I believe this is the case (with the current protocol), yes. 2: I think this looks pretty well-thought-out, quite complete 3: I have sought confirmation in the protocol whether I could answer the questions asked with sufficient or good. I also looked, with a blind eye, too see if there are insufficient parts, this was not the case. Through the protocol I have received confirmation that it is good. 5: Well my opinion is that it can give a good overview. A good tool, it is reasonably compact and manageable. I sometimes miss information. If you fill out such a list, it can sometimes be that it can be freely interpreted. But it seems to be a very good basis in my opinion. 6: I recognize a lot of thing from which I think it is super valuable, you do have to take a good look at that. 7: Part A in particular I think is very useful.
Design opinion	1: Functioneel, maar het mag wel ietsje sexyer qua vormgeving. Het ziet er goed verzorgd uit. Hoe het is opgebouwd is heel intuïtief eigenlijk, zoals je in een applicatie zou zeggen. 2: Ik vind het wel helder. Ik denk ook wel dat het een hele smak werk is als ik dit zo bekijk. Misschien kan je wel gewoon zeggen is het er wel/is het er niet. Het is niet zo'n sexy onderwerp. (…) Als ik hier naar kijk zie ik 8 standaarden met elke een aantal punten erin, ik denk is het dan niet allemaal een beetje veel. Een beetje overdreven. 3: Als ik kijk naar de lay-out, dan liet het protocol zich gemakkelijk scannen. Als	1: Functional, but it may be a bit more sexier in terms of design. It looks well-cared for. It looks pretty intuitive, as you would say in an application. 2: I think it is clear. I also think it is a lot of work if I look at it like this. Maybe you could just say whether it is there or not. It is not so much a sexy subject (…). When I look at this I see 8 standards with underlying criteria, I then think isn't it a bit too much. A little over the top. 3:When I look at the layout, the protocol was easily scanned. As a manager I do not read it in detail, but it is very accessible. The categories are

	leidinggevende lees ik het niet minutieus door, maar het is heel toegankelijk. De categorieën zijn herkenbaar, ik snap waar het over gaat en er zijn steeds de drie antwoord categorieën. 5: Ja, lijkt mij prima. Het is niet iets wat je dagelijks doet. Het is een bewustzijnsscan, security. Ja het is prima. Wat ik zelf weleens gebruik is een soort spinnenweb systeem, waarbij je dan uitkomt op bepaalde thema's. Maar goed het zijn hier 8 dingen, dus het is niet wereldschokkend, dus volgens mij is het goed hanteerbaar.	recognizable, I understand what it is about and there are always the same three answer categories. 5: Yes, I think it's fine. It is not something you will do every day. It is a security consciousness scan. Yes it is fine. What I sometimes use is a spider web system, in which you then end up with certain themes. But there are only eight standards here, so it is not ground shaking, so I think it's easy to handle.
Goal feasibility	1: Ja. 2: (…) er zal altijd een kans op fraude bestaan. Je zal niet kunnen voorkomen dat er fraude is. Of je hier mee kan meten? Ik denk dat als je hier, allemaal naar gaat kijken, dat je er alles aan hebt gedaan om enigszins de fraude die je op voorhand kan verwachten, dat je die enigszins onder ogen hebt gehad en dat je kan proberen het te verbeteren. Dat denk ik zeker wel. 3: Ja, want de belangrijkste functie die het protocol wat mij betreft heeft is bewustwording. Natuurlijk zal de opsteller, dit zal jij zelf zijn, willen dat het protocol wordt gebruikt als een instrument. Maar als je het mij vraagt is er een hogere orde, namelijk dat je bewust wordt van de veiligheid, betrouwbaarheid en integriteit, zodat je met je product de hoogst mogelijke kwaliteit kan leveren. Het gaat om het maximeren van het examen of toets kwaliteit. Als het gaat om high-stakes, dan moet het niet 'zo goed mogelijk' zijn, maar moet het top zijn, klasse, optimaal. 5: Ja omdat ook elke processtap is benoemd, en dat je per stap kunt benoemen of hier aan voldaan wordt. Dus dat lijkt me prima. 7: Het expliciet vastleggen hoe je het doet of regelt, dat vond ik wel heel mooi.	1: Yes. 2: there will always be a chance of fraud. You will not be able to fully prevent fraud. Whether you can measure this? I think when you look at all these, that you will have done everything in your power to minimalize the chance of fraud that you have been somewhat faced with and that you can try to improve. I certainly think so. 3: Yes, because the main function that the protocol has for me is awareness. Of course the author, that will be you, will want the protocol to be used as an instrument. But if you ask me, there is a higher order, namely that you become aware of safety, reliability and integrity, so that you can deliver the highest possible quality with your product. It is about maximizing the exam or test quality. When it comes to high-stakes, it should not be 'as good as possible, it should be top-class, optimal. 5: Yes, because every process step has also been mentioned, for each step it is possible to describe whether or not it can be scored sufficient. So that seems fine to me. 7: To explicitly record how you do it or arrange it, I found that very nice.
Current standards	1: Nou ik heb daar van de week al eens globaal over nagedacht, maar ook toen kon ik niets bedenken. Als je naar de	1: Well I have thought about it during the last week, but even then I could not think of anything. If you are looking at the

	inhoudt gaat kijken heb je het proces aardig dekkend zou ik maar zeggen. 2: Nee 3: Die 8 standaarden zijn relevant. Die gaan over datgene wat er nodig is voor fraude preventie en detectie daarvan. En ze zijn relevant. (…) Ja, eentje, dat zijn namelijk de mensen die verantwoordelijk zijn voor de uitvoering van dit proces. 5: Ik kon zo snel niets verzinnen. Ik zal nog even in mijn aantekeningen kijken. Nee ik mis zo snel geen stappen. Je kunt nog denken aan, dit gaat uit van digitaal toetsen. Soms zie je bij digitaal toetsen dat je deze laat beoordelen door beoordelaars, en dat je daar natuurlijk ook allerlei soorten van fraude kunt zien. En dat zag ik hierin nog niet zo snel terug. 6: Even kijken, als ik kijk naar de volledigheid miste ik nog de rol van een beoordelaar bijvoorbeeld. (…) ik zou even kijken naar het examenproces als je dat volgt dan zou ik het wel als aparte standaard ook, uhm. Dat vraagt echt iets anders 7: Wat ik eigenlijk wel echt mis is een standaard voor fraude afhandeling. Er staat nergens 'indien er fraude wordt geconstateerd dan gebeurt er dit of dat'. (…) Dat is misschien nog wel op een wat algemeen niveau. Maar daar ontkom je misschien ook niet aan	content I believe the process is adequately covered. 2: No (after being asked if there are standards missing) 3:these 8 standards are relevant. They are about what is needed for fraud prevention and detection. They are relevant. (…) Yes, one, namely about the people who are responsible for the implementation of the process 5: I could not think of anything. I will take a look at my notes. No, I do not miss any standards I believe. You can still think of, this is based on digital examinations, sometimes you see exams being assessed by assessors. You might encounter all sorts of fraud there. And I did not see that in the protocol yet. 6: Let's see, if I look at the completeness, I am still missing the role of an assessor. (…) I wouls just look at the exam process, if you follow those steps I would have the assessor as a separate standard. That really requires something else. 7: what I actually really miss is a standard for fraud handling. There currently no information about 'if fraud is detected, then this or that happens'. (*here I point out that there are criteria which relate to this*). This may still be on a somewhat general level. But you might not get around that.
Comparison	1: Bij mijn weten zijn die er niet. Wat je wel ziet zijn protocollen, handleidingen, keurmerken dat soort zaken met betrekking tot het proces, waarbij men dat ook wel probeert door audits ook wel meetbaar te maken en controleert of dat nageleefd wordt. Maar echt het detecteren van fraude daar is volgens mij echt helemaal niks voor 3: Nou, dat kan ik mij eerlijk gezegd niet bedenken. Dit was al een vraag die in mijn hoofd kwam nadat ik het had gelezen. Zo had ik de vraag, hebben wij binnen de organisatie iets waardoor dit protocol overbodig is? En het antwoord daarop is nee. Dat heb ik niet. Er is wel van alles geregeld waardoor ik kan ja	1: To my knowledge they are not there. What you do see are protocols, manuals or qualifiers with regard to the exam process. Some of these also try to make prevention measurable by means of audits or checks based on what is observed. But to really detect fraud, I believe there is nothing so far. 3: Well, I honestly cannot think of that. This was already a question that came into my mind after reading the protocol. So I asked the question 'do we have something within the organization that makes this protocol redundant? The answer to that was 'No'. I do not have that. We have arranged everything so I can nod yes to the questions asked, but

	knikken. Maar zo'n protocol hebben we niet. Dus ik ben wel geïnteresseerd. 5: Niet dat ik weet. Maar wij hebben onze eigen interne werkprocessen rondom beveiliging. Interne protocollen, die zijn vrij in detail. 6: Ja wij zijn ISO gecertificeerd dus wij hebben van alles in handboeken en die handboeken die houden we ook up-to-date. Dus als je kijkt naar het examen ontwikkelproces dan zitten daar heel veel van dit soort stappen in ook voor het veilig opslaan, dus daar zie ik dubbelingen in met wat wij dan in ons handboek hebben en in ons eigen kwaliteitsmanagementsysteem.	we do not have such a protocol. So I am interested. 5: Not to my knowledge. But we have our own internal documents with regard to security. Internal protocols, which are quite in detail. 6: Yes we are ISO certified so we have everything in manuals which we keep up-to-date. So if you look at the exam development process, there are a lot of these steps that relate to safe storage, so I see duplications with what we have in our manuals and in our own quality management system.
The ideal protocol	1: Ja dit protocol zou in mijn optiek naar de opdrachtgevers, naar gebruikers van het protocol. Voor die partijen zou het protocol inzichtelijk moeten maken waar zitten de hiaten. 2: Om een bewustwordingsproces in gang te zetten. Vooral ook vanuit reputatie is het wel goed om dit te doen. 3: De belangrijkste functie die het protocol wat mij betreft heeft is bewustwording. 5: Nou je kunt alle stakeholders bewust maken van de stappen. En aansturen dat ze bewuste keuzes maken. Het is voor ons niet van belang, maar dat je bewust dingen beoordeeld 6: Dit protocol kunnen wij naast onze handboeken leggen en dan kunnen we gaan afvinken he. Je hebt ook een onderdeel audit erin zitten. Als ik kijk naar onze audit partij, die zou dit bijvoorbeeld kunnen gebruiken, maar die hebben al hun eigen protocol natuurlijk om ons te auditen en hun eigen checklist. (…) Dus om waardering aan te geven	1: Yes, this protocol should be able to, in my view, provide the users of the protocol with insight into possible security gaps. 2: To initiate an awareness process. Especially considering reputation, it is good to do this. 3: The most important function that the protocol has for me is awareness. 5: Well you can make all stakeholders aware of the steps. And directing them to make conscious choices. It is not important to us, but that you consciously judge things. 6: We can place this protocol next to our manuals and then we can start a check. You also have an audit component in there. If I look at out audit party, they could use this for example, however they already have their own protocols to audit us, and their own checklists. (…) So it can be used as a scoring guide.
Use or recommendation	1: Ja, vanwege alles wat ik gezien heb buiten dit document om hé. Het gekke is zelfs als je kijkt naar de security guideline waarop die gebaseerd is denk ik dat dit veel werkbaarder is. Hoe het opgebouwd is. Dan zie je gewoon dat het veel herkenbaarder is, dus het is niet zo'n enorme brok in één keer maar het is mooi duidelijk afgebakend. Dat maakt gewoon dat het werkbaar is. En daarnaast	1: Yes, because of everything I've seen outside of the document. The crazy thing is even if you look at the security guideline on which this is based, I think this is much more workable. How this is built up. Then you just see that it is not such a huge lump at once but it is nicely divided into parts. That simply makes it workable. And in addition, the

	de combinatie met data forensics, dat is een heel krachtig meetmiddel. 2: Ja, en dan met name aanraden. Want ik ben er zelf niet zo mee bezig binnen mijn werkzaamheden, maar wel aan onze proces regisseur. Maar ook voor mij staan er punten in dat ik zeg 'oja'. 3: Ja dit is de moeite waard. Ik wil in ieder geval met dit protocol het gesprek met mijn managers aangaan. 5: Ja, dus ik denk dat dat zinvol is ja. Die delen neerleggen bij de stakeholders die over dat onderdeel gaan. Het is een goed vinklijstje, dus ja. 6: Jazeker, omdat ik vind dat het nog steeds wel onderbelicht is wat het risico is van al die content. Als je daar niet genoeg in zit heb je echt niet door wat het risico is als er iets op straat komt te liggen. Dat het imago wordt bedreigd. Dus ik vind het heel goed als daar aandacht voor is. En als dat werkt met een protocol wat meestal wel zo is dan prima. 7: (…) het protocol mits wat explicieter gemaakt, denk ik dat het zeer zinvol is. En ik denk dat we daar als universiteit wel iets uit kunnen leren.	combination with data forensics, which is a very powerful measuring tool. 2: Yes, and especially recommend it. Because I'm not really working on this topic myself in my work. But to our process director. However, also for myself there were points in it that I say 'Oh, right'. 3: Yes, this is worth it. At least I would want to get the conversations started with my managers. 5: Yes, so I think this is useful. Involve those stakeholders who are responsible for specific parts. This is a good checklist, so yes. 6:Yes, because I think in general the risks are underestimated. If you are not fully aware of the consequences if content gets compromised. The image could be compromised. So I think that it is good if there is attention for that. And it that works through with a protocol, which often is the case, then that is fine by me. 7: (…) If the protocol would be a bit more explicit, I think it would be very useful. And I think we can learn something for it as a university.

Overview of the interview input on category 2: The protocol content

Theme	**Dutch quote**	Translated quote
Standard 1	1: Als je kijkt naar het security plan dat is natuurlijk iets waar je mee start, dus daarmee zeker met deze richtlijnen geef je denk ik opdrachtgevers een goede start 2: Als je nou goed op security plan scoort, ben je dan beter tegen fraude gewapend, ja ik denk het wel. 3: Als het gaat over high-stakes examens, en over het maken van opgaven, het drukken of invoeren van examens in computersystemen dan is veiligheid dan een onmisbare factor. '…' Dus veiligheid is een fundamenteel punt. 5: Ja ik vond het duidelijk. Maar ik zou qua securityplan daarin meer detail informatie in willen verwerken. Als ik kijk naar het plan kan dit vrij breed zijn. Het 'locken' van de pc bijvoorbeeld, of een detectie poortje, het wijzigen van je wachtwoord. Dat zijn algemene	1: if you look at the security plan, that is of course something you'll start with. So this would provide a good starting point. 2: if you'd score well on security plan, you would be better protected against fraud. Yes I do think so. 3: When it comes to high-stakes exams, and about taking assignments, safety is an indispensable factor '…' So security is a fundamental point. 5: Yes, I thought it was clear. But I would like to include more detailed information in terms of the security plan. If I look at the plan this can be quite broad. Locking the Pc for example, or a detection gate, changing your password. These are general measures and are part of your security plan.

	maatregelen en zijn eigenlijk ook onderdelen van je security plan.	
Standard 2	1: Ik denk een security team … vaak missing is. Ik vind het dedicated toewijzen van deze taken tot mensen in je organisatie vind ik ongelooflijk belangrijk zodat het ook blijft leven en het niet meer alleen op basis van incidenten gaat leven 2: Je hebt mensen, waarbij je kan zeggen een security team. Maar je hebt ook een team onder proces en maintenance, rekenkundige taaltechnici etc. misschien hoort het team daar ook wel weer tussen. Suggereer je nu misschien dat het een apart sec. team moet zijn. Want je hebt mensen die houden zich alleen bezig met het ontwikkelen van het examen. Anderen alleen met de afname etc. 3: Deze raakt dus al ook aan het punt wat ik had benoemt, van de bewustwording. Hier zou ik dus die integriteitsscan en integriteit bewustwording die zou ik daarbij willen betrekken. 5: Wat mij opviel is dat er wordt gesproken over een security team. Een officer heb je gewoon nodig, maar team klinkt als heel groot, en vaak niet haalbaar. Verderop praat je over data forensics, en dan vraag ik mij af of dat bij dit team thuis hoort. Maar dat is misschien de naamgeving. Bij ons wordt het op een manier ingericht dat dit nooit zou passen. 6: ik zie ook de rol van security officer die heb je ook al als het goed is rondom de wet privacy. Dat is dan deze security officer ten opzichte van de officer die je bijvoorbeeld hebt rondom privacy. Dus ik ben op zoek naar de rol in de organisatie van de security team en audit team dan en hoe ze ten opzichte van elkaar verhouden. '…' Oké. Dan heeft dat nog opheldering nodig wellicht. Dat mensen eenduidig snappen wat ermee bedoeld wordt.	1: I think security team is often missing. I think that assigning these tasks to people in your organization in incredibly important, so that it is continually part of practice and it no longer only starts based on an incident. 2: you have people, you call it a security team. But you also have a team under process and maintenance. Among others, psychometricians are also part of a security team. Currently you seem to suggest these are separate teams. People are concerned with separate parts of the exam process, to which team do they relate. 3: this also touches on the point that I just mentioned, about awareness. So here I would like to include the integrity scan and integrity awareness into the protocol. 5: what struck me is that we are talking about a security team. You just need an officer, but a whole team sounds big and often not feasible. Later on you talk about data forensics, I wondered is this should be part of the security team. But that is perhaps the naming of the standard. With us these thing are arranged in such a way that the criteria would not fit. 6: I also see the role of security officer, which is also directed to the privacy law right. What is the security officer compared to the privacy officer. So I am still searching for the role of the security team and the audit team. '…'. Alright, then perhaps this need clarification. So that people understand unambiguously what is meant by team.
Standard 3	1: exam development proces en maintenance daarvan zie je dat partijen die niet werken met een doordacht plan wel eens iets horen roepen maar eigenlijk geen idee hebben hoe ze nou dat deel van die keten mee moeten nemen 2: Ik zou iets toevoegen bij het examen development. Dat je een bepaald niveau van awareness hebt overal. Het is vooral operationeel allemaal. Awareness zou ik er inderdaad tussen zetten 3: Nee ik denk dat het zoals het er nu staat, dat dit wel handig is. 5: Wat je hier zegt is dat de itembank dusdanig groot moet zijn dat ik meerdere varianten moet kunnen maken. Maar dit klinkt heel erg als een oplossing, en dat	1: exam development process and maintenance, you see that parties that do not work with a well thought-out plan sometimes hear something but have no idea how they should take that part of that chain 2: I would add something to the exam development. That you have a certain level of awareness everywhere. It is mainly operational. I would indeed include awareness in the protocol 3: No, I think it is as it is now, that this is useful. 5: What you are saying here is that the item bank must be so large that I have to be able to make several variants. But this sounds very much like a solution, and that

	hangt erg van de situatie af. (geeft voorbeeld van kleutertoets). Ik vind dit punt echt te specifiek.	depends a lot on the situation. (gives example of toddler test). I think this point is too specific.
Standard 4	1: examen security daar hoeven we natuurlijk niet zoveel over te zeggen. Ik denk dat dat logisch is. Als de afname niet goed is hoef je de rest ook niet in te richten. 3: Ja ik kijk daar, ik kan geen indicatoren bedenken de nog aanvullend kunnen zijn, punt 5: Ik heb opgeschreven dat de security heel erg samen hangt met de ‘stake’ van het examen. Dus dit is niet altijd relevant. En wat mij niet helemaal duidelijk is, was het gebruik van ‘unauthorised materials’. Deze komt twee keer terug. Dat begreep ik niet helemaal. Die staat bij proctoring en bij use of materials. 6: Even kijken hoor, daar zit wat mij betreft een dubbeling in bij ‘use of materials’. Ik wist niet wat daar dan het verschil in was.	1: exam security of course we do not have to say much about it. I think that makes sense. If the examination is not secure, you do not have to organize the rest. 3: Yes, I cannot think of any indicators that can be added, period. 5: I have written that the security is very much related to the 'stake' of the exam. So this is not always relevant. And what is not entirely clear to me was the use of 'unauthorized materials'. This comes back twice. I did not quite understand that. It stands for proctoring and use of materials. 6: Let's see, well, as far as I am concerned, there is a doubling in 'use of materials'. I did not know what the difference was in that.
Standard 5	1: Security of results ik denk dat, dat goed is omdat er inderdaad er niet bij stilgestaan wordt dat je examen stelsel ook een bedreiging heeft ook al is het examen volledig betrouwbaar afgenomen 3: Het aardige hiervan is dat je met een aantal heel kernachtig geformuleerde indicatoren je het hele aspect, in dit geval results afgedekt lijkt te zijn. 5: Ik zou die activiteiten eerder koppelen aan de rol en niet aan de functie. Het wordt nu beschreven dat het onderdeel is van de taken van iemand van het security team. Bij CITO zou je dat bij een analist neerleggen of zelfs automatiseren.	1: Security of results I think that is good because it is not considered that your exam system also has a threat even though the exam is completely reliable. 3: The nice thing about this is that with a number of very crisply formulated indicators you seem to be covering the whole aspect, in this case results. 5: I would link these activities to the role rather than the function. It is now described that it is part of the duties of someone from the security team. At CITO you would put that down with an analyst or even automate.
Standard 6	1: Ik vraag me alleen af of je internet screening in een apart topic moet doen. Dat zou je ook onder security of exam kunnen laten 3: Zit het hacken dan ook hier in? Zou dat hier bij horen. Met andere woorden, ga je zelf ook actief op zoek naar eventuele commissies in je eigen systemen 5: Alsof je mijn opmerking kunt lezen. Want ik heb inderdaad opgeschreven ‘hangt af van stake en geheimhouding’. Bij ons doen we dit voor een aantal producten gestructureerd en bij een aantal producten doen we dit niet.	1: I only wonder if you should do internet screening in a separate topic. You could also include this in security of examamination 3: Is hacking in here? Would that be part of this. In other words, you also actively look for possible gaps in your own systems 5: As if you can read my comment. Because I have indeed written down 'depends on stake and secrecy'. In our organizations, for some products we do this regularly, but for other products we do not do this.
Standard 7	1: Incident respons is denk ik goed om apart te benoemen. Want dat is vaak de vraag ‘ja wat gebeurd er nou?’ als je dan iets ontdekt en dan? Dat zie je nu heel veel dat er van alles ontdekt wordt maar dan geen idee wat moet je er dan mee. 3: Deze is heel procedureel, en dat is heel belangrijk. Inclusief het afleggen van	1: Incident response is, I think, good to name separately. Because that is often the question 'yes, what happens now?' If you then discover something and then? You see that a lot now that everything is being discovered but then no idea what to do with it.

	verantwoordelijkheid. Procedures moeten op orde zijn inclusief verantwoording 5: Ja wat ik hierover opmerkte is, dit is algemeen noodzakelijk, zeker gezien al het voorgaande. Ook zoals de data forensics. Dat je het kunt detecteren is één ding, maar dat je er ook op gaat acteren is ja, dat lijkt me heel erg belangrijk. Je moet gewoon je processen klaar hebben liggen van hoe je hier mee om gaat. Deze is namelijk heel erg belangrijk want als je hier geen aandacht aan besteed heeft de rest geen zin. En volgens mij heb je de criteria daarvoor benoemd. 6: Even kijken ik had bij 7 opgeschreven, neem op wanneer dit dan van toepassing is, voor, tijdens of na het examen..	3: This is very procedural, and that is very important. Including the responsibility. Procedures must be in order including accountability 5: Yes, what I said about this is, this is generally necessary, especially in view of all the foregoing. Also like the data forensics. That you can detect it is one thing, but that you are also going to act on it is yes, that seems very important to me. You just have to have your processes ready for how you deal with this. This is very important because if you do not pay attention to it, the rest does not make sense. And I think you have set the criteria for that. 6: Just looking at what I wrote down at 7. Include when this applies, before, during or after the exam.
Standard 8	1: De security audit die zou je eventueel onder het security team kunnen hangen 3: Eigenlijk is dit een vraag die je in verlegenheid brengt, als je het niet weet schiet je dan te kort. Het gaat er vooral om dat je moet weten hoe de processen lopen om te kunnen bepalen hoe vaak je een audit zou moeten uitvoeren. Wat wij doen als we nieuwe processen implementeren bijvoorbeeld, is dat we direct na implementatie een audit laten uitvoeren. Dan heb je een soort 0-meting, vervolgens doen we het vooral naar behoefte. En dat is meestal als we een incident hebben of ruiken. 5: Ja ik zie dit ook als handhaving van alle afspraken. Dus dat je samen bespreekbaar maakt hoe je op een hoger niveau komt, dus dit is een, uh, dit lijkt mij belangrijk. Ik had hier verder geen opmerkingen over behalve dat het heel belangrijk is. En ik zie verder geen dingen die ik mis. 6: '…'Ja dat zie ik ook wel. Bij acht had ik verder niks staan.	1: The security audit you could include in the security team 3: Actually, this is a question that embarrasses you, if you do not know, you will fall short. The main point is that you have to know how the processes runs in order to determine how often you should perform an audit. What we do when we implement new processes, for example, is that we have an audit carried out immediately after implementation. Then you have a kind of base analysis, then we do it mainly according to need. And that is usually when we have an incident or smell one. 5: Yes, I also see this as maintaining all agreements. So that you can discuss together how you get to a higher level, so this is a, uh, this seems important to me. I had no further comments on this, except that it is very important. And I do not see anything else that I miss. 6: '...' Yes I do see that too. At eight I did not have anything else.
Current criteria	1: ik heb de neiging om dan gelijk te gaan zoeken naar wat ik nog mis. Maar daarvoor geldt dat ik niet zo kan bedenken wat er nog missing zou zijn. 3: Dat weet ik eerlijk gezegd niet, of ze helemaal dekkend zijn. Ik vind ze allemaal heel plausibel, wat ik moeilijk vind om te overzien is of ze voor elk examen in wat voor instelling dan ook. Bijv. CITO t.o.v. Malmberg, daar zit toch een verschil tussen qua toetsen, van een andere orde, van een andere kwaliteit. Bij allemaal horen er processen bij die de veiligheid betreffen. Hoe goed ze dekkend zijn, dat kan ik, en dat geldt overigens bij de meeste standaarden, dat kan ik zelf niet zo goed overzien.	1: I tend to start looking for what I am still missing. But that is why I cannot think of what would be missing. 3: I honestly do not know if they are complete. I find them all very plausible, what I find difficult to see if they are for any exam in any institution whatsoever. E.g., CITO compared to Malmberg, there is a difference between the tests, of a different order, of a different quality. All processes are associated with safety. How well they are covering, I can, and that applies to most standards, I cannot oversee that.

Criteria missing	1: Dus in mijn optiek zijn ze wel compleet. 2: Waar ik aan zit te denken ik praktische voorbeelden, geen gegevens delen, geen informatie op een stick dat soort dingen komen in mij op. Dat zit hier wel impliciet in, ik zou het expliciet noemen. Ik denk dat mensen wel erg gevoelig zijn voor voorbeelden. Een veel gehoord proces bij ons is 'scheduling', misschien heet dat hier planning and acting. Maar het op de juiste momenten openzetten van de toetsing is wel heel belangrijk, dus misschien is dat iets wat je kan toevoegen? 3: Wat mij telkens opvalt, is dat het aantal indicatoren beperkt is. En dat is prettig. Ik ben ook wel eens procedures tegen gekomen waarbij je pagina's lang moest doornemen met allerlei indicatoren. Het aardige hiervan is dat je met een aantal heel kernachtig geformuleerde indicatoren je het hele aspect, in dit geval results afgedekt lijkt te zijn. Dat is de kracht van dit model, dat maakt het heel hanteerbaar.	1: So in my view they are complete. 2: What I think of is practical examples, no data sharing, no information on a stick that kind of things come to mind. That is implicit in this, I would make it explicit. I think people are very sensitive to examples. A much heard process with us is 'scheduling', maybe that's planning and acting here. But starting the assessment at the right time is very important, so maybe that's something you can add? 3: What repeatedly strikes me is that the number of indicators is limited. And that's nice. I have also come across procedures where you had to go through pages with all sorts of indicators. The nice thing about this is that with a number of very crisply formulated indicators you seem to be covering the whole aspect, in this case results. That is the power of this model, which makes it very manageable
Equivalency	1: Ik denk ook dat ze alle vier even belangrijk zijn. 2: Ik zou zeggen ja, misschien 8 en 1, die zou ik wat lager waarderen dan de anderen, de reden is, '…' 1&8 zijn meer voor de buitenwereld dat je het op orde hebt de anderen zijn meer intern. 5: Nee dat klopt. En, wat ik daar aangeef is, het geeft een beeld in ieder geval. Je moet bewust kunnen afwijken op onderdelen denk ik, en het geeft je bij het invullen een beeld van waar je op zou moeten focussen. Dus ik verwacht ook niet en keihard punten aantal. Het is geen harde wetenschap, dus je kan niet zeggen bij zoveel punten scoor je een voldoende. Als je ergens wilt verbeteren, dan kun je hiermee ook verbetering meten. Het weging geven aan alle onderdelen zou voor mij ook niet een doel zijn voor deze lijst (red. protocol). Het gaat namelijk over bewustwording, dus als je te diep in gaat op een weging schiet je volgens mij je doel voorbij.	1: I also think that all four are equally important. 2: I would say yes, perhaps 8 and 1, I would rate these a little lower than the others, the reason is, '...' 1 & 8 are more for the outside world to show you have it in order, the others are more internal. 5: No, that's right. And, what I mean is, it provides an image in any case. You have to be able to deviate consciously on parts, I think, and it gives you a picture of what you should focus on when completing. So I do not expect points. It is not a hard science, so you cannot say that you score a sufficient number of points. If you want to improve somewhere, you can also measure improvement with this. Giving weight to all parts would not be a goal for me for this list (Protocol). It is about becoming aware, so if you go too deep into a weighting, I think you'll miss your goal

Overview of the interview input on category 3: Grading of the protocol

Theme	**Dutch quote**	**Translated quote**
Grading system	1: Ja ben ik erg blij mee. Maar volgens mij zijn dat de 3 zaken waar het om gaat. Je wil kunnen constateren doet een organisatie voldoende of niet voldoende	1: Yes, I am very happy with it. But in my opinion, these are the three issues that matter. You want to be able to ascertain that an organization does

	en je wilt constateren of een organisatie misschien iets extra's doet dan wat er gevraagd wordt. En dat onderscheid dat mag je ook wel inzichtelijk maken. Dat stimuleert dan namelijk organisaties. (…) Je wilt ook dat er vanuit je protocol een stimulerende werking gaat. 2: Even kijken hoor. De onvoldoende zie ik en die is ook goed, dan is het vaak 'iets is er niet'. V&G, ik ga even bij security plan kijken. (…) Ik weet niet of jij dit hebt gemaakt Christiaan, maar ik vind het er goed uitzien als ik dit zo bekijk. Als ik er zo door heen scrol dan is er wel op een goede onderscheidende manier in beeld gebracht wat het onderscheid is tussen sufficient en good. 3: Ik moet vooral kunnen zeggen dat de zaak deugt. Dus ik kan mij voorstellen dat een keuze tussen 0 en 1 kan volstaan. 5: Dat lijkt mij zeker relevant. (…) Dit geeft je de mogelijkheid tot groei zeg maar. Uhm, mogelijk zou ik dan ook nog het knopje 'niet van toepassing' willen toevoegen. Dan is het per criteria mogelijk om niet van toepassing aan te geven in plaats van dat iets meteen onvoldoende is, dat terwijl het wellicht niet relevant is dat het er is. 6: Ja prima vind ik dat. Voor het doel om te checken of een onderdeel werkt zijn die drie categorieën prima.	enough or not enough, and you want determine if an organizations puts in extra effort that what is asked for. And you can also make that distinction clear. That is what stimulates organizations. (…) You also want to give a stimulating effect with the protocol. 2: Let's see. The insufficient, I can imagine this and I believe it is appropriate, then often you can say something is not there. Sufficient and good, I'm going to look at security plan (…) I don't know if you made this Christiaan, but I think it looks good while reading it. If I scroll through it, we can see a clear distinctions between what can be considered sufficient and good. 3: Above all, I must be able to say that the matter is being done correct. So I can imagine that a 0 or 1 score would suffice. 5: That seems relevant to me. This gives you the opportunity to show growth. I might also want to add the option 'not applicable'. Then it is possible to indicate per criteria if it is o.k. to leave it out without receiving an insufficient score, while it may not be relevant that it is there. 6: Yes, that is fine. For the purpose of checking whether a part works, those three categories are fine.
Concreteness	1: Uhm nou volgens mij hoef je dat niet zo heel concreet te maken. Kijk je moet concreet maken wat voldoende en onvoldoende is. En bij de vraag of iets goed is moet je eigenlijk maar 1 vraag hoeven stellen, namelijk doen ze extra bovenop wat wij als voldoende betrachten. Volgens mij moet je het zo simpel houden eigenlijk. Want hoe duidelijker je het gaat beschrijven hoe concreter je het maakt hoe moeilijker het wordt. 2: Laat ik het anders zeggen, ik werk ook met rubrics. En het is geworstel. Er is geen correct antwoord op deze vraag, dus dan val je terug op 'keep it simple'. Dit is misschien een soort van aanwijzing, maar daar heeft iedereen weer andere ideeën over. Dus dan zou ik er toch voor kiezen om het simpel te houden, want ik zie hier hele redelijke dingen staan. Dan moet je	1: Well, I do not think you'll have to make that very concrete. What is considered insufficient and sufficient should be made concrete. When it comes to the question whether something is good, you really need to ask only one question, namely did then do extra on top of what we consider sufficient. I think you should keep it that simple. Because the more clear you describe it, the more concrete you make it, the harder it will be to use. 2: Let me say it differently, I also work with rubrics. It can be tough. There is no correct answer to this question, so you fall back to 'keep it simple'. This can provide a clue, but everyone has an opinion on these matters. So therefore I would say keep it simple, because I see very reasonable thing here. So focus on

	het vooral nog betrekken op impact. Daar kun je meer over zeggen. Dat hoeft voor mij niet in percentages ofzo. Ik vind het wel een logisch geheel. 3: Laten we die verantwoordelijkheid maar leggen bij de security officer. Die zal binnen zijn team moeten bepalen wat er moet gebeuren. Ik denk dat hij daarin een beoordelende rol heeft. Om dat maar meteen te koppelen aan het protocol, hij heeft hiermee een prachtig handvat om het halfjaarlijkse of jaarlijkse gesprek vorm te geven. Want over elk van deze indicatoren kan je zelf bedenken waarover je het gesprek wilt voeren met je mensen. En kun je ze bevragen over de mate van zekerheid binnen de organisatie. Het moet hanteerbaar blijven. 5: Ja. Dat zal van de organisatie afhangen. Ik denk niet dat je heel SMART kunt formuleren wat er in je security plan moet komen te staan. Dus misschien moet je het wel gewoon algemeen formuleren. 6: als je hem groter gaat maken lever je weer in aan de bruikbaarheidskant, dus uhm. Dat kan ik mij prima voorstellen. En weetje, ik zou ook zeggen als je het gaat gebruiken, ga het eerst eens even proberen en dan is het mooi als het zou gaan werken zeg maar. En dan kun je daarna nog wel weer finetunen.	the impact. You can say more about that. I do not need a percentage or something. As a whole I believe it looks logical. 3: Let's put that responsibility with the security officer. He will have to decide what needs to be done within his team. I believe he has an evaluating role in this case. To link that to the protocol directly, he has a great tool to shape the conversation. Because about each of these standards you can think of what you want to talk to your people about. And you can question them about the level of certainty within the organization. It must remain manageable. 5: Yes, that will depend on the organization. I do not think that you can formulate all the criteria very SMART, considering what should be included in your security plan. So maybe you should just formulate it in general terms. 6: If you are going to make the criteria more concrete you will lose in terms of usability, so. I can easily imagine this. And you know what, I would say in the first place go out there and use the protocol, see if it works. Afterwards you would still be able to do some finetuning.
Rubrics	1: Nou die scores hebben vooral een meerwaarde. Ja, sowieso omdat je daarmee uitdrukt onvoldoende, voldoende of goed. Dus daar hebben ze sowieso een meerwaarde in, dus wat we net al zeiden. Beloningssysteem. Of het waarschuwingssysteem. Hoe die dan uitpakt. 2: Even kijken hoor. De onvoldoende zie ik en die is ook goed, dan is het vaak 'iets is er niet'. V&G, ik ga even bij security plan kijken. (…) Ik weet niet of jij dit hebt gemaakt Christiaan, maar ik vind het er goed uitzien als ik dit zo bekijk. Als ik er zo door heen scrol dan is er wel op een goede onderscheidende manier in beeld gebracht wat het onderscheid is tussen sufficient en good. Wel niet regelmatig, ja wel goed.	1: These scores have added value. Yes, in any case because you describe what can be considered insufficient , sufficient or good. So they have added value anyway. A reward system, or a warning system depending of the outcome. 2: 2: Let's see. The insufficient, I can imagine this and I believe it is appropriate, then often you can say something is not there. Sufficient and good, I'm going to look at security plan (…) I don't know if you made this Christiaan, but I think it looks good while reading it. If I scroll through it, we can see a clear distinctions between what can be considered sufficient and good. 3: If I score good, I will be happy. Do I need to be happy or should I at least be able to score sufficient. Above all, I must

	3: Van goed word ik blij, van sufficiënt daarmee ben ik tevreden. Wil ik elke keer blij worden, of moet de zaak vooral deugen. Ik moet vooral kunnen zeggen dat de zaak deugt. Dus ik kan mij voorstellen dat een keuze tussen 0 en 1 kan volstaan. 5: Nou ja als er een paar schalen niet van toepassing blijken te zijn dan wordt daar nu dus een score aan toegekend. Dan zou het wel mooi zijn als je er een interactief document van kan maken, dus dat wanneer je NVT invult je automatisch een andere totaal score krijgt	be able to say that the matter is being done correct. So I can imagine that a 0 or 1 score would suffice. 5: Well, if several criteria turn out to be 'not applicable', in the current form I will get an low score. So then it would be nice is you can make the scoring interactive, so if I would score a not applicable to a criterium you would automatically get a different total score.
Security labels	1: Ja. Maar (..) laat ik het zo zeggen het is zo dat de ene criterium heeft een hoger risico impact dan het andere. En ik denk wel dat het belangrijk is om dat in je weging mee te laten nemen. Want daarmee bewerkstelling je ook dat organisaties die daadwerkelijk hiermee aan de slag gaan ook als ze snel naar een voldoende willen gaan sneller die zwaarwegende issues aanpakken. 3: Daar kan ik wel in komen, ik denk ook dat de organisatie die feedback krijgt obv de 0/1 scores zelf de inschatting kan maken of er echt sprake is van een medium of high security risk. Als ik allemaal 1tjes scoor, heb ik een low security risk, want no security risk bestaat niet. Hierbij vind ik low risk goed gekozen. Het is in dit geval dan weer de officer die obv de uitkomst van de audit de uitkomsten zal moeten reflecteren op de omvang van het risico medium/high. 5: Kijk security, als iemand spiekt. Dan is het een risico. Maar de vraag is of al deze punten voor risico's staan. Ja, ik denk het eigenlijk wel. Het is maar een label, maar volgens mij dekt het de lading wel. Zolang je die Niet Van Toepassing er dan maar tussenuit haalt. 6: Ja ik ben heel erg gewend om met die termen te werken. Het sluit gewoon aan bij de gangbare termen van risicomanagement. Dus wat mij betreft is dat dan, zijn dat prima labels.	1: Yes. However, let me put it this way, some criteria contain a higher security risk impact than the other. And I think it is important to include that in you weighing. In that way, organizations who aim to score a easy sufficient overall score, will tend to invest in criteria that weight heavier. 3: I can relate to that, I also think that the organization that receives feedback based on the 0 or 1 scores can make the assessment whether there really is a medium or high security risk. When I score sufficient on all account it still can only be a low security risk, because no security risk in nonexistent. I think that low risk is well chosen. In this case it is the officer who, on the basis of the outcome of the audit, will have to judge whether to security risk is medium or high. 5: Looking at security, if someone cheats. Then it is a risk. But the question is whether all these criteria represent security risks. Yes, I think so. It is only a label, but I think it covers the content. As long as you leave 'not applicable' criteria out of the grading. 6: Ye, I am very much used to work with those terms. It simply fits the usual terms of risk management. So far as I'm concerned these are fine labels.
Security assessment	1: Ja. Maar (..) laat ik het zo zeggen het is zo dat de ene criterium heeft een hoger risico impact dan het andere. En ik denk wel dat het belangrijk is om dat in je weging mee te laten nemen. Want daarmee bewerkstelling je ook dat	1: Yes. However, let me put it this way, some criteria contain a higher security risk impact than the other. And I think it is important to include that in you weighing. In that way, organizations who aim to score a easy sufficient overall

	organisaties die daadwerkelijk hiermee aan de slag gaan ook als ze snel naar een voldoende willen gaan sneller die zwaarwegende issues aanpakken. 2: (…) als ik op 1 willekeurige criteria een nul scoor heb ik een high security risk. (…)Ja ik denk dat ik daar wel in mee zou kunnen gaan, dat je dan zegt ik heb een high risk. Misschien, ik zie hier die data forensics, dat je dan ook een high security risk hebt, daar wordt nog niet zoveel gebruik van gemaakt. Dat is wat groot gezegd, ik zou daar een beetje aarzelen om te zeggen dat je een hoog risico loopt. 3: Inderdaad, dus het onderscheid tussen 0 & 1 is wat mij betreft véél groter dan tussen de 1 en de 2. Dat onderscheid geeft mij te weinig meerwaarde. Ik hoef als (company) niet te excelleren, ik moet deugen. 5: Dat is natuurlijk. Bij 'insufficient' ziet het er nu uit dat je er niet over na hebt gedacht, en voor mij is het dan prima dat je een hoog risico, of nou ja 'wellicht' een hoog risico loopt. Het is onduidelijk of je een risico loopt, maar dat is hetzelfde als een risico lopen. 6: (…) Misschien heeft het meer te maken met welke standaard je dan betrekt. Als je kijkt naar 'exam development', als het daar niet goed is. Daar heb je een groot risico dat wanneer je het niet goed doet dat dan je onderdelen op straat komen te liggen. (…) Ja dus kans maal effect he, dus hoe groot is de kans dat.. ja dus welke kans heb je.	score, will tend to invest in criteria that weight heavier. 2: (…) if I'd score a 0 on a single criteria, I would have a high security risk. (…) Yes, I think I could agree with that, that you would say I'd have a high risk. Perhaps, I see data forensics over here, because not everybody does it on a usual basis, I'd say the conclusion in this case would be heavy. I would be hesitant to say that you'd have a high security risk. 3: Indeed, so the difference between 0 & 1 weight far heavier in my opinion than between 1&2. The later distinction gives me to little added value. I do not have to excel as (a company), I have to be good enough. 5: That is natural, at insufficient it now looks like you'd have not thought about it. If this is the case, for me it is fine to say you'd possibly have a high risk. It remains unclear whether you are at risk, but that is the same as being at risk. 6: (..) Perhaps it has more to do with a specific standard. If you look at 'exam development', if this is not done securely. Here the security risk could be high if you have not done it correct. The content could become know. (…) Yes, so chance versus effect, so how big is the chance… so what chance do you have.

Overview of the interview input on category 4: Data Forensic standards

Theme	**Dutch quote**	**Translated quote**
Goal feasibility	4: Ja. In de zin van dat het voorlichting is inderdaad. En met uitgewerkte voorbeelden van casussen kan dat het dan zijn. Iets wat wij ook merken is, de volgorde van fraude detectie, begint bij een signaal, uit een afname door een betrokken partij, en dat je daarna mogelijkheden hebt om naar de data te kijken. Ik denk dat dat een vrij essentiële is, en als je andersom gaat redeneren het best wel een moeilijk kwestie is. Of kunnen we ook gewoon zonder waarschuwing naar de data kijken. Dat laatste vraagt misschien om een ander protocol.	4: Yes. Is the sense that it is aimed to provide information. And with detailed examples of cases it can be just that. Something we also notice, the sequence of fraud detection starts with a signal after examination. This offers you the opportunity to look at the data. I think this is pretty essential, if you would go at it the other way around, it probably a difficult issue. Can we simple look at the data without signals. The latter may require a different protocol. 7: Currently it is quite, well, compelling.

	7: Het is nu best wel een soort van, nou ja, dwingend.	
Current standards	4: ze lijken erg op elkaar, dus de vraag is een beetje of de driedeling is vanwege het gedrag, dus de soort fraude. Of vanwege het geen je met de data kan laten zien. En nu heb ik het idee dat ze nog een beetje gecombineerd waren. Zo heb je de voorbereiding he, pre-knowledge. Maar die tweede slaat veel meer op wat je in de data ziet (*criteria 2 van std.1*). Dus ik zit met de vraag van wat is nou fraude, want het type fraude is bij de eerste voorkennis, en staat bij 3 hoe kom ik aan die voorkennis, maar criteria 2 is niet een type fraude, maar een gevolg van 'oké we doen het (met zijn allen) samen. Dus die rubricering is niet identiek. 7: Nou de derde begreep ik niet helemaal. De eerste dus voorkennis, dus alles gaat om ongeoorloofde voorkennis. Dat snap ik dus dat lijkt mij ook zinvol. Zeker bij adaptief testen met grotere itembanken enzo. Hoewel dit dan niet perse fraude is.	4: they are very similar, so the question is whether the distinction is because of the behavior, so the kind of fraud. Or because of what the data can tell you. Currently I have the idea that it is a combination of the two. This one (std1, cri 1) is how you prepare, pre-knowledge. But the second one (criteria 2) refers more to what you can see in the data. So I am concerned about what can be called fraud. Because at standard one it is pre-knowledge, standard three describes how you can get pre-knowledge, but criteria 2 does not refer to fraud but is a consequence of collaborating. So the way it is classified is currently not identical. 7: well the third I did not quite understand. The first one is about pre-knowledge, so everything is about unauthorized knowledge. I understand that, so that makes sense. Especially with adaptive tests with larger item banks. Although this is not necessarily fraud.
Completeness	4: (…) Hmm, ik snap wat je zegt, maar ik weet niet of ik dat nou zo logisch terug vind hierin. Ik zit even te kijken naar, want je wilt een standaard opzetten voor hoe detecteer ik, dan kan ik mij voorstellen dat je zegt hoe detecteer ik als er van te voren fraude is gepleegd, maar die laatste is dan wezenlijk anders, want dan zeg je iets over de effecten ervan. Dat vind ik best wel een moeilijke, want de eerste is een soort fraude, de tweede is in feite een soort fraude, maar de derde is een gevolg van de soort fraude. 7: (…) respons times is eigenlijk een belangrijke factor en dan meer in het algemeen he, dus dit zijn dingen die je kunt doen maar of dat nou echt standaarden zijn zou je, je af kunnen vragen. Waar ik me dan bijvoorbeeld erg zorgen over zou maken is, al die statistics hebben niet altijd heel veel power. En de base rate is vaak laag. Dus hoe weet je dat je niet teveel false positives hebt	4: (…) I understand what you're saying, but I don't know if I find it logical. I'm just looking at, because you want to set a standard for how to detect, than I can imagine that you would want to detect pre-knowledge, but the third standard is essentially different, because then you say something about the effects. I think that is a difficult one, because the first is a kind of fraud, the second one is in fact a kind of fraud, but the third is a consequence of the type of fraud. 7: (…) response time is actually a important factor, in general, so these are thing you can do but whether that really should be standards, that is the question. For example, what I would worry about is that the statistics do not have a lot of power. Also the base rate is often low. So how do you know that you won't end up with too many false positives.
Types of fraud	4: Even kijken hoor, wij meten de scores op antwoorden, de tijd op de antwoorden, de standaard test theory of IRT die we draaien, dat draai je zonder dat je iets af weet van fraude signalen. Als we wel signalen krijgen dan ga je groepen vergelijken, of naar het gedrag van 1 leerling. Wat ik zie is dat die er allemaal wel in staan.	4: Let's see, we measure the scores based on the answers, the response time, the standard test theory or IRT that we run, those can be run without signals of fraud. If you do receive these signals you will compare groups or look at the individual. What I see is that these are all in here.

	5: Wat ik zelf zag, het corrigeren door een beoordelaar. Wat ik vaker hoor, identiteitsfraude. Bij afname van de docent speelt het niet zo, maar het is toch wel een type van fraude. Dus dat. Dat zag ik er niet snel tussen komen. 7: Even kijken hoor. Nou kijk dat colluding with others heeft een vrij omvangrijke vorm. Dat kan natuurlijk van alles betekenen, maar wat hier niet echt in staat is gewoon spieken.	5: What I saw myself, correcting by an assessor. What I hear more often, identity fraud. This is often not the case when the lecturer is responsible, but it still is a type of fraud. I did not see these things. 7: Let's see. Well, colluding with others is fairly broad. That can mean anything, but what it not say now I looking directly at the work of someone else.
Missing indices	4: Uhm, nouja. De standaard parameters uit de IRT, de P & ritwaarde voor kwaliteitscontrole voor je examen is de basis voor detectie. Ja dat hoort in je 'plan' je staan, maar je moet het ook doen, want als het alleen in je plan staat maar je doet het niet, ben je niet volledig. Dus ook de uitvoer hoort in je cyclus. 5: Voor de hand liggende analyses, de items, de responsietijd die worden genoemd. Uhm, volgens mij ook een gelijk itempatroon.(…) Voor data forensics kwam ik niet zoveel dingen tegen. Die exposure, naarmate die groter wordt gaat de 'p'waarde omhoog , maar wat je ook wilt is dat wanneer items heel vaak worden aangeboden, is dat je ze dan on-hold kan zetten. Want dan is je exposure gewoon te groot geworden. En is je items te bekend geraakt. En dan zie je veel kandidaten zo'n item goed maken. Dat gaat dus over langere tijd hè. Exposure control. Dus dan gaat een item er uit als deze te vaak gebruikt wordt. 7: Nee ik denk dat het gewoon een algemeen probleem is met indices. Kijk je hebt per kandidaat niet heel veel informatie. Maar bij een beperkte aantal items moet je vaak wel heel veel of heel bijzonder spieken. Dus ik denk dat het vooral een algemeen probleem is bij deze indices. (…) Ja inderdaad, ja ja. En nee ik zou het niet iedereen screenen met dit soort indices.	4: Well, the standard parameters from the IRT, the P-value for quality control for your exam is the basis for detection. These should be in your security plan, but you also really have to act on these things. So the way you act should also be part of the cycle. 5: Obvious analyses, the items, the response time, those are mentioned. According to me, an identical pattern (…) In terms of data forensics I can't really name things. About exposure, when it becomes higher the P-value goes up as well, but you want to be able to put these items on hold if they are offered a lot during examination. Otherwise these items become to familiar. What you then see is many candidates passing these items. So that is an analysis over time. Exposure control. So then an item is taken out of the item bank if it is used to often. 7: No, I think it's just a general problem with indices. Look, you do not have a lot of information per candidate. But with a limited number of items (per exam) you would have to cheat on a large scale. So I it is mainly a general problem with these indices. (…) Yes indeed, yes. And no, I would not screen everyone with these kind of indices *(referring to my question about only using these indices after getting fraud signals).*

Appendix G – The EDF Protocol

EDF Protocol

For Quality Assurance around Fraud Prevention and Detection

Dutch: De definitieve versie van het EDF-Protocol is gratis te verkrijgen via www.xquiry.nl
English: The final version of the EDF-Protocol kan be freely requested via www.xquiry.com