

# Chapter 20

## The Theory and Practice of Educational Data Forensics



Sebastiaan de Klerk, Sanette van Noord and Christiaan J. van Ommering

**Abstract** Although the testing community has been trying to prevent test fraud through multiple practices and methods (e.g., applying strict security practices or testing protocols), test fraud is still a ubiquitous problem. Exact numbers are unknown, but self-report studies show that up to 85% of students admit to committing test fraud at least once during their school career (Lin and Wen, *Higher Educ* 54:85–97, 2007; Hughes and McCabe, *Can J Higher Educ* 36:1–12, 2006; Berkhout et al., *Studie and Werk* 2011. SEO Economisch Onderzoek, Amsterdam, 2011). Research on the statistical detection of test fraud, also called educational data forensics (EDF), already exists since the 1920s (Bird, *School Soc* 25:261–262, 1927), but the body of research started growing considerably since the 1970s (e.g., Angoff, *J Am Stat Assoc* 69:44–49, 1974). Nowadays, many methods and models are presented in the literature. Two of those models are the Guttman error model (Guttman, *Am Soci Rev* 9(2):139–150, 1944; Meijer, *Appl Psychol Measur* 18(4):311–314, 1994) and the log-normal response time model (Van der Linden, *Psychometrika* 80(3):689–706, 2006). In the first part of this chapter, both models will be discussed. In the second part of this chapter, an empirical study on the functioning of the Guttman- and response time model will be presented. In the final part of the chapter, the design, development and validation of a protocol on the use of EDF will be presented.

### 20.1 Educational Data Forensics

Educational data forensics (EDF) is the statistical analysis of test takers' response data with the aim of the detecting aberrances that potentially indicate test fraud (Wollack and Fremer 2013). In this chapter EDF is discussed from several angles. We will start with a short historical context and introduction to the most used EDF methods and models, followed by a discussion on how EDF can be used in practice. In part two and three of this chapter, we will present two empirical studies on EDF.

---

S. de Klerk (✉) · S. van Noord · C. J. van Ommering  
eX:plain, Amersfoort, The Netherlands  
e-mail: [s.dklerk@explain.nl](mailto:s.dklerk@explain.nl)

© The Author(s) 2019  
B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, [https://doi.org/10.1007/978-3-030-18480-3\\_20](https://doi.org/10.1007/978-3-030-18480-3_20)

### ***20.1.1 Historical Context, Theory, and Practice of EDF***

Test fraud can be defined as deliberately influencing the test process with the aim of improving the results on one or more tests. This puts different methods of committing test fraud within a broad spectrum. Less harmful methods like cheat sheets or test peeking on the one hand of the spectrum and more harmful methods like item theft or identity fraud on the other hand of the spectrum. The administration of a test yields test taker response data that can be used to detect potential test fraud. The statistical methodology used to this end is part of the EDF domain. Although most of the current methods are exclusively applicable to computer-based test (CBT), the first methods discussed in the literature were used to detect test fraud on paper-based tests (PBT). For example, Bird (1927) already studied the number of identical mistakes in the exam papers of pairs of test takers (i.e., test takers who were seated next to each other in the exam room). He reasoned that if the number of identical mistakes exceeded ‘the chance threshold’, then it was very likely that test fraud had been committed. However, the threshold is difficult to determine, especially for pairs of essays. In the following decades, the multiple-choice test made its entrance in the field of testing. Response data from mc-tests enabled researchers to use more advanced EDF methods on detecting collaboration between two or more test takers or so-called ‘copying behavior’. Saupe (1960) was one of the first researchers that published an empirical model that could be used to determine the likelihood that two test takers collaborated on their mc-test. The statistical procedures on copying behavior in mc-tests were further elaborated and discussed by Bellezza and Bellezza (1989). More recently, Van der Linden and Lewis (2015) introduced Bayesian statistics, opposing the traditional hypothesis testing approach, for detecting copying behavior on tests. Since 1927 many statistical indices on copying behavior or similarity analysis have been presented in the literature. All these indices revolve around the rationale that two or more test takers have a strong similarity in their answers, both correct and incorrect.

Obviously, the higher the correspondence between answers, the more likely it is that test fraud has been committed. Yet, research is still not consonant about when there is enough correspondence to irrefutably prove that test fraud has been committed. It is questionable if this will ever be possible at all. There are many potential causes to aberrations in response patterns (Reise and Due 1991), and there is always a chance that a test taker produces a very unlikely yet honestly produced response pattern. Furthermore, similarity analysis models may become obsolete very quickly. The introduction of CBT has also increased the use of randomly composed tests (e.g., on the fly testing or computerized adaptive testing), which do not allow test takers to copy answers because each test taker is presented a unique sample of questions from the item bank. Of course, randomized testing contributes to the prevention of test fraud, but it does not make it impossible. Other methods of test fraud (e.g., the use of technological devices, preknowledge, or collusion) can still occur. The fact that similarity analysis lacks the (statistical) power to detect all forms of test fraud and because new testing methods (e.g., CAT) do not allow similarity

analyses to be performed has led researchers to investigate other ways of detecting test fraud.

For example, statistical methods to assess the congruence between an item response pattern and a specified item response theory (IRT) model started to proliferate since the 1990s (Reise and Due 1991). These models are also called ‘person fit’ models. The general idea is that response patterns can be defined a priori using these models. Aberrations to these patterns that are determined a posteriori may point to potential test fraud. A well-known person fit model is the Guttman error model. Guttman (1944) invented an item response model in which there is always a chance of 1 that you can demonstrate an ability when you possess that ability, and a chance 0 that you cannot demonstrate an ability when you do not possess that ability. An aberration to such a pattern is called a ‘Guttman error’ (Meijer 1994). Simply put, in a testing context, a Guttman error indicates that you have answered an item correctly that is beyond your ability level. Of course, this happens rather often—coincidentally a test taker knows the answer to a difficult question or guesses correctly a couple of times. Nonetheless, a strong aberration in the number of Guttman errors that a test taker produced during the test, potentially indicates that test fraud has been committed. The standard Guttman error is the point of departure for many different EDF indices (Van Noord 2018).

Still other EDF methods rely on the analysis of the number of erasures or revisions that a test taker made during the test (McClintock 2015; Van der Linden and Jeon 2012). Interestingly, these methods can be used on both PBT and CBT. In PBT, highly sensitive scanners can discriminate reliably between erasure marks and marks made to indicate the test takers’ answers (McClintock 2015). In CBT, when programmed correctly, the test administration software can log whether a test taker (or someone else) changed their answers to the test. To an increasing extent, test organizations are routinely conducting erasure analyses on their tests. The number of wrong-to-right (WTR) changes seem to be most promising in detecting potential test fraud (Maynes 2013). Qualls (2001) was one of the first researchers that empirically investigated how often erasures occur during a test, and what can be regarded as an aberrant number of erasures. Traditionally, and this also holds for the number of Guttman errors or the number of identical mistakes (similarity analysis), researchers check for test takers that are beyond three standard deviations from the mean in their number of erasures and label these test takers as aberrant. As already mentioned before, more recent and advanced Bayesian models are opposing this view, also in erasure analysis (Van der Linden and Jeon 2012; Van der Linden and Lewis 2015). Several comprehensive test fraud cases in the United States have been revealed through erasure analysis (see for example <https://nyti.ms/2k6osyd>). Remarkably, it was discovered that teachers had changed the test takers’ answers to make themselves and their schools look better.

Response time analysis is also used to detect potential test fraud (Van der Linden 2006). However, these analyses can only be performed in CBT, as the system must log a test takers’ response time. There are several ways to investigate the authenticity of response times, but the central idea is that fraudulent test takers show aberrances in their response time. Fast response times may indicate preknowledge of the test items and slow response times could point to the use or misuse of (unauthorized) resources,

such as a smartphone or a modified calculator that can help in answering test items (Van der Linden 2006; Van der Linden and Jeon 2012; Van der Linden and Lewis 2015). Basically, except for extreme aberrations, response times are difficult to use in detecting test fraud. This is because there is high variance in the response time of test takers, both within and between test takers. Between test takers, because some test takers are just naturally slower or faster than others, and within test takers because of intrapersonal factors such as concentration, indecisiveness, etc. Therefore, simply using an overall test taker mean and then labeling test takers three SD under or above the mean for several items will not work. Van der Linden (2006) presents a log normal response time model for the analysis of response times. In his log normal model both the test taker's working speed and the characteristics of the test time are accounted for, when determining whether a test taker displays aberrancies. For example, a fast response time is not necessarily conspicuous when it concerns a test taker who is naturally very fast and is answering to a concise and closed-ended 'factual' question.

Above, the historical context of EDF and the most used methods have been discussed. The counter side of these methods is that they are in many cases difficult to apply, especially by practitioners. Often, big data log files need to be analyzed, preferably through automatized algorithms. Easy to use and hands-on software does not exist for EDF. Researchers can resort to *R* (2018) packages, but most test agencies do not employ researchers or people that know how to work with *R*. We have, therefore, been working on developing a hands-on software application that can automatically analyze large data sets for potential test fraud. We have called the software *EDF Monitor*. *EDF Monitor* can, through a transformation database and a communication protocol, be connected to every test administration program. Test takers' response data are uploaded into the software daily through an automatic job. The analysis is run nightly, and aberrant test takers are 'flagged' automatically, so that the practitioner can use the information the following day. *EDF Monitor* currently encompasses nine algorithms, and we are planning to expand. The algorithms are based on similarity analyses, the Guttman error person fit model, and the log normal response time model. In the future, erasure analysis will also be part of *EDF Monitor*. Of course, we were interested in the empirical functioning of *EDF Monitor* and the analyses. We have therefore performed a study on the detection rate (i.e., the number of true positives) and the reliability (i.e., the number of false positives) of the software.

## 20.2 An Empirical Study on the Reliability of *EDF Monitor*

The main objectives of the study were to establish the highest detection rate using different and manually set combinations of the algorithms in the software (i.e., the true positive rate), and the reliability of the algorithms (i.e., the true negative rate). The focus of the study was on the functioning of the Guttman error based indices, the log normal response time model, and a combination of both. In total a combination of six indices were researched. The first three indices were the standard Guttman

error model ( $G^*$ ), and the Guttman error relative to the distance of the vector of item responses. In the standard Guttman error model, a Guttman error exists in a vector of item responses that are ordered from easy to difficult, when a more difficult item is answered correctly after an easier item has been answered incorrectly. In the adapted Guttman error models, that take relative distances into account, a Guttman error only counts as an error when a specific number of positions in the vector have been skipped or when the distance in  $p$ -value (derived from Classical Test Theory) has crossed a pre-specified  $p$ -value threshold. The fourth index is the standard Guttman error model combined with the log normal response time model. The fifth and sixth indices are the adapted Guttman error indices combined with the log normal response time model.

### 20.2.1 Method

The study was conducted using mixed methods, with a research design that covers a confirmatory and exploratory phase. In the first phase the control group and the experimental group were compared, and in the second phase the different experimental conditions were compared. In the experimental conditions, the participants were instructed to commit test fraud in several ways: (1) using a smart device (smartphone), (2) internal collaboration (proctor leaves exam room), (3) proctor assistance, (4) using cheat sheets (notes on paper), and (5) pre-knowledge. Participants in the control condition were regular test takers which were put under enhanced surveillance (i.e., increased number of proctors).

#### *Participants*

The control group consisted of 37 participants (age:  $M = 39.30$  years,  $SD = 13.70$ ; 14 females). The experimental group consisted of 80 participants, distributed over the different conditions: smartphone ( $n = 18$ ; age:  $M = 17.83$  years,  $SD = 2.85$ ; 7 males), internal collaboration ( $n = 16$ ; age:  $M = 17.38$  years,  $SD = 1.02$ ; 8 males), proctor assistance ( $n = 21$ ; age:  $M = 16.95$  years,  $SD = 1.69$ ; 7 females), cheat sheet ( $n = 8$ ; age:  $M = 17.38$  years,  $SD = 1.19$ ; all males), and pre-knowledge ( $n = 17$ ; age:  $M = 51.25$  years,  $SD = 3.58$ ; 5 times male). Due to value of the test items, participants in the fifth condition (i.e., pre-knowledge) could not be actual test takers. For test security reasons it would be highly undesirable to have regular test takers gain pre-knowledge of the test items of an exam. Therefore, four co-workers with deep knowledge of the content of the item bank were selected to take five trial exams each, on the condition that they were familiar with the item bank.

#### *Materials*

Randomly drawn items from the Basic Competence Legal Knowledge for Extraordinary Detective Officers item bank (linear on the fly) were used (230 items). The unique tests consisted of 50 items and had an average  $p$ -value of 0.69, an average

RIT-value of 0.28 and an average reliability of  $\alpha = 0.76$ . All tests were trial exams and test takers would not be certified when they passed the exam.

### *Procedure*

The experiment was run in several individual and group sessions of approximately 90 min. Groups consisted of participants in the same experimental condition or the control condition. Participants in different conditions were never combined in one exam room. Only the participants in the fifth condition (e.g., pre-knowledge) were run in individual sessions for practical reasons, as the group setting had no additional value. The participants in this condition were also not proctored, all others were by two or more proctors, including a researcher. All participants were informed about the research goals, method, and consequences of participation, after which the participants signed the informed consent form. Participants in the control group were strictly monitored, to ensure nobody in this group cheated. Respondents in the experimental groups were instructed to cheat in a specific way to attain the best grade possible on the trial exam.

In the first condition (i.e., smartphone) the participants were told to consult their smartphone during a period of three minutes. This moment was announced after the first 15 min of the exam had passed. The violators were not allowed to talk out loud or make other noises during this time. In the second condition (i.e., internal collaboration) the group of violators was left alone by their proctors for three minutes, allowing them to consult each other for answers. The violators were notified of the occurrence and instructed to use their time wisely. They were not told how long the proctor would be gone, as this would also be unknown to them, were this an event to occur in reality. In the third condition (i.e., proctor assistance) the violators were instructed to consult the proctor at least once when they struggled with an item. The proctor, a teacher from the same educational institution as the test takers, was instructed to provide more information than usually allowed, or, if this information did not help, the actual answer to the question. Since the participants were to wait their turn to ask a question, they were encouraged to mark the respecting item and continue with the exam while the proctor took turns. In the fourth condition (i.e., cheat sheets) the violators were asked to bring a cheat sheet to the exam. The proctor was instructed to ignore the sheets. In the fifth condition (i.e., pre-knowledge) the violators completed the exam without interference of a proctor or researcher.

### *Data analysis*

First, it is tested whether the mean score for the six indices in the cheating conditions is significantly higher than in the non-cheating condition. This is evaluated using independent *t*-tests. Secondly, the optimal true positive detection ratio and true negative ratio are established. In addition, and for an explorative purpose, for each condition the detection ratio was determined through a Kruskal-Wallis ANOVA, combined with Mann-Whitney U and Wilcoxon Signed Ranking tests.

**Table 20.1** Descriptive and results from the initial data forensics analyses

Index	Not cheating			Cheating					
	(N = 37)			(N = 80)					
	M	SD		M	SD		t(115)		p
$G^*$	0.08	1.16		0.88	1.08		-3.65		<001
			Mean rank			Mean rank	U	z	
$G^*_{d20}$	0.52	1.41	49.07	0.99	1.51	63.59	1112.50	-2.16	0.031
$G^*_{d0.20}$	-0.40	0.91	48.96	0.02	1.10	63.64	1108.50	-2.19	0.029
$G^*_{rt}$	0.01	1.04	39.34	1.60	2.15	68.09	752.50	-4.28	<0.001
$G^*_{rtd20}$	0.41	1.10	46.64	1.87	2.75	64.72	1022.50	-2.71	0.007
$G^*_{rtd0.20}$	-0.34	0.75	46.22	0.75	2.29	64.91	1007.00	-2.84	0.005

Note In this table *p* is reported two-tailed, and *z* is corrected for ties

### 20.2.2 Results

First, regarding the comparison between the non-cheating and cheating conditions, results showed that the scores on all six indices were higher in all five cheating conditions as compared to the control condition (see Table 20.1).

Secondly, regarding the optimal true positive and true negative ratio, we found that a combination of combined indices leads to highest ratios. That is, when the  $G^*_{rt}$  index (the standard Guttman error model combined with the log normal response time model), was combined with the  $G^*_{rtd12}$  (the adapted Guttman error model with 12 skipped positions in the item response vector that was ordered from the easiest item to the most difficult item), the  $G^*_{rtd0.21}$  (the adapted Guttman error model for a *p*-value threshold of 0.21 combined with the log normal response time model), and the  $G^*_{rtd20}$ , it leads to the optimal true positive and true negative ratio for this data set. The true positive ratio was in this case 37.5%. That is, of the 80 students in the cheating conditions, 30 were indeed detected as cheaters. 50 of 80 students who were in one of the cheating conditions were not detected, which is translated in a false negative ratio of 62.5%. The true negative ratio was 97.3%—or a false positive ratio of 2.7%. That is, of the 37 students in the non-cheating condition only one was wrongly detected as a cheater (a type I error).

Using the combination of combined indices discussed above, our explorative analyses showed that especially cheating through proctor assistance and pre-knowledge could best be detected (see Table 20.2). Yet, four Kruskal-Wallis ANOVAs indicated that there were no significant differences between the cheating conditions for analysis with the  $G^*_{rt}$  index,  $H$  (corrected for ties) = 8.030,  $df = 4$ ,  $N = 80$ ,  $p = 0.090$ , Cohen’s  $f = 0.336$ , the  $G^*_{rtd12}$  index,  $H$  (corrected for ties) = 8.453,  $df = 4$ ,  $N = 80$ ,  $p = 0.076$ , Cohen’s  $f = 0.346$ , the  $G^*_{rtd0.21}$  index,  $H$  (corrected for ties) = 3.571,  $df = 4$ ,  $N = 80$ ,  $p = 0.467$ , Cohen’s  $f = 0.218$ , and the  $G^*_{rtd20}$  index,  $H$  (corrected for ties) = 8.006,  $df = 4$ ,  $N = 80$ ,  $p = 0.091$ , Cohen’s  $f = 0.336$ .

**Table 20.2** Mann Whitney U statistics determining the validity of the indices for every experimental group

Experimental index	Group	Control condition	$U$	$z$	$p$	$r$
	Mean rank	Mean rank				
<i>Smart phone</i>						
$G_{rt}^*$	36.19	24.01	185.50	-2.66	0.008	0.36
$G_{rid12}^*$	36.25	23.99	184.50	-2.69	0.007	0.36
$G_{rid0.21}^*$	30.75	26.66	283.50	-0.93	0.353	0.13
$G_{rid20}^*$	33.25	25.45	238.50	-1.72	0.085	0.23
<i>Internal collaboration</i>						
$G_{rt}^*$	31.31	25.14	227.00	-1.35	0.178	0.19
$G_{rid12}^*$	30.75	25.38	236.00	-1.18	0.239	0.16
$G_{rid0.21}^*$	27.59	26.74	286.50	-0.19	0.849	0.03
$G_{rid20}^*$	26.56	27.19	289.00	-0.14	0.890	0.02
<i>Proctor assistance</i>						
$G_{rt}^*$	41.98	22.42	126.50	-4.25	<0.001	0.56
$G_{rid12}^*$	42.60	22.07	113.50	-4.49	<0.001	0.59
$G_{rid0.21}^*$	35.88	25.88	254.50	2.24	0.025	0.29
$G_{rid20}^*$	39.26	23.96	183.50	-3.36	0.001	0.44
<i>Cheatsheet</i>						
$G_{rt}^*$	31.75	21.11	78.00	-2.09	0.037	0.31
$G_{rid12}^*$	32.63	20.92	71.00	-2.31	0.021	0.34
$G_{rid0.21}^*$	28.44	21.82	104.50	-1.34	0.201	0.20
$G_{rid20}^*$	27.25	22.08	114.00	-1.03	0.327	0.15
<i>Pre-knowledge</i>						
$G_{rt}^*$	38.03	22.66	135.50	-3.35	0.001	0.46
$G_{rid12}^*$	33.68	24.66	209.50	-1.97	0.049	0.27
$G_{rid0.21}^*$	34.68	24.20	192.50	-2.36	0.018	0.32
$G_{rid20}^*$	35.21	23.96	183.50	-2.47	0.013	0.34

Note In this table  $z$  is corrected for ties, except for the Cheatsheet experimental group, and  $p$  is two-tailed

### 20.2.3 Discussion and Conclusion

The aim of this study was to assess the detection strength of several data forensics indices. The response data of a control group of highly supervised examinees and an experimental group of instructed cheaters were analyzed for that purpose. We found that a combination of the Guttman error model, two adaptations to the Guttman error model, and the log normal response time model yielded the highest detection



rate (37.5%) combined with the lowest type I error rate (2.7%). It could therefore be said that the strength of these indices lies in their combination. Or, response time analysis is a better behavioral indicator when combined with the Guttman error model than when the Guttman error model is used separately. Although there were no significant differences in the indices between the five methods of cheating, a trend seems to lean towards proctor assistance and pre-knowledge to be best detectable with this combination of indices. The goal of a larger and improved follow-up study is to investigate whether this is indeed the case.

## 20.3 The Design and Development of the *EDF Protocol*

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

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 therefore also answered 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?

### 20.3.1 *Method*

The *EDF Protocol* 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 *EDF Protocol* 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 et al. (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.

For the main topic of the study several search terms were used (see Table 20.3). Boolean search operators were also used during this step (e.g., AND, OR, NOT,

**Table 20.3** 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

and \*). The initial search findings were reduced by 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.

### *Step 2—Developing an EDF-Protocol prototype*

The insights gathered in the literature search were used in the development of the first set of standards of the prototype (Part A), as well as a corresponding grading system. The intention was to make the standards (concerning prevention of misconduct during the process of examination) as complete as possible before starting the interviews.

The development of part B (i.e. the standards and criteria for detection of misconduct by means of using data forensics) took 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 comparative studies, and thus often lacking proper directions for practical implementation. 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 and a set of standards for detection (by means of data forensics) because these goals do not always coincide in practice.

### *Step 3—Validating the EDF-Protocol standards Participants*

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. 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*

Systematic expert interviews offer the possibility to identify strengths and weaknesses in the content (McKenney and Reeves 2012; Piercy 2015). This method is a valuable source of data collection, particularly when establishing the correctness (e.g., validating) of the 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 of an interview 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?”

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.

After the interviews, all the recordings have been converted to verbatim transcripts to keep statements in their proper context. 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.

#### *Step 4—Adjustment of the EDF prototype and final protocol*

In the fourth step, the statements from the experts were used to transform the prototype into a final version of the *EDF Protocol*.

#### *Step 5—Implementation of the EDF Protocol*

In the fifth step of this design research the final *EDF Protocol* was used to determine if there was a possible security risk within an exam program.

## **20.3.2 Results**

### *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 search was carried out in 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. Figures 20.1 and 20.2 show the steps of the search processes for both queries.

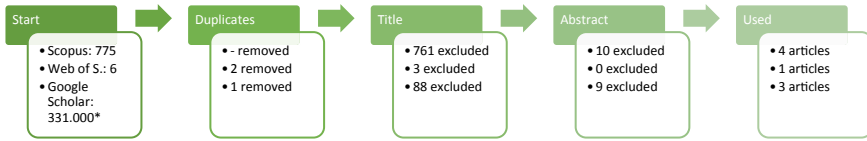


Fig. 20.1 PRISMA flow chart of the search process in the query of test security



Fig. 20.2 PRISMA flow chart of the search process in the query of data forensics

<b>Part A – Standards for Fraud Prevention</b> .....	<b>1</b>
1. Security plan.....	1
2. Security Team: tasks and responsibilities.....	2
3. Exam development process and maintenance.....	3
4. Security of Examination.....	4
5. Security of Results.....	5
6. Internet Screening.....	6
7. Security incident response.....	7
8. Performing Security Audit.....	8
<b>Part B – Standards for Fraud Detection through Data Forensics</b> .....	<b>9</b>
1. Detecting Preparatory Fraud Threats: Pre-knowledge and Item Compromise.....	9
2. Detecting Test Score Similarity and Answer copying.....	10
3. Detecting Unusual Gain Scores and Test Tampering.....	11

Fig. 20.3 Body of content of the EDF protocol prototype

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. Figure 20.3 shows the steps of the search process.

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. Because 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 (n = 20).

### *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 Fig. 20.3. The prototype standards each relate to the most commonly used and accepted guidelines on test security: The Security Guidelines by NVE (Dutch Association for Assessment) and Caveon (2016), the guidelines by SURF (ICT cooperation for education and research in the Netherlands, 2014), and the Standards for Educational and Psychological Testing (AERA et al. 2014) as well as other literature to support the inclusion of these criteria in the prototype.

### *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 ( $n = 6$ ). 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 EDF prototype and final protocol*

The interview statements were summarized 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 Fig. 20.4.

#### *General protocol adjustments*

The first adjustment removes the 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, the result is 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 streamlined in general and the content more consistent.

<b>Contents</b>	
<b>EDF-Protocol Standards</b> .....	<b>1</b>
1. Security plan .....	1
2. Involved personnel: tasks and responsibilities .....	2
3. Exam development process and maintenance .....	3
4. Security of Examination.....	4
5. Security of Results.....	5
6. Data Forensics: detecting aberrant patterns in test data.....	6
7. Security incident response.....	7
8. Internet Screening.....	8
9. Data Forensics: following a suspicion of fraud.....	9
10. Performing Security Audit.....	10

**Fig. 20.4** Body of content EDF protocol

The second adjustment was the introduction of an evidence table for each standard. This adjustment was based on two potential advantages. 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 initiate a discussion. For example, to check for potential security risks, or to determine if a change in practice is needed. 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.

### *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. 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 assessor integrity 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). This choice is based on the

fact 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. Reasons 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 a criterion. 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, therefore the latter (use of materials) was excluded.

### *Grading adjustments*

In all interviews, on various topics, several experts stated that 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. In order to make the protocol more manageable, only a clear description of the ‘sufficient’ level was now included in the rubric. The descriptions of the other levels have become fixed, namely: (1) Insufficient: the described criteria are not met; (2) Good: the criteria are amply met/demonstrates how 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.

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 criterion. 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 block an ‘X’ will automatically be applied. This makes the protocol slightly more user-friendly and more intuitive.

### *Step 5—Implementation of the EDF Protocol*

During the fifth step, the *EDF Protocol* was used to evaluate and measure possible security risks within an exam program. 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 organized to implement the *EDF Protocol*. 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.

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 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. 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.

### 20.3.3 Discussion and Conclusion

This design research started on the premise of developing a set of standards, enabling practitioners to prevent and detect possible fraud 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.

By means of the five design steps carried out in this study, the main research question is unambiguously answered by stating that the *EDF Protocol* provides enough 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. 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.

## 20.4 General Discussion

Two empirical studies on educational data forensics have been presented in this chapter. The first study was an empirical study in which the effectiveness of a combination of EDF indices was tested. The main finding was that the combination of Guttman error model indices and log normal response time model indices were able to differentiate between non-cheating and cheating students. Using the right combination and the optimal cut-off scores for the indices, we were able to detect 37.5% of the cheating students, at a false positive ratio of 2.7% (the type I error). Finally, the findings indicated that especially preknowledge and the assistance of the proctor during the test were best detectable.

The second study had a more qualitative character. The goal of the study was to establish an educational data forensics protocol, consisting of evidence-based standards, that could be used by both researchers and practitioners. Through five design and development steps we were able to develop an *EDF Protocol* that consisted of ten standards, and a grading system that can be used to evaluate potential test security risks and breaches.

To conclude, we see EDF more as a continuous improvement process for test organizations, rather than a set of techniques or methodologies that are deployed ad hoc. Test fraud continues to advance, and improvements in test security should keep pace. In that sense, the *EDF Monitor* and the *EDF Protocol* can work in tandem. Potential breaches detected through the protocol can point to where to look in the data for test fraud, using the monitor. Conversely, the origin of detected test fraud with the monitor can be found by applying the protocol. Although test fraud will probably never be fully eradicated, the use of monitoring and security protocols allows test organizations to protect against misconduct.

## References

- 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.
- Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association*, 69, 44–49.

- Bellezza, F. S., & Bellezza, S. F. (1989). Detection of cheating on multiple-choice tests by using error-similarity analysis. *Teaching of Psychology, 16*(3), 151–155.
- Berkhout, E. E., Van der Werff, S. G., & Smid, T. H. (2011). Studie & Werk 2011. Amsterdam: SEO Economisch Onderzoek. Retrieved from [http://www.seo.nl/fileadmin/site/rapporten/2011/2011-29\\_Studie\\_en\\_Werk\\_2011.pdf](http://www.seo.nl/fileadmin/site/rapporten/2011/2011-29_Studie_en_Werk_2011.pdf).
- Bird, C. (1927). The detection of cheating in objective examinations. *School & Society, 25*, 261–262.
- Cassell, C., Denyer, D., & Tranfield, D. (2006). Using qualitative research synthesis to build an actionable knowledge base. *Management Decision, 44*(2), 213–227.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139–150.
- Hughes, J. M. C., & McCabe, D. L. (2006). Academic misconduct within higher education in Canada. *The Canadian Journal of Higher Education, 36*, 1–12.
- Lin, C. H. S., & Wen, L. Y. M. (2007). Academic dishonesty in higher education: A nationwide study in Taiwan. *Higher Education, 54*, 85–97.
- Maynes, D. (2013). Educator cheating and the statistical detection of group-based test security threats. In A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 187–214). New York, NY: Routledge.
- McClintock, J. C. (2015). Erasure analyses: Reducing the number of false positives. *Applied Measurement in Education, 28*(1), 14–32.
- McKenney, S., & Reeves, T. C. (2012). *Conducting educational design research*. New York, NY: Routledge Education.
- Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311–314.
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269.
- Piercy, K. W. (2015). Analysis of semi-structured interview data. *Department of Family, Consumer, & Human Development, 1–16*.
- Qualls, A. L. (2001). Can knowledge of erasure behavior be used as an indicator of possible cheating? *Educational Measurement: Issues and Practices, 20*(1), 9–16.
- R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>.
- Reise, S. P., & Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*(3), 217–226.
- Saupe, J. L. (1960). An empirical model for the corroboration of suspected cheating on multiple-choice tests. *Educational and Psychological Measurement, 20*(3), 475–489.
- Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181–204.
- Van der Linden, W. J., & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180–199.
- Van der Linden, W. J., & Lewis, C. (2015). Bayesian checks on cheating on tests. *Psychometrika, 80*(3), 689–706.
- Van Noord, S. (2018). *Using data forensics to detect cheating in randomized computer-based multiple-choice testing* (Unpublished master thesis).
- Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. London, UK: ACM.
- Wollack, J. A., & Fremer, J. J. (2013). Introduction: The test security threat. In A. Wollack & J. J. Fremer (Eds.), *Handbook of test security* (pp. 1–13). New York, NY: 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 Studiën, 88*, 23–40.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

