MASTER THESIS

USING DATA FORENSICS TO DETECT CHEATING IN RANDOMIZED COMPUTER BASED MULTIPLE-CHOICE TESTING

STUDENT
Sanette van Noord
Educational Science and Technology

GRADUATION COMMITTEE
dr. Hans (J.)W. Luyten
dr. Sebastiaan de Klerk
prof. dr. ir. Bernard P. Veldkamp

DATE
May 2018

UNIVERSITY OF TWENTE.

Table of Contents

Acknowledgements

The research for this master thesis was initiated by eX:plain, a testing agency situated in Amersfoort, The Netherlands. The client provided accommodation for two days a week for the period of 11 months, and assigned their scientific advisor to supervise the project, Sebastiaan de Klerk. I would like to thank him first for his time, strong input and guidance, which was a perfect combination of independence of the researcher and support in the research process. The master thesis was also supervised by Hans Luyten and Bernard Veldkamp from the University of Twente, whom I'd like to acknowledge for their interest in the study and valuable scientific input from a perspective other than that of those involved in the actual project. Additionally, I would like to thank all my colleagues at eX:plain for helping me during the research process, from organizing events, proctoring, and distributing forms, to reading drafts and listening to my complaining. And lastly, I would especially like to thank those colleagues that participated in this study, along with all other participants, and the educational institutions that provided their students and time. This research, unique in its field, could not have been completed without the efforts of eX:plain, their partners and employees.

Public Summary

Cheating is an existential problem in the testing industry; especially in high-stakes testing examinees are tempted to cheat, endangering the value of obtained credentials. Data forensics, data analysis methods to identify aberrant behaviour patterns that could be classified as cheating by an examinee, have been around for decades. However, most of these methods focus on fixed paper-based exams. The Data Forensics Tool software, developed by eX:plain, is able to analyse randomized computer-based multiple-choice tests. Using and adjusting the Guttman model (1944), eX:plain has developed six indices that detect behavioural patterns. Judging strong deviation from the regular behaviour pattern as potential cheating, analists at eX:plain have started investigations in the past year. Nonetheless, the quality, true detection rates and reliability of measurement, of such indices and software has rarely been investigated in practice; as it is impossible to know who did or did not actually cheat, researchers mainly use simulated data. To that matter the design of the current study is unique in its field: having known and instructed cheaters along with a control group of honest examinees take an existing test, with up to 2 years of information for benchmarking of normal and aberrant behaviour. Aside from being able to determine the true quality of the Tool, this provided the opportunity to finetune the software to detect 37.5% of all cheaters, with 96.8% reliability within the detected sample. The study also revealed the Guttman model itself to be rather unreliable to detect cheating, unless it is adapted. Based on current adjustments and results, further improvements of the software were suggested, including suggestions for automation of the analysis procedure and adapted behaviour measurements of response times and answer selection.

Keywords: test security, data forensics, cheating, randomized computer based multiple-choice testing

Introduction

Suppose we could ensure that nobody ever cheated on exams. And, if a candidate would obtain a test score through inappropriate means, that testing agencies would always be able to detect it and act on it. The validity and value of credentials, licenses, diplomas, and alike could be greatly improved and ensured (Impara & Foster, 2011; Kingston & Clark, 2014). Unfortunately, this is not yet the case. Testing agencies, such as eX:plain, take strong measures to prevent cheating. Randomization of items and, for multiple-choice, alternatives has proved to discourage cheating greatly, and with the introduction of Computer Based Testing (CBT) it is even possible to have candidates in the same room answer completely different sets of items (Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014). However, as test security progresses, cheating does too.

Cheating is committed on a larger scale than one might suspect. Evaluations of the extent of cheating in high schools, colleges and universities all over the world present numbers such as 85% of Taiwanese students admitting to copying (Lin & Wen, 2007), and 58% of Canadian high schoolers admitting to serious test cheating (Hughes & McCabe, 2006). A large scale evaluation in The Netherlands was done in 2010 and 2011 for college and university students. In a sample of 7000 students, 15% of college students and 10% of university students admitted to cheating at some point (Berkhout, van der Werff, & Smid, 2011). In other words, the problem of cheating is easily underestimated and must not be taken lightly.

Especially in high-stakes testing, examinees are tempted to cheat (Impara & Foster, 2011; Foster, 2013). High-stakes testing environments are “situations where decisions and interpretations from test scores have important, critical, and direct consequences for the test-taker (...)” (Association of Test Publishers, 2002, p. 3). For instance, eX:plain deals with professional certification, a referred example of a high-stakes testing environment. Candidates are often tempted to cheat as time can be limited and pressure to obtain the certificate is high (to get or keep a job). Commercial exam centers, where the exams are administered, often do not benefit from trying to prevent cheating attempts (Foster, 2013). The companies that provide the candidates for the exams (i.e., their employees) also pay the bills, and are more likely to continue to do so if there are no complications and the candidates acquire the certification fast enough (Impara & Foster, 2011). However, cheating prevention is crucial to the value of the obtained certification (Foster, 2013; Kingston & Clark, 2014), as well as overall proficiency of employees and safety on the job.

Therefore, eX:plain has been trying to find ways to detect cheating reliably themselves. With the introduction of CBT, statistical investigation of response patterns on the items in exams has gained popularity. In line with this, they designed their own cheating detection software tool: the Data Forensics Tool (DFT), as part of their test security services project Xquiry. The DFT incorporates several data forensics algorithms for computer based multiple-choice testing, and can be used to analyse students’ test response patterns for potential cheating. However, the reliability of tools like the DFT has hardly been validated in scientific research (Wollack & Fremer, 2013). The indices are partially deduced from theory, however they are hardly tested in practice, other than with simulated data (Zopluoglu, 2017; e.g., Belov, 2015, Meijer, 1994). This simulated data can be considered biased as they are developed with the theoretical aberrant behaviour patterns of examinees in mind, while this pattern is only theoretical and has not been proven to identify cheating in real life. Other indices have been designed by specialists at eX:plain, also solely based on theory, and have therefore never been studied before. This research produces real data of cheating behaviour patterns that offers great advancements in the field of data forensics, reviewing the current theories and improving them.

This research focusses specifically on the quality of the indices that comprise the DFT software. It needs to be determined what the predictive validity is of the data forensics algorithms and which constraints provide the most accurate estimation. The reliability of the DFT needs to be established by investigating and minimalising the misclassification error. Furthermore, the most distinctive patterns for cheating need to be isolated, exploring the future usability of the tool. In a more practical sense, this research provides eX:plain with data on the performance of the current tool and recommendations for improvement.

Theoretical Framework

In this section, cheating, cheating detection using data forensics, and the data forensics software central to this research are discussed extensively. This study focuses specifically on randomized computer based multiple-choice testing; the theoretical framework will focus thereon.

The Context

There is no set of guidelines to administer computer based multiple-choice tests. The administration is based on the existence of corresponding infrastructure; computers and software. In case of randomized computer based exams, this software assembles different sets of test items for each candidate (Marianti et al., 2014. Further test security measures are determined by the school or testing agency itself. Many studies recommend certain measures, such as appropriate physical distance between examinees, secure browsers, identification of individuals, and prohibition of private appliances (Clark & Kingston, 2014). Examinees are usually supervised by a proctor; some institutions choose to install additional camera supervision.

EX:plain conforms to most known measures. Computers in the exam room are situated in a physically and electronically supervised room, at an appropriate distance from each other. Candidates are assigned their seat, and required to be able to identify themselves with proper documentation. They are not allowed to bring any other objects into the room, such as phones, calculators, watches, large jewelry, pens, or paper. They receive an extensive instruction about the procedure of the exam an use of the software from their proctor before they start their timed exam. Each candidate gets a different set of items that is assembled from a large (usually 1000+) item bank and randomized by a digitally secured computer program. There is no talking during the exam, and restroom breaks are not allowed.

Cheating

While papers on cheating detection usually refer to the detected as "test fraud", the term is best divided in two categories: cheating and piracy (Foster, 2013; Impara & Foster, 2011). This study focusses solely on cheating, as the extend of test fraud as a whole exceeded the scope of this study. Piracy entails methods of item theft; a reflection on piracy is found in the discussion section.

A comprehensive definition of cheating is: "obtaining a test score through inappropriate means" (Impara & Foster, 2011, p. 93). Cizek (1999) roughly summarizes all literature on cheating published between 1970 and 1996, and thereby provides a comprehensive overview of cheating methods candidates have found and developed over the years. As his work is almost 20 years old and this research focuses solely on randomized computer based multiple-choice testing, this section builds on his taxonomy, adjusting it to the context, recent developments, and new threats. A comprehensive overview can be found in Table 1.

Table 1

A Comprehensive Overview of Potential Cheating Methods for Randomized Computer Based Multiple-Choice Testing, based on Cizek's (1999) Taxonomy and edited with New Perspectives from Recent Literature

Tag	Description
Giving, Taking, and Receiving information (GTR)	
Using code	Using a beforehand discussed sign language to sign answers to other candidates (GTR5+7). Using technology to sign code, such as laser pens to point out answers on posters on exam room walls (GTR8).
External collaboration	A companion resides outside a window and they communicate the questions and answers with each other (GTR12)
Using a smart device	Using a smart device that enables the candidate to send or/and receive information (GTR15), such as smart phones or calculators using infrared (Impara & Foster, 2011; Foster, 2013).
Pre-knowledge	Having detailed knowledge of test items that are possibly on the test and studying them in preparation of the test. Compromised items are either publicly attainable on the internet, received through peers, or bought from those who illegally obtained them (Impara & Foster, 2011; Ferrara, 2017; Foster, 2013).
Forbidden Materials (FM)	
Using cheatsheets	Using a hidden paper cheetsheet (FM1; Ferrara, 2017; Foster, 2013), written notes on personal articles (FM4), a desk, chair or the floor (FM5), the body

	(FM9), decorative accessories (FM12), tissues, or gumwrappers (FM13). A cheetsheet can also be an audio recording, being played on a concealed earphone (FM18).
Taking Advantage of the Process (TAP)	
Identity fraud	A substitute taking a test for another person (TAP4; Ferrara, 2017). This proxy candidate may be a friend or acquintance of the original candidate, or may be provide by proxy test taking services (Foster, 2013).
Internal collaboration	The candidates discuss the test as the proctor momentarily leaves the room (TAP7). Or leaving notes, answers, or questions in restroom facilities for others to find (TAP22).
Proctor assistance	Getting information out of the proctor by asking questions (TAP9; Foster, 2013).

Note. Even though paper and smart devices are not allowed, candidates smuggle them in. Also, proctors can be persuaded to allow restroom breaks, even though they are not supposed to.

Data Forensics

The term ‘data forensics’ refers to a collective of data analysis methods to identify cheating, defined as aberrant, therefore potentially fraudulent, response behaviour (Simon, 2014, p. 83). The statistical methods used vary from simple univariate methods to model-based multivariate and nonparametric techniques. The method used is usually adjusted to the problem at hand, hence the great variety. The earliest publication on data forensics methods are dated back to the 1920s, examining the similarity of response pattern as evidence of copying and collusion (Clark & Kingston, 2014; e.g., Bird, 1927, 1929; Crawford, 1930). The analyses focussed on error-similarity in pairs; candidates copying each other would not only copy correct but also incorrect answers. In 1974 Angoff published his highly influential work on comparing the effectiveness of several copying indices that had been introduced in the past time. Publication of copying detection methods in the 70s, 80s and 90s of the 20th century largely built on Angoff’s findings and indices were introduced using similarity in incorrect as well as correct answers (e.g., Frary, Tideman, & Watts, 1977; Bellezza & Bellezza, 1989; Holland, 1996; Wollack, 1997).

Until the year 2000 most cheating detection efforts focussed on copying. With greater awareness, the introduction of the No Child Left Behind Act in the United States of America (USA), and CBT in (high-stakes) assessment, new approaches were explored at the start of the 21st century (Clark & Kingston, 2014; Mroch, Lu, Huang, & Harris, 2014). For paper-based tests erasure-tracking became largely important, as it was discovered that teachers in US schools would change student answers after the test, to upgrade the school’s national performance (Mroch et al., 2014). An abnormal rate of erasures in a group could be an indication of cheating. Furthermore, research in the field of data forensics started focussing on detecting pre-knowledge, by tracking changes in performance over time (e.g., Belov, 2005; Impara, Kingsbury, Maynes, & Fitzgerald, 2005), and response-time modeling, exploring the new possiblities of the additional behaviour that could be measured in CBT (Marianti et al., 2014; Van der Linden, 2006). The new methods offer opportunities of detecting cheating by indivuals, not just in groups.

In the context of randomized computer based multiple-choice testing, the similarity analysis methods are considered irrelevant, as copying has become almost impossible due to the randomization of items and alternatives (Marianti et al., 2014). Most CBT software randomizes the adminstered exam, and only logs the final answer submitted by the candidate and the corresponding response time. Little research has been done for cheating detection in randomized exams, as the technique is fairly new. The DFT includes those methods relevant to the random computer based multiple-choice exams, appropriate within the available data. They are discussed in the next section.

Data Forensics Tool Psychometric Indices

The DFT software was designed by eX:plain based on several different scientific publications (i.e., Guttman, 1944; Meijer, 1994; Van der Linden, 2006) on data forensics relevant to randomized computer based multiple-choice testing, and contains six indices to measure behavioural patterns. Analyses are done on an individual level; the software is to identify aberrant patterns on the distribution of aggregated individual data.

The indices are all based primarily on the Guttman sequence, the ordering of items in a test according to difficulty (least to most difficult). The difficulty of each item is determined by the proportion correct (*p*-value) of the items (Guttman, 1944; Sirotnik, 1987; Meijer, 1994). The software computes these values over a two year dataset of recorded responses. A simple worked example of a test and calculations based on the Guttman theory is used in this section to illustrate the equations of the indices, which are mere presentations of the algorithms in the software. Running one algorithm is referenced as one analysis. Detecting cheating with the DFT software is most likely to involve running several algorithms, doing serveral analyses, and combining the outcomes.

Worked example. Consider a group of five persons taking a ten item exam. Their correct (1) and incorrect (0) answers are represented in Table 2. The items of this test are ordered to difficulty, determined by the *p*-value; the first item is the easiest and the last one is the most difficult. To work with the Guttman sequence, the base of all indices, the items should be ordered as such (Guttman, 1944). In reality, exams are not generally organized this way, therefore the software constructs the sequence itself.

Table 2

Results of Example Sample of Ten Candidates completing a Ten Item Test

Person Item	A	B	C	D	E
1	1	1	1	1	1
2	1	1	1	0	0
3	1	0	1	0	1
4	1	1	0	0	1
5	1	0	1	0	0
6	1	1	1	1	1
7	0	1	0	1	1
8	1	0	1	1	1
9	0	0	0	1	0
10	0	1	0	1	1

The Guttman error. The Guttman sequence is used to compute the raw Guttman error (*G*), the sum of all correct responses after each incorrect response in the sequence, as in Equation 1 (Meijer, 1994, p. 312). In the equation, *k* is the number of items involved, *g* is the number-correct on item g, and *h* is the number-correct on item h. *G* occurs when a respondent posits an incorrect response, when a correct response was expected based on the ability level of the individual and the difficulty of the items. Equation 1 is merely the programmed algorithm in the software

$$G = \sum_{g=1}^{k-1} \sum_{h=g+1}^{k} fgh \qquad (1)$$

By hand *G* is simply computed by counting all correct responses after each incorrect response in the sequence and adding them up. Consider Person A (Table 2). From the top, item 7 has the first incorrect response. Out of the following items, 8 to 10, only item 8 was answered correctly, so count one Guttman error, a correct response after an incorrect response. The next incorrect response in the sequence is on item 9, however there are no correct responses after item 9, nor after item 10. The raw Guttman error for Person A will be *G* = 1. A low Guttman error indicates that the sequence proceeded how one would expect it to; easy questions answered correctly until the ability level (a *p*-value somewhere on the sequence) is reached and the candidate starts submitting incorrect responses because the items have become too difficult. This assumes correct responses after incorrect responses to be irregular, as the incorrect response should indicate that the ability level has been reached.

Consider Person D for instance. The first incorrect response on the sequence is on item 2. After item 2 follow five correct responses, on items 6 to 10; count five Guttman errors. Again following the sequence, the next incorrect response is on item 3. Again, five correct responses follow, on items 6 to 10; count another five Guttman errors. The raw Guttman error is now ten, adding up the errors on items 2 and 3. The next item in the sequence is the incorrect response on item 4, followed by again five correct responses, as well as for the incorrect response on item 5. Finally, the raw Guttman error adds up to 20, which is quite large. It is highly suspicious that a candidate would answer many relatively easy questions incorrectly, while submitting correct responses to the more difficult items. This is therefore considered an aberrant behavioural pattern according to the Guttman model (1944), a potential indication that this person has cheated, represented by a high Guttman error (Meijer, 1994).

The Guttman score. However, results cannot be compared on an individual level by using *G*, because the distribution of the number of items is unclear and different for every error. Therefore, the raw Guttman error is solely input for the first data forensics index in the software: the Guttman score (*G**) as computed in Equation 2 (Meijer, 1994, p. 312). *G** is *G* corrected for the number of items (*k*) on the exam and the number-correct score of a person (*r)* as the computation of *r(k-r)* equals the maximum *G* for the individual's number-correct score.

$$G* = \frac{\sum_{g=1}^{k-1}\sum_{h=g+1}^{k} fgh}{r(k-r)} \qquad (2)$$

The Guttman score for Person D on the ten item test would be: $G* = 20/(6(10\text{-}6)) = 0.83$. A high Guttman score still indicates an aberrant pattern, but the computations are now comparable to each other.

The Guttman score corrected for distance. The second and third indices are *G** corrected for the distance between items in the Guttman sequence, as computed in Equation 3. The distance can be indicated in a number of positions between items (second index; $G*_{dpositions}$) or *p*-value between items (third index; $G*_{dpvalue}$). The distance is set manually in the software by the assigned data analist before running the algorithm, and can be run with any given value within the acceptable range; no more positions than are actually in the sequence, or no more *p*-value than the maximum of the sequence. It is yet unclear what exact settings are ideal for detecting cheating.

$$G*_{d} = \left(\frac{\sum_{g=1}^{k-1}\sum_{h=g+1}^{k} fgh}{r(k-r)} \middle| distance\right) \qquad (3)$$

The Guttman score corrected for response time. The fourth index corrects the *G** for response time, see Equation 4.

$$G*_{rt} = \left(\frac{\sum_{g=1}^{k-1}\sum_{h=g+1}^{k} fgh}{r(k-r)} \middle| response\ time\right) \qquad (4)$$

The computation of the response time involves a formula (see Equation 5) based on the log-normal model for response times by Van der Linden (2006). The model considers the individual features of the candidate, comparing the actual total time on the test to the total time expected, represented by the average workpace per item (*Wp*). This workpace is then multiplied by the mean response time of the item, again based on a two year dataset, as the expected item time ($E(Item_{time})$), to calculate the expected item time for the specific candidate. By subtracting the observed item time (O_{time}), the aberrance of the response time per item in the sequence can be identified and corrected for by the index.

$$\frac{Wp*E(Item_{time}) - O_{time}}{Wp*E(Item_{time})} \qquad (5)$$

The Guttman score corrected for response time and distance. The fifth and sixth indices combine the previous indices, correcting *G** for response time as well as distance, as seen in Equation 6. The fifth index combines the correction for distance in positions with that for response time ($G*_{rtdpositions}$), the sixth in *p*-value ($G*_{rtdpvalue}$).

$$G*_{rtd} = \left(\frac{\sum_{g=1}^{k-1}\sum_{h=g+1}^{k} fgh}{r(k-r)} \middle| distance, response\ time\right) \qquad (6)$$

Detection of aberrant behavioural patterns. For each of the six indices a mean and corresponding standard deviation over all candidates are computed as a benchmark for the 'normal behavioural pattern measured by the index. For each candidate it is determined how many standard deviations from the mean their index scores are to get an impression of how aberrant their behavioural pattern is. The cheating detection of the software is based on these *z*-scores for the index. The *z*-scores will be referred to as "deviation scores" for the remainder of the thesis, to mark their purpose in detecting aberrant behaviour. It is theorized that an index score further away from the mean, is more likely to indicate cheating; the further away the deviation score is from 0, the more aberrant the behavioural pattern. Thereby the analysis resonates the statement that aberrant behaviour patterns can be an indication of potential cheating (Simon, 2014).

However, the cut-off (deviation) score is yet to be established for each index; it is unclear how aberrant a behavioural pattern is to be an indication of cheating. Angoff (1974) was the first author to propose a cut-off score in his data forensics research on similarity analyses. He reasoned for a very conservative cut-off deviation score of 3.72 *SD*. However, according to Maynes (2017) it is currently a decision that "should adhere to accepted scientific practice and abide by the organization's goals and responsibilities" (p. 54). Also, sustaining just one cut-off score for all analyses is presumably too simplistic. The totality of the individual evidence, the consequence of misclassification errors, and the organization's ability to implement policy decisions should be taken into account in accusing candidates of cheating.

The deviation scores are computed even though the Guttman score (*G**) itself was supposed to make individuals comparable to each other. This is because the Guttman model is deterministic (Guttman, 1944), the theoretical patterns are ideal and almost never realised in practice. A pattern displayed by Person C in Table 2 of the worked example is probably more realisic than that of Person A. Therefore it is useful to calculate a mean pattern, and a standard deviation, to be able to differentiate between examinees that are simply displaying a realistic pattern that is solely aberrant from the ideal, or a pattern seriously aberrant from reality. This is, of course, based on the assumption that most examinees (the mean) do not cheat.

Research Questions and Model

Analists at eX:plain have been using the DFT to analyse for potential cheating on the random computer based multiple-choice exams they regularly administer. Detected examinees are currently not accussed of cheating directly, as the functionality of the DFT has not been established. Instead, detected examinees are compared in search of similarities such as testing time, location, or proctor. Upon the information found, additional investigations are set in place, researching the causes of the abnormalities detected. This way, irregularities in examinations have already been tracked down to suspicious exam locations and shady proctoring activities. However, the analists have the urging question what the analyses that they are currently conducting truely detect. Did all the detected examinees really cheat? And does the DFT detect all cheaters that are actually out there? Therefore, the first research question for this study reads:

> Research Question 1: *What is the quality of the current data forensics analyses conducted with the DFT?*

This will put the data forensics analyses that have been done in the past in the correct perspective.

In further assessment and exploration of the DFT, with the goal of improvement of the current analyses and software, two additional research questions complete the research model of this study:

> Research Question 2: *What manual setting(s) and interpretation(s) of the indices in the DFT comprise the best quality data forensics analyses?*
> Research Question 3: *How could various methods of cheating be detected by the data forensics analyses with the DFT?*

Method

This research was conducted using mixed methods. The research design covers two phases; an assessment and an explorative phase. Within these phases the participants in this study were grouped differently in order to answer different questions in the research model. In the assessment phase, visualized in Figure 1, the first research question was addressed; what is the quality of the current data forensics analyses conducted with the DFT? The control group consisted of regular, highly supervised examinees and the experimental group consisted of violators, counterfeit examinees instructed to cheat. Both were evaluated by the DFT, to see whether the violators were, or could be, succesfully distinguished from the regular test takers. This also provided input for the second research question: what manual setting(s) and interpretation(s) of the indices in the DFT comprise the best quality data forensics analyses?

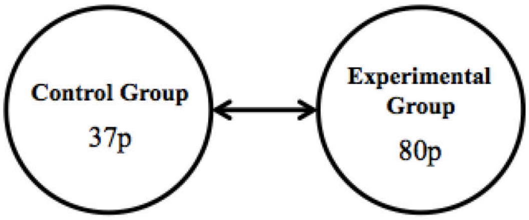

Figure 1. The assessment phase of the research design, comparing the control group with the experimental group.

In the explorative phase of the research design, visualized in Figure 2, qualitative data in a questionnaire on how cheating occured in the experimental group was used to regroup the participants, to answer the third research question on the detection of various methods of cheating. Five cheating methods were simulated within the experimental group: (1) using a smart device (smart phone), (2) internal collaboration (proctor leaves examroom), (3) proctor assistance, (4) using cheatsheets (notes on paper), and (5) pre-knowledge. In this phase, results from the first phase were evaluated in detail.

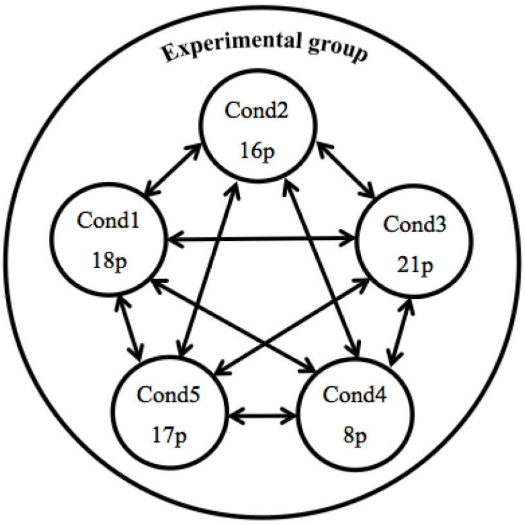

Figure 2. The explorative phase of the research design, comparing the different cheating conditions within the experimental group of the assessment phase.

Participants

Control condition. Participants in the control condition were recruited by means of an information form handed out before each exam. Those willing to participate were to sign the informed consent form (Appendix A). Underage (18-) exam candidates were not allowed to sign the form. In case they did, these forms were annulled. In the initial sample (n = 52), fifteen candidates were found to have taken a different test from the one designated for this research, possibly due to a logistical mix up at the testing location. The concerning participants were hence removed from the sample, resulting in a final sample of 37 participants (age: *M* = 39.30 years, *SD* = 13.70; 14 females) in the control condition. Considering all participants requested to take the Dutch exam, it was assumed they were fluent or either proficient in that language.

Experimental condition. To recruit participants for the experimental condition, several Dutch educational institutes were offered a trial exam for Dutch students in the relevant course. The knowledge and skills of these students were assumed to be insufficient to complete the exam successfully, since they had not completed the course at this time. This was deemed beneficial to the experiment, as actual cheaters would not attempt to cheat if they were confident their knowledge was sufficient to complete the exam a fair way (Miller, Murdock, Anderman, & Poindexter, 2007). However, due to existential value of the test items, participants in the fifth condition (i.e., pre-knowledge) could not be actual students. For test security reasons is would be highly undesirable to have regular students gain pre-knowledge of the test items of an official exam. Therefore, four employees at eX:plain were selected to take five trial exams each, on the condition that they were familiar with the item bank.

Participants in all experimental conditions signed the informed consent form (Appendix A). Parents and/or caretakers of underage students in the experimental condition were consulted for passive consent through their educational institute. After a brief review of the completed questionnaires on cheating, thirty-one participants were found to have disregarded the extensive instructions on cheating. They either participated in other activities (*"I wanted to see whether I could pass the test checking all the longest answer options."*), indicated they did not try to cheat as it would not have helped them perform better on the exam (*"Non of the items asked about the information on my cheatsheet."*), or just completed the exam without cheating (*"I completed the exam all by myself."*). The remaining sample (*n* = 80) included all five conditions, smart phone (*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), cheatsheet (*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).

Ethical concerns. Upon registration, all participants were assigned a candidate number by eX:plain's registration system, based on their date and time of entry. These numbers were used to subtract the appropriate data from the DFT. For optimal anonymisation and client protection, the list of candidate numbers was randomized and reassigned participant numbers unrelated to registration, and/or research condition.[1]

Materials

Test. Actual Dutch test items from the Basic Competence Legal Knowledge for Extraordinary Detective Officers (Dutch: Basisbekwaamheid Rechtskennis voor buitengewoon opsporingsambtenaren) item bank were used (230 items). The tests administered in the period the research was conducted had an average *p*-value of .69, an average RIT-value of .28 and an average reliability of α = .76 (this computation does not account for the variation within the individual exams because each set is unique). The items were selected randomly from the item bank by Questionmark Perception (QMP) software, with the usual restrictions based on the test matrix, and presented in the secure browser. That way, all participants each took a different test of 50 items. The trial exams in the experimental condition were no different from the exams in the control condition, except for that the candidates in the experimental condition were not eligible for the certificate that is normally acquired through passing this test.

Questionnaire. The participants in the experimental group completed a questionnaire (Appendix B) following their trial exam, to collect information on the execution of the instructed cheating method. The participants were asked in Dutch to identify their method of cheating and describe the situation as detailed as possible, given hints on what information was expected.

Procedure

The experiment was run in several individual and group sessions of approximately 90 minutes. Groups consisted of participants in the same experimental condition or the control condition. Participants in different conditions were never combined in one examroom. Only the participants in the

[1] This research was approved by the Ethical Commitee of the University of Twente before it was conducted.

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 the researcher.

All participants were informed on the research goals, method, and consequences of participation, after which the participants signed the informed consent form (Appendix A). 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., smart phone) the participants were told to consult their smart phone during a period of three minutes. This moment was announced after the first 15 minutes of the exam had passed. The violators were not allowed to talk outloud 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 occurance 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 where 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 students, 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., cheatsheets) the violators were asked to bring a cheatsheet 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.

After the participants in the experimental conditions finished their trial exam, they were handed the questionnaire. Names were registered on the informed consent form as well as the questionnaire to be able to link the information to the recorded data. After filling out the questionnaire, the experiment had finished and the participants were allowed to leave the room. Some participants in the pre-knowledge condition failed to fill out the questionnaire, as the sessions were not proctored. Therefore these participants were informally questioned orally after all trial exams were finished.

Data Analysis

Assessment phase. The data used was collected in the last quarter of the year 2017. Because in this quarter only 450 tests were administered, of which 111 were trial exams, it was deemed undesirable to compute deviation scores based on the means and standard deviations over this period. Therefore, the second and third quarter of the same year were used as the benchmark to base reliable means and standard deviations on to conduct all analyses in this study. The actual means and standard deviations used, calculated over a set of 1385 different tests administered, cannot be reported, as the examinees concerned have not signed informed consent. It is also undesirable to publish any information to help fraudsters elude cheating detection methods such as those researched in this study. Participant scores on each index were translated to a deviation score, the number of standard deviations from the mean. These scores were then used to compute three quality assessment parameters.

Quality assessment parameters. All analyses were assessed using three parameters. First, it is tested whether the mean deviation score in the fraudulent condition is significantly different from the non-fraudulent condition. This is evaluated using independent *t*-tests ($\alpha = .05$). This parameter indicates the validity of the analysis and is a condition of usage. Unvalid analyses are not assessed any further. The quality of usage of the valid analyses is described in detail by the other two parameters.

The second parameter, by name of 'detection', is the percentage of participants in the fraudulent condition that was successfully detected by the analysis according to the determined cut-off score. The detected sample of each analysis is determined by simply sorting the participants by their assigned deviation score from most to least, labelling each participant with a deviation score equal or higher than the cut-off score as fraudulent. The goal is to finely represent the chances of cheaters being detected by the analysis.

The third parameter, by name of 'reliability', is the percentage of true positives, rightfully detected cheaters, in the detected sample. It is important to note that the reliability does *not* counter represent the percentage of participants in the control group that would be wrongfully accused. The reliability percentage is based on the *total detected sample*, composition of which is descibed in the previous paragraph. To illustrate; as the detected sample increases but the number of falsely detected participants within this sample remains the same, the reliability of the analysis increases, as the percentage of falsely accused participants in the detected sample decreases. It was determined that this was the most informative way to represent false detection by the analyses.

Procedure. The analysis procedure for the assessment phase is responsive; the steps taken are based on earlier results in order to find the best possible combination of cheating detection analyses. First, the initial analyses that are regularly executed by eX:plain analists were conducted on the data in this research. The cut-off scores of these analyses were then altered to gain better results. In follow-up analyses the settings of the distance of the indices in either positions or *p*-value were altered to try to find even better results. These follow-up analyses were used to reach a conclusion on the best combination of cheating detection analyses possible with the current software. For the $G*_{dpositions}$ index 21 different settings were assessed (5 to 25 positions), for the $G*_{dpvalue}$ index 20 different settings were assessed (*p*-values .01 to .20), for the $G*_{rtdpositions}$ index another 21 settings were assessed (5 to 25 positions), and for the $G*_{rtdpvalue}$ index a total of 25 different settings were assessed (*p*-values .01 to .25).

In case the assumption of normality for the independent *t*-test for the validity parameter of an analysis was violated, the non-parametric equivalent, a Mann-Whitney *U* test, was conducted instead, in case transformations of the data did not produce considerable results (Field, 2013). Out of the 89 analyses assessed, for only four indices the deviation scores data distribution appeared to be normal, therefore their results for the normality tests are reported instead. The distribution of the data for the *G** index appeared to be normal in both the control condition, $D(36) = .11$, $p > .200$, and the fraudulent condition, $D(79) = .09, p > .200$. For the $G*_{d5}$ index (Guttman score corrected for distance in 5 positions) the distribution of data also appeared to be normal in the control condition, $D(36) = .11, p >.200$, as well as the fraudulent condition, $D(79) = .09$, $p = .16$. On the $G*_{d0.01}$ index (Guttman score corrected for distance in .01 *p*-value; *p*-value displayed with an additional 0 to facilitate better differentiation from the indices with positions settings in reading) this was also the case for both the control condition, $D(36) = .12, p = .199$, and the fraudulent condition, $D(79) = .08, p > .200$. And finally, the distribution of the data on the $G*_{d0.02}$ index appeared to be normal in the control condition, D(36) = .13, $p = .111$, as well as in the fraudulent condition, $D(79) = .09$, $p = .088$. All other data appeared to be non-normally distributed according to Kolmogorov-Smirnov Goodness-of-Fit Statistics ($p < .05$; see Appendix C)

Explorative phase. Grouping the experimental condition according to Figure 2, a MANOVA was conducted to determine whether the different kinds of cheating can be identified differently by the analyses established in the previous research phase. However, the assumption of normality could not be met for all groups for all analyses (see Appendix D), so multiple Kruskal-Wallis ANOVAs were conducted instead, since no non-parametric substitute is available for MANOVA (Field, 2013).

In addition, each group was assessed seperately, reviewing the detection rate within the sample for each analysis, and the validity of the analyses per group determined by Mann-Whitney *U* tests. To try to determine what valid analyses are potentially better than others to identify a specific group, a Friedman two-way ANOVA, as the non-parametric replacement of a Repeated Measures ANOVA (Field, 2013), was performed on the valid analyses for each group, if possible followed up by pairwise comparisons using, also non-parametric, Wilcoxon Signed Rank tests. In the case only two analyses were found valid, this pairwise comparison was performed directly. A short summary of the answers submitted in the questionnaire was provided for further context to the division between the detected and the undetected sample in each experimental group.

Results

Assessment Phase

Initial analyses. The first analyses in the assessment phase were conducted as they are normally conducted by eX:plain analists. Settings for distance in positions (20) and *p*-value (.20) were both used for the indices corrected for distance as well as those corrected for response time and distance. The cut-off score used for all analyses is 3 *SD*. All seven analyses indicated significant validity, as presented in Table 2. If all candidates scoring above three standard deviations in the analyses were to be flagged as cheaters, 28.8% of all cheaters in the sample would be correctly identified with a reliability of 92.0%, the percentage of rightfully convicted candidates in the detected sample.

In order to construct the best possible procedure with the current analyses, it was estimated that the G^*, the G^*_{d20}, and the $G^*_{d0.20}$ indices should be removed from the set for adding more negative value (violating reliability) than positive value (increasing detection). Altering the cut-off score, the G^*_{rt} index was estimated to provide the most valuable analysis, adding an initial 30% detection, with 96% reliability for a cut-off score of 1.9 *SD*. To order, the indices G^*_{rtd20} (2.5 percent points additional detection, cut-off score 3.1 *SD*), and $G^*_{rtd0.20}$ (1.3 percent points additional detection, cut-off score 1.9 *SD*), comprise a set of analyses with 33.8% detection combined and 96.4% reliability.

Follow-up analyses. Settings for the G^* index, previously determined to be invaluable, and the G^*_{rt} index, the base of the previously established procedure, could not be further adjusted to improve detection or reliability of the data forensics analysis. Therefore, analyses with different settings of the other four indices were assessed to this purpose, while retaining the G^*_{rt} index analysis. Validity of two follow-up analyses could not be established, leaving them excluded from further assessment. In the G^*_{d25} analysis the deviation scores in the control group (*Mean Rank* = 50.59) did not differ significantly from the deviation scores in the fraudulent group (*Mean Rank* = 62.89), *U* = 1169.00, *z* = -1.84 (corrected for ties), *p* = .066, two-tailed, *r* = .17. Neither did the deviation scores in the control group (*Mean Rank* = 50.88) differ significantly from those in the fraudulent group (*Mean Rank* = 62.76) for the $G^*_{rtd0.23}$ index, *U* = 1179.50, *z* = -1.82 (corrected for ties), *p* = .069, two-tailed, *r* = .17. Validity for all other settings of the indices was confirmed (see Appendix E).

For the validity of all remaining analyses was established with the deviation scores in the fraudulent group significantly higher than those in the control group (see Appendix E), a positive cut-off score was retained for the remainder of the process. In addition, the cut-off score for all analyses was lowered to 2 *SD*, as a lower cut-off score seemed more beneficial in most of the previously assessed analyses. Cut-off scores were only adjusted further if the reliability of the additional sample were 50.0% or more. The highest possible cut-off score is always preferred.

All follow-up analyses are assessed using graphs (Figures 3 to 6), displaying the change in detection rate for each setting if it were added to the set of initial analyses selected previously (G^*_{rt}, G^*_{rtd20}, and $G^*_{rtd0.20}$), and similarly the changes in the reliability of the set. They are compared to the original detection and reliability, to assess whether their addition would be benefitial. The G^*_{rtd20} index is removed from the initial set to assess all settings of the $G^*_{rtdpositions}$ index, the $G^*_{rtd0.20}$ index is removed from the set in assessment of all settings of the $G^*_{rtdpvalue}$ index.

Table 3

Descriptives and Results from the Initial Data Forensics Analyses

	Not cheating (N = 37)			Cheating (N = 80)								
Index	*M*	*SD*		*M*	*SD*		t(115)		*p*	Cohen's *d*	% Detection	% Reliability
G^*	0.08	1.16		0.88	1.08		-3.65		<.001	.76	5.0	80.0
			Mean Rank			*Mean Rank*	*U*	*z*		*r*		
G^*_{d20}	0.52	1.41	49.07	0.99	1.51	63.59	1112.50	-2.16	.031	.20	7.5	75.0
$G^*_{d0.20}$	-0.40	0.91	48.96	0.02	1.10	63.64	1108.50	-2.19	.029	.20	3.8	100.0
G^*_{rt}	0.01	1.04	39.34	1.60	2.15	68.09	752.50	-4.28	<.001	.40	23.8	100.0
G^*_{rtd20}	0.41	1.10	46.64	1.87	2.75	64.72	1022.50	-2.71	.007	.25	26.3	91.3
$G^*_{rtd0.20}$	-0.34	0.75	46.22	0.75	2.29	64.91	1007.00	-2.84	.005	.26	13.8	100.0

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

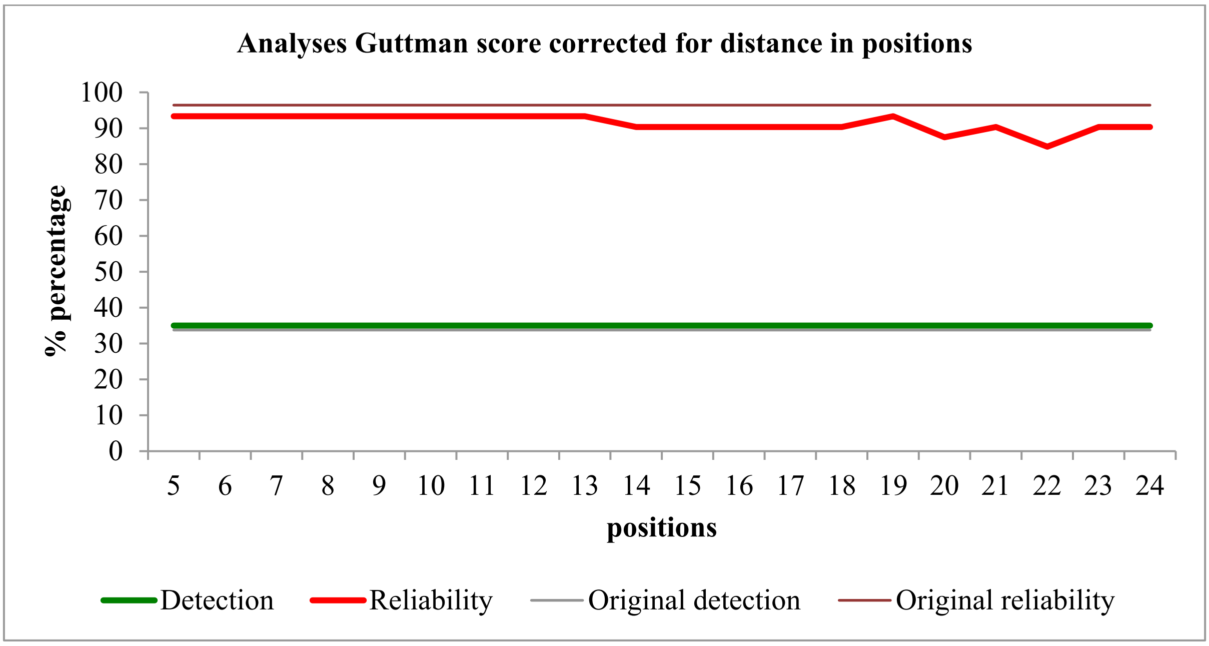

Figure 3. Graph displaying the results of analyses with the Guttman score corrected for distance in positions index with 20 different settings and a cut-off score of 2 SD, compared to the results of the initial analyses determined previously, which excludes this index originally.

Evaluation of the detection and reliability of 20 different analyses with the $G*_{dpositions}$ index with a cut-off score of 2 *SD* (see Figure 3), thus assessing 20 different position settings, determined that this index could not improve detection (green) without deteriorating the reliability (red) of the final analyses. Evaluating analyses with the same index with *p*-value settings (see Figure 4) proved that the $G*_{dpvalue}$ index could not be of any value to cheating detection at all, as none of the analyses increased detection of the initial analyses set and some only decreased its reliability.

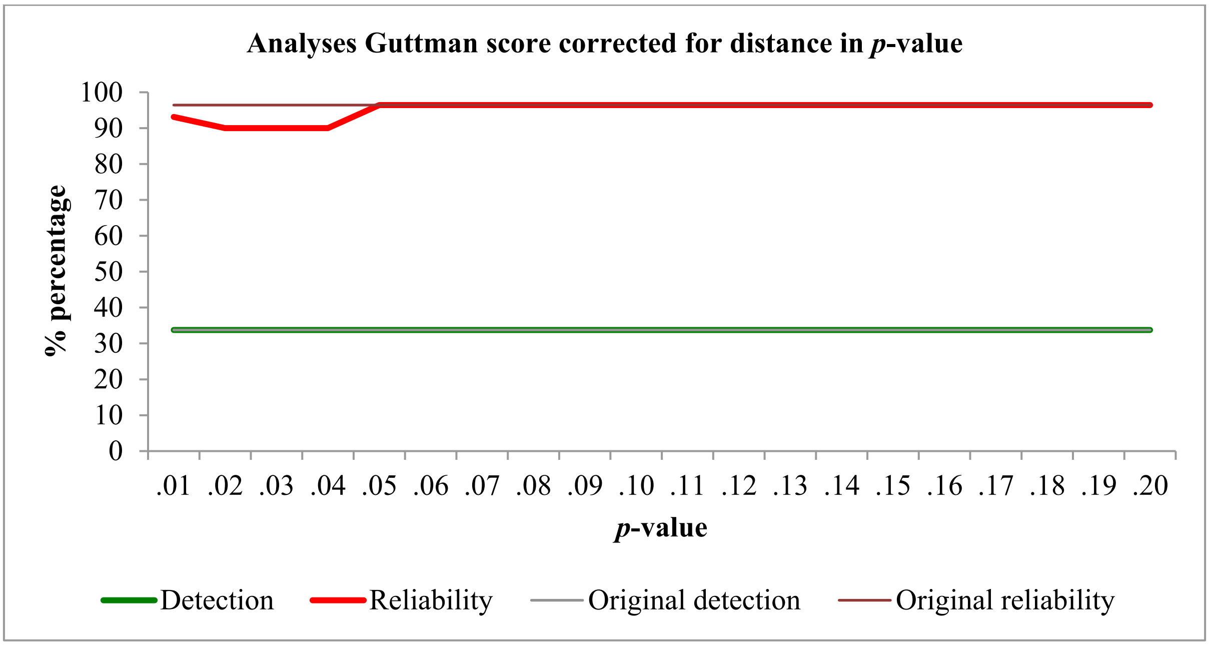

Figure 4. Graph displaying the results of analyses with the Guttman score corrected for distance in *p*-value index with 20 different settings and a cut-off score of 2 SD, compared to the results of the initial analyses determined previously, which excludes this index originally.

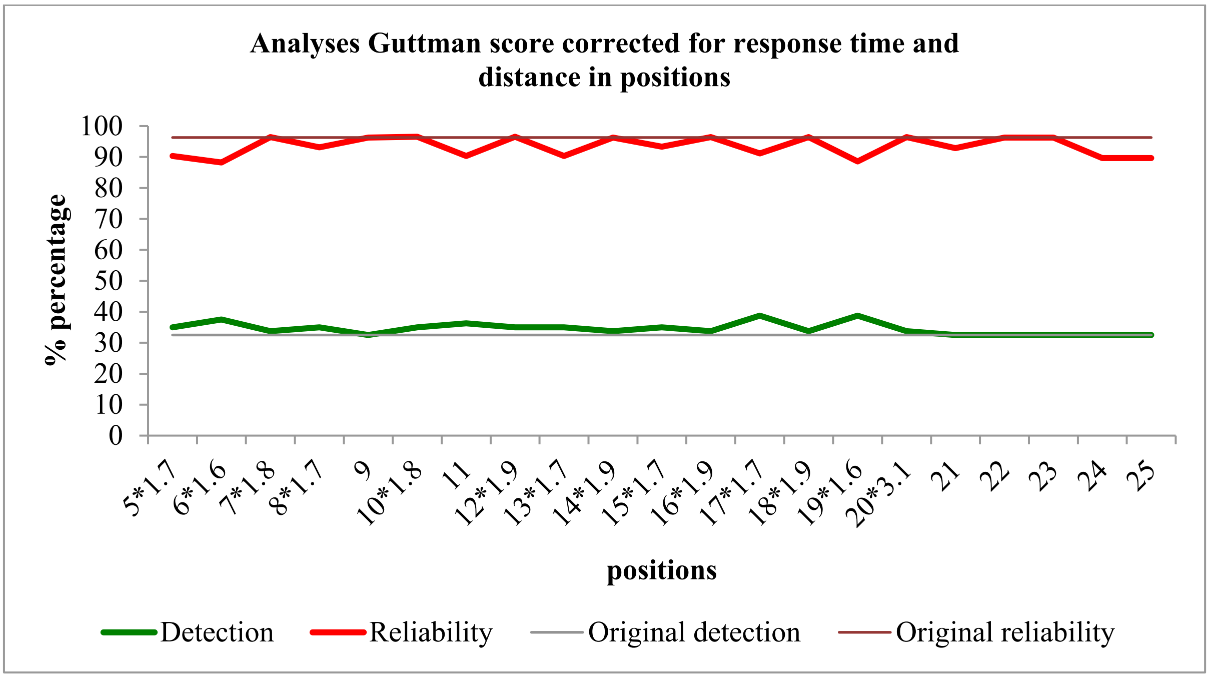

Figure 5. Graph displaying the results of analyses with the Guttman score corrected for response time and distance in positions index with 21 different settings and a cut-off score of 2 SD, compared to the results of the initial analyses determined previously excluding the initial index settings. Aberrant cut-off score reported after (*) for respecting position.

Evaluation of detection and reliability for different settings for the $G*_{\mathrm{rtd}}$ index proved more fruitfull. Figure 5 shows that the initial procedure (excluding the index in the current setting) could be improved to 35.0% detection with the $G*_{rtd12}$ index and a cut-off score of 1.9 *SD*, slightly improving reliability of the entire procedure to 96.6%. However, distance settings in *p*-value could not contribute more detection than it already contributed in the previously established set of analyses (see Figure 6) initially containing the $G*_{rtd0.20}$ index.

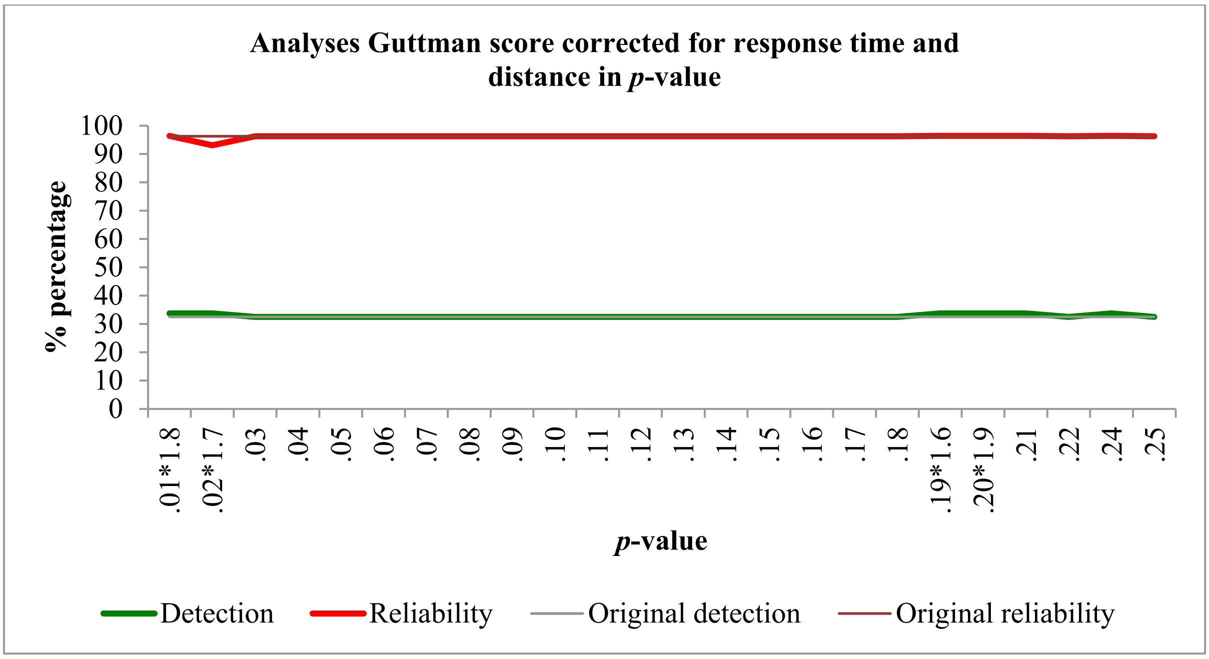

Figure 6. Graph displaying the results of analyses with the Guttman score corrected for response time and distance in *p*-value index with 24 different settings and a cut-off score of 2 SD, compared to the results of the initial analyses determined previously excluding the initial index settings. Aberrant cut-off score reported after (*) for respecting *p*-value.

Final analysis procedure. Within the current data forensics software two potential sets of analyses can be composed, one with the best possible reliability (≥ 95%; favoring this quality assessment parameter), and an alternative with the best possible detection (favoring this quality assessment parameter) with reasonable reliability (≥ 90%). The first consists of the previously established $G*_{rt}$ index (cut-off score 1.9 *SD*), combined with the $G*_{rtd12}$ (cut-off score 1.9 *SD*) and $G*_{rtd0.21}$ (cut-off score 2 *SD*) indices; the latter supersedes the 0.20 setting with identical results but a higher cut-off score. Closer inspection revealed this procedure could be further improved by adding analysis with the $G*_{rtd20}$ index (cut-off score 3.1 *SD*), for the cheaters detected with this index are different from those detected by the 12 positions setting. This final procedure provides 37.5% detection with 96.8% reliability.

The alternative set of analyses builds once again on the $G*_{rt}$ index (cut-off score 1.9 *SD*), combined with the previously established $G*_{rtd20}$ (cut-off score 3.1 *SD*) and $G*_{rtd0.21}$ (cut-off score 2 *SD*) indices. In addition, analyses are conducted with the $G*_{rtd17}$ index (cut-off score 1.7 *SD*), adding 6.3 percent points detection, but with only 71.4% reliability within the additional sample, and the $G*_{rtd11}$ index (cut-off score 2 *SD*), adding another 1.2 percent points detection, but with only 50.0% reliability within the additional sample. Also the previously abolished $G*$ index (cut-off score 1.9 *SD*) could provide an additional sample, improving detection with 1.3 percent points, but with 66.7% reliability within the additional sample. All together this alternative set of analyses provides the maximum of 43.8% detection with 92.1% reliability.

Explorative Phase

Table 4 presents the descriptive statistics on the indices in the established set of analyses for the groups within the experimental condition, the differentiation central to this research phase. A Kruskal-Wallis ANOVA indicated that there were no significant differences between these groups for analysis with the $G*_{rt}$ index, *H* (corrected for ties) = 8.030, *df* = 4, *N* = 80, *p* = .090, Cohen's *f* = .336. Such test also indicated no significant differences for analysis with the $G*_{rtd12}$ index, *H* (corrected for ties) = 8.453, *df* = 4, *N* = 80, *p* = .076, Cohen's *f* = .346. Nor for analysis with the $G*_{rtd0.21}$ index, *H* (corrected for ties) = 3.571, *df* = 4, *N* = 80, *p* = .467, Cohen's *f* = .218. Finally, another Kruskal-Wallis ANOVA indicated also no significant differences between the experimental groups for analysis with the $G*_{rtd20}$ index, *H* (corrected for ties) = 8.006, *df* = 4, *N* = 80, *p* = .091, Cohen's *f* = .336.

Overall, the participants indicated to the researcher that they found the test rather difficult. This is reflected by the fact that none of the students in the first four experimental groups passed the exam. Their frustration demotivated them. Some also expressed being uncomfortable having to cheat (*"I was very uncomfortable using my phone during the test (...)"*), while others used multiple methods to cheat (*"(...) I also consulted a peer real quick."* – smart phone group participant).

Table 4

Descriptive Statistics for the Deviations in Each Experimental Group on the Indices in the Established Procedure, including Mean Ranks for the Kruskal-Wallis ANOVA Tests

Index	Experimental group	*M*	*SD*	*Mean Rank*	*n*
$G*_{rt}$	Smart Phone	1.46	2.16	36.97	18
	Internal Collaboration	0.56	1.35	29.00	16
	Proctor Assistance	2.10	1.90	49.62	21
	Cheatsheet	1.58	2.28	39.75	8
	Pre-Knowledge	2.14	2.74	44.15	17
$G*_{rtd12}$	Smart Phone	1.58	2.17	39.44	18
	Internal Collaboration	0.51	1.36	29.13	16
	Proctor Assistance	2.19	2.08	50.93	21
	Cheatsheet	1.84	2.66	43.13	8
	Pre-Knowledge	1.85	3.34	38.21	17
$G*_{rtd0.21}$	Smart Phone	0.61	2.18	37.67	18
	Internal Collaboration	-0.24	1.00	32.69	16
	Proctor Assistance	0.57	1.82	43.43	21
	Cheatsheet	1.44	3.83	43.19	8
	Pre-Knowledge	1.24	2.79	45.97	17
$G*_{rtd20}$	Smart Phone	1.68	2.31	38.67	18
	Internal Collaboration	0.49	1.45	28.31	16

Proctor Assistance	2.08	2.17	47.60	21
Cheatsheet	-0.77	2.96	37.25	8
Pre-Knowledge	3.18	4.03	46.68	17

Smart phone. The detection within the group that used their smart phone to cheat is 33.3%. As presented in Table 5, Mann-Whitney *U* tests indicated that only the G^{*}_{rt} index and the G^{*}_{rtd12} index are valid analyses to detect this group. Wilcoxon Signed Rank pairwise comparison of the two valid analyses indicated that there was no significant difference between the deviation scores on the G^{*}_{rt} index over those on the G^{*}_{rtd12} index, $T = 46$, $z = -1.72$ (corrected for ties), N – Ties = 18, $p = .085$, $r = .41$, with 14 participants scoring higher deviations on the latter index (*Sum of Ranks* = 125).

Feedback from the questionnaire on cheating (see Appendix F for detailed answers) in this group revealed mainly that the participants managed to search several questions during the time they used their device. The main source for information was Google, and several participants indicated using the function of marking an item for later review in order to speed up the process of finding relevant information during the time of using their device. There seems to be no substantial difference between the methods of cheating in the detected sample over the undetected sample. One finding that stood out is the revelation by an undetected participant that he or she also consulted a peer, which went unnoticed by both proctors.

Internal collaboration. The detection within the group that collaborated with each other is 18.8%. Mann-Whitney *U* tests (see Table 5) indicated no valid analyses to detect this group, hence no further comparison of the indices could be made.

The questionnaires on cheating (see Appendix F) revealed that the participants consulted several peers, either asking their neighbours or calling to the entire group outloud. They indicated checking each others screens and questions. The most notable method was that of one participant finishing the test so the incorrectly answered items would appear, along with the correct answer. All methods were used in both the detected sample and the undetected sample of the group.

Proctor assistance. The detection within the group that received assistance from their proctor is 42.9%. The Mann-Whitney *U* tests (see Table 5) indicated that all established analyses were valid to detect this group. A Friedman two-way ANOVA indicated that rankings of the deviation scores varied significantly across the four analyses, $\chi^2(3) = 32.37$, N – Ties = 21, $p < .001$. Follow-up pairwise comparisons with Wilcoxon Signed Rank tests indicated strongly that the deviation scores on the $G^{*}_{rtd0.21}$ index (*Mean Rank* = 1.14) are significantly lower than those for the G^{*}_{rt} index (*Mean Rank* = 2.95), $T = 17$, $z = -3.42$ (corrected for ties), N – Ties = 21, $p = .001$, $r = .75$, as well as the G^{*}_{rtd12} index (*Mean Rank* = 3.19), $T = 18$, $z = -3.39$ (corrected for ties), N – Ties = 21, $p = .001$, $r = .74$, and the G^{*}_{rtd20} index (*Mean* Rank = 2.71) for this group, $T = 19$, $z = -3.36$ (corrected for ties), N – Ties = 21, $p = .001$, $r = .75$. The deviation scores on the G^{*}_{rt} index did not differ significantly from those for the G^{*}_{rtd12} index, $T = 91$, $z = -0.85$ (corrected for ties), N – Ties = 21, $p = .394$, $r = .19$, nor for the G^{*}_{rtd20} index, $T = 103$, $z = -0.44$ (corrected for ties), N – Ties = 21, $p = .664$, $r = .10$. Neither did the latter two, $T = 71$, $z = -1.55$ (corrected for ties), N – Ties = 21, $p = .121$, $r = .34$.

The participants in this group indicated in their questionnaires (see Appendix F) that they mainly consulted the proctor for concepts which were unknown to them, or sought answers through rephrasing the question. One detected participant also indicated consulting a peer during the test, which went unnoticed by three proctors. Apart from this, there was no notable difference between the methods of cheating in the detected and the undetected sample.

Cheatsheet. The detection within the group that used a personal cheatsheet is 25.0%. Mann-Whitney *U* tests (see Table 5) indicated that only analyses with the G^{*}_{rt} index and the G^{*}_{rtd12} index were valid to detect this group. Wilcoxon Signed Rank pairwise comparison of the two valid analyses indicated that there was no significant difference between the deviation scores on the G^{*}_{rt} index over those on the G^{*}_{rtd12} index, $T = 6$, $z = -1.68$ (corrected for ties), N – Ties = 8, $p = .092$, $r = .59$, with six participants scoring higher deviations on the latter index (*Sum of Ranks* = 30).

All participants in this group indicated answering questions using information in their cheatsheet that they did not know without it in their questionnaire on cheating (see Appendix F). One detected participant expressed "only" answering seven questions using the sheet; others did not give such indication. Several participants did imply that there were questions about content that had not been yet discussed in class, declaring their sheets could not help them on these items.

Pre-knowledge. The detection within the group that had pre-knowledge of the items on the test is 52.9%. Mann-Whitney *U* tests (see Table 5) indicated that all analyses were valid to detect this group. The Friedman two-way ANOVA indicated that rankings of the deviation scores varied significantly across the four analyses, $\chi^2(3) = 19.24$, *N* – Ties = 17, $p < .001$. Again, Wilcoxon Signed Rank pairwise comparisons strongly indicated that the deviation scores on the $G^*_{rtd0.21}$ index (*Mean Rank* = 1.35) are significantly lower than those for the G^*_{rt} index (*Mean Rank* = 3.18), $T = 27$, $z = -2.34$ (corrected for ties), *N* – Ties = 17, $p = .019$, $r = .57$, as well as the G^*_{rtd12} index (*Mean Rank* = 2.76), $T = 29$, $z = -2.25$ (corrected for ties), *N* – Ties = 17, $p = .024$, $r = .55$, and the G^*_{rtd20} index (*Mean* Rank = 2.71) for this group, $T = 6$, $z = -3.34$ (corrected for ties), *N* – Ties = 17, $p = .001$, $r = .81$. The deviation scores on the G^*_{rt} index did not differ significantly from those for the G^*_{rtd12} index, $T = 45$, $z = -1.49$ (corrected for ties), *N* – Ties = 17, $p = .136$, $r = .36$, nor for the G^*_{rtd20} index, $T = 75$, $z = -0.07$ (corrected for ties), *N* – Ties = 17, $p = .943$, $r = .02$. Neither did the latter two, $T = 73$, $z = -0.17$ (corrected for ties), *N* – Ties = 17, $p = .868$, $r = .04$.

In their interviews the participants in this group indicated that the first trial exam presented the most unknown items. This was reflected by the undetected sample including the first trials exams of three out of four participants. In the trial exams that followed, more and more items became familiar, which was again reflected partly by the data. The participants had handled the items on the test about 1.5 years back and had not reviewed them since. They expressed that knowing or recognizing an item does not imply knowing the answer also. Due to the informal office setting of the individual sessions, participants were sometimes interrupted or lost focus during the exams.

Table 5

Mann-Whitney U Statistics Determining the Validity of the Indices for Every Experimental Group

Experimental Group		Control condition				
Index	*Mean Rank*	*Mean Rank*	*U*	*z*	*p*	*r*
Smart Phone						
G^*_{rt}	36.19	24.01	185.50	-2.66	.008	.36
G^*_{rtd12}	36.25	23.99	184.50	-2.69	.007	.36
$G^*_{rtd0.21}$	30.75	26.66	283.50	-0.93	.353	.13
G^*_{rtd20}	33.25	25.45	238.50	-1.72	.085	.23
Internal Collaboration						
G^*_{rt}	31.31	25.14	227.00	-1.35	.178	.19
G^*_{rtd12}	30.75	25.38	236.00	-1.18	.239	.16
$G^*_{rtd0.21}$	27.59	26.74	286.50	-0.19	.849	.03
G^*_{rtd20}	26.56	27.19	289.00	-0.14	.890	.02
Proctor Assistance						
G^*_{rt}	41.98	22.42	126.50	-4.25	<.001	.56
G^*_{rtd12}	42.60	22.07	113.50	-4.49	<.001	.59
$G^*_{rtd0.21}$	35.88	25.88	254.50	-2.24	.025	.29
G^*_{rtd20}	39.26	23.96	183.50	-3.36	.001	.44
Cheatsheet						
G^*_{rt}	31.75	21.11	78.00	-2.09	.037	.31
G^*_{rtd12}	32.63	20.92	71.00	-2.31	.021	.34
$G^*_{rtd0.21}$	28.44	21.82	104.50	-1.34	.201	.20
G^*_{rtd20}	27.25	22.08	114.00	-1.03	.327	.15
Pre-Knowledge						
G^*_{rt}	38.03	22.66	135.50	-3.35	.001	.46
G^*_{rtd12}	33.68	24.66	209.50	-1.97	.049	.27
$G^*_{rtd0.21}$	34.68	24.20	192.50	-2.36	.018	.32
G^*_{rtd20}	35.21	23.96	183.50	-2.47	.013	.34

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

Discussion

The aim of this study was to assess and explore the data forensics analyses possible with the DFT software designed at eX:plain, researching the quality of the current analyses, finding the best quality composition of analyses within the currect tool, and exploring the possibilities of detecting specific methods of cheating. A control group of highly supervised examinees and an experimental group of instructed fraudulent examinees were evaluated by the software to see how well the currently conducted analyses could detect cheating. All initial analyses conducted with the six indices in the software were found valid, and the analyses added to 28.8% detection of cheating with 92.0% reliability. Cancelling indices, due to negative value or redundant positive value of analyses, and adjusting the settings of others, applying adequate interpretation of the analyses, led to two best quality compositions of data forensics analyses; one with the highest possible reliability (96.8%; detection: 37.5%), and one with the highest possible detection (43.8%; reliability: 92.1%). Favoring the first set, the following best set of data forensics analyses was established:

G^*_{rt} with cut-off score 1.9 *SD*;
G^*_{rtd12} with cut-off score 1.9 *SD*;
$G^*_{rtd0.21}$ with cut-off score 2.0 *SD*; and
G^*_{rtd20} with cut-off score 3.1 *SD*.

Although the analyses present different detection rates and reliability, their strength is in the combination of them. Overall, the G^*_{rt} index has the highest detection rate (30.0%; reliability: 96.0%), while the $G^*_{rtd0.21}$ index and the G^*_{rtd20} index have a smaller detection rates (18.8%; 25.0%), but with 100% reliability.

In review of the detection of different methods of cheating, it can be concluded that none of the indices prove exceptionally well fit to detect a specific method of cheating simulated in this research. Collaboration in a room during an exam could not be detected by any of the established data forensics analyses, as none of these were valid for this group. The G^*_{rt} and G^*_{rtd12} index analyses appeared fit to detect all other methods of cheating simulated in this research, namely the use of a smart phone, assistance of a proctor, using a paper cheatsheet, and having pre-knowledge of the test items. Although the $G^*_{rtd0.21}$ and G^*_{rtd20} indices can only be used to detect proctor assistance and pre-knowledge. It must be noted that the $G^*_{rtd0.21}$ seems less fit to do so; the scores on this index are on average lower than those of other indices, hence the detected sample will not be large.

With 37.5 to 43.8% detection and 96.8 to 92.1% reliability the DFT is certainly useful to detect cheating, as statements on the occurance seem rather reliable. It is hard to put those findings in perspective, as studies published in the field of data forensics are usually not based on real data. However, in a rather new publication, Mueller, Zhang & Ferrara (2017) reported about the evaluation of a real data set with known cheaters by several different data forensics programs. On average 13.9% of cheaters were detected, while reliability of the analyses is at a maximum of 99.2% for a 19.8% detection rate, down to 67.1% reliability linked to 50% detection. The reliability in this case is based on the proportion of non-cheaters truely labelled as such; in the current research this would remain 97.3% independent of the detection rate. Compared to the comparisons made by Mueller et al. (2017), the DFT software could be considered rather successful at detecting cheating.

An overall conclusion that can be drawn from the results is that the response times of examinees are a better behavioural indicator when combined with the Guttman score than the Guttman score itself. This is an essential finding for the field of data forensics, that has been focussing on response time modeling in the past. The explanation for the negative outcomes on the Guttman score most likely originates in the original theory of the model (Guttman, 1944). According to the basic literature, the model assumes unidimensionality, which is one of the main critiques that were voiced after Guttman's publication. The model was originally designed for measurement of attitudes, which is generally much better to scale than achievement (Sirotnik, 1987). To illustrate; it is common to use various questions to measure one attitude, it is much less common to measure only one construct in an achievement test. Take for instance an English test in high school; in reading a text quote students' vocabulary as well as their grammar is tested, in assessment of comprehension of the text an open item could be used, which additionally tests these vocabulary and grammar skills in writing. The question raised is what ability level is essentially assessed by this achievement test; vocabulary, grammar, reading, or writing? The conclusion is most likely the combination of these; the measurement is multidimensional.

Multidimensionality of the test assumes that there is not one ability level of the examinee that is to be measured. The examinee has a different ability level for each construct that is measured; the test result is simply a compromise. However, the Guttman model is based on the assumption of unidimensionality, and the Guttman score therefore increases proportionally due to the multidimensionality of the achievement test used. Even if a candidate shows ideal Guttman sequences for each dimension, the dimensions combined show many Guttman errors due to unavoidable differences in ability level in each dimension. This problem, which is reality for most achievement measurements, was likely partly overcome by the measurement of deviation scores instead of the true Guttman scores. However, this may not have produced as realistic measurements of behaviour patterns as the designers of the software would have hoped, perhaps because each individual differs on the ability levels in their own way.

Although the problem of multidimensionality seems hard to overcome since it is so inherent to complex achievement measurement, there are hopeful alternatives. The opportunities lie in the detailed professional design of the exams administered by test agencies such as eX:plain. The multidimensionality is undoubtedly partly fixed in the test matrix that is at the foundation of the exam, designed by content and test experts, evaluated by factor analysis, and used by QMP software to randomize the exams. Computing the Guttman error, score and other indices based on the sum of the seperate sequences per construct in the test matrix, rather than the combined one, might provide new opportunities for cheating detection in randomized tests. This enforces the Guttman model as one of the best options for data forensics in randomized tests, as most other models (mainly similarity analysis) are still only fit for fixed tests. It is recommended to program the DFT software so that is computes the Guttman score from multiple Guttman sequences, one per dimension in the test matrix.

With the response time measurement as the best behavioural measurement for cheating detection, and yet undetermined differences between the detected and the undetected samples found, the response times measurements were further explored in a small-scale additional research project reported in Appendix G. Hereby it can be concluded that QMP does not properly log the response time of examinees. The response time recorded is the time from first opening the item to closing it after a first answer is selected, and should therefore better be referred to as display time. Reopening the item and/or changing the answer was not recorded, nor were behaviours such as marking an item of any influence on the logging on the display times.

These conclusions help find an hypothesis for why some cheaters were detected, while others were not. The questionnaire distributed amongst the instructed fraudulent examinees did not provide statements on this subject. It is hypothesized, based on the disclosed logging of display times by QMP, that the undetected cheaters might have dodged the data forensics software through the loopholes that the log creates. A cheater might not have been detected simply because he or she reopened an item that was already answered and possibly showed it to a peer, or reanswered it using information found on their smart phone. This additional display time, and thus the cheating, was is those cases not logged in the display times provided by QMP, and therefore not detected by the DFT software.

This theory provides additional hypotheses on why some of the explored methods of cheating were better detected than others. Proctor assistance was a method that was rather well detected, compared to for instance internal collaboration and the use of a smart phone. It can be hypothesized that the differences between the groups are partly caused by the incentive to submit at least a momentary response for a test item on the first try, possibly influenced by the confidence the participants had in their method of cheating. The participants in the smart phone group and the internal collaboration group could have been uncertain whether they would actually be able to profit from cheating; would they be able to find their answer online within the time limit? Or would their peers even be able to help them? This uncertainty may have moved them to submit a first temporary answer, more so than the participants in the proctor assisstance group, who were much more certain they would get useful help. The additional display time and change of answer by the smart phone and internal collaboration groups during or after the moment of cheating, was not logged by QMP in case a temporary response was already submitted.

However, it must be emphasized that the conclusions of the main study are based on the disclosed loggings of QMP; response times used in the data forensics analyses are logged the same for all participants in this study. As QMP has been logging the display time this way all along, conclusions about past analyses still stand. Even with the limitations of the information used by the data forensics software, the established analyses still indicated to be able to detect 37.5% of cheaters in the sample with 96.8% reliability. In order to improve the analyses however, it is strongly recommended that QMP starts logging the actual total display time of the items, as stated (Questionmark, 2018). Confidence in one's cheating method should never be a factor in whether or not one is caught. The loopholes in the current display time loggings might even provide cheaters a method of eluding the system. For now the practices were probably merely accidental, however if the information were to be seized by professional cheaters they could easily develop a method to dodge the system. It is essential that the time the cheating occurs is logged by QMP. Therefore it is not recommended to start logging true response times (time until selecting the answer), as cheaters would again be able to develop methods to elude the system. It would be simple to just select the answer at an appropriate time (logged as response time) and consult a peer, proctor, smart phone or cheatsheet after, while the item is still displayed. Even going through previous items to show them to others would be left undetected.

The proposed amplification of the Guttman score computations and the improvement through the logging of the true display times by QMP are expected to improve the detection of the DFT. For future development and research, it is valuable to explore even more possiblities for improvement of the software. Considering the behaviour measurements available; ability level and display time, with the latter, based on current results, expected to be most useful, it would be appropriate to implement the original response time index and the seperate lognormal model by Van der Linden (2006). In the initial design of the indices of the software, the Guttman score corrected for response time and the computation of the lognormal model by Van der Linden were seperately stored; they were combined by the programmer of the software. The Guttman score corrected for response time index was initially computed similar to the corrected for distance indices; items in the sequence were supposed to be selected based on their deviation from the mean response time on each item; for instance those with a deviation higher than 3 *SD*, a setting in the software to be adjusted manually by data analists. A seperate computation of the lognormal model could reveal the exact response time patterns that possibly distinguish (methods of) cheating in detail. This can only be based, of course, on proper display time logging by QMP.

Besides measurement of item display times and ability of candidates, other behavioural measurements are relevant to develop data forensics indices. It is relevant to seek opportunities to record item selection behaviour of candidates: whether or not they changed their answer and whether or not they profited from this (Van der Linden & Jeon, 2012). Erasure tracking is a frequently exercised measure to detect cheating in selection behaviour. It was initially used for paper-based tests, but can be extended to computer-based tests, if the behaviour is logged. The model Van der Linden & Jeon (2012) present is based on the probability that a candidate will perform a wrong-to-right answer change. The model would be exceptionally fit to detect internal collaboration, and, in the current study, the occurance of a moment of smart phone use. Furthermore, Van der Linden & Jeon (2012) mention potential improvement by using computer registered response times, the description of which is similar to the display times argued in the current paper. However, the model has only been researched in simulations; it will have to be reviewed with realistic data, such as collected for the current study.

Apart from information that could possibly be provided by QMP logging, in research other measures of behaviour are being explored for cheating detection as well. An interesting development is the measurement of head pose when cheating. In using cheatsheets, the head poses of cheaters significantly differ from non-cheaters (Chuang, Craig, & Femiani, 2017). Head pose measurements would however require the installation of the appropriate materials, software, cameras, and the permission of candidates to be filmed. In practice, it is usually a long process to come to such measures in exam centres. However, as distance proctoring becomes more dominant in the field of assessment, it is wise to eye such developments in research and take steps to 'keep up' and, for instance, start preparing functional software.

With the improvement, expansion and development of the software, and the growing importance and awareness of cheating and cheating detection in assessment, it is important that the DFT becomes easier to use. Automation of the data forensics analyses or even real-time cheating detection should be central to further development of the software. This study provides essential information to start the automation: it is now established what indices, with which settings, should be run. The software should be programmed to compute means and standard deviations itself and flag all those candidates with a deviation score higher or equal to the established cut-off score as cheaters. The software should produce a clear overview of what candidates could be identified as cheaters with almost 97% certainty.

Future automation of the analyses by the software could help minimize a minor limitation that the benchmarking in the current study presented. In the current study, it is unknown how many cheaters were included in the frame of reference. These cheaters will have influenced the mean and standard deviation for each analysis. A fairly large sample was used, to minimize the problem. However, the software could be able to compute better means and standard deviations, by using the two-year data set it already uses to establish the *p*-values and mean response times for items. This large sample should be adapted to exclude as many cheaters as possible, using the established analyses. This will provide even more accurate benchmarking.

To establish the true reliability of all statements made in this paper, the study should be replicated. Only retesting can determine the true quality of the software, the detection rates and reliability in the current paper can be used as an estimation until then. It is especially important to consider the non lineair relationship the settings of the different indices displayed in the Figures in the results section. It is uncertain whether a retest will reveal the same patterns. Other limitations of this research should also be addressed.

Sample sizes presented one of the greatest limitations to this study. The experimental group using cheatsheets is so small no strong conclusions can be drawn from the results. Also, the control group was considerably smaller than the experimental group in total. This was due to the low season in regular test taking at the location the control group was recruited. The initial goal for the study was to recruit 100 participants for each condition (20 per experimental group), but this goal was also not met for the experimental condition, due to problems motivating the students to follow the instructions on cheating close enough. In replicating this study it is strongly advised to at least surpass the pre-determined goals, to get the previously addressed better estimation of the true reliability of the analyses. It will therefore possibly require a larger timeframe.

The realistic setting of the cheating attempts, although a distinctive feature for research in the field of data forensics, offered its own limitations and context. The participants in the first four experimental conditions expressed that they were demotivated by their lack of ability to complete the test on their own. This is underlined by the fact that none of these students passed their trial exam.The pre-knowledge condition presented various problems; the group was not similar in demographics to the other experimental groups, and statements on cheating could not be linked directly to a trial exam (because there we several per participant).

In the process of reviewing the questionnaires completed in the experimental condition of this study, an important issue was raised that will also need consideration in future research; whether or not cheating should be successful to be detected. Only a few participants indicated whether their method of cheating was actually helpful; whether they found any useful information on their phone, or them consulting a peer actually helped answer an item correctly. For this research it was determined that the succes of cheating could not be a distinctive feature, since there was no evidence provided by the participants on the subject. However, it would be useful to include an additional question to the questionnaire to assess this feature in the future. It is useful to know whether the software detects successful and non-successful cheaters, or just the successful ones. This in turn would provide an additional explanation for the detection rate in the currect study.

The pre-knowledge group presented a method of cheating that is very hard to simulate in a realistic setting; it is highly undesirable to have regular students gain pre-knowledge of the test items. The current research focussed solely on methods of cheating, test fraud as a whole includes methods of piracy that are rather undesirable to simulate realistically also. However, to reach future conclusions on the potential to detect test fraud, including cheating *and* piracy, it is important to consider more methods. Piracy involves the theft of test items or entire item banks, in order to publish or sell them, or to use them to create more succesful educational programs, directed at teaching the test (Foster, 2013). The practice is extremely harmful to testing agencies as their intellectual properties are highly valuable. Hacking item banks or filming one's screen while completing an exam are rather extreme but not uncommon methods of piracy that will most likely not be detectable with data forensics (Foster, 2013; Impara & Foster, 2011). However, the practice of harvesting; trying to remember as many items as possible and potentially purposely failing the test to get a retake and remember more, is just as harmful, but could possibly be detected as aberrant behaviour. It is strongly recommended to research the possible detection of this piracy method in the future.

The detection of recurring cases of cheating should provide some tools in the detection of incidents of piracy already. Although the detection rate found for individuals is 37.5%, for grouping variables the chances of being investigated for possible fraud increase with every cheating candidate. For instance, if an educator uses stolen test items to teach just 10 candidates, 3 to 4 of these candidates with pre-knowledge would be detected by the DFT. The similarity on the grouping variable (educator) between these detected candidates could induce an investigation to what caused the irregularities detected. It is very likely an educator will teach more candidates; they teach for a profit, and with that it is more and more likely the practice will get detected. Other grouping variables should be considered as well; exam location, testing time, and proctor.

To finalize, it should be mentioned that several incidents in the realistic settings created awareness in the research team for the pressing problem of cheating. While some participants in the experimental conditions indicated that they felt very uncomfortable cheating, others went creative and applied several methods of cheating at once. In one situation, three proctors were monitoring approximately 10 participants in a rather small room. There was no talking or noises aloud (the participants were using cheatsheets). In the questionnaire, reviewed afterwards, a participant, sadly without a cheatsheet, stated to have consulted a peer at the next computer for answers, in an effort to cheat anyway. None of the proctors had noticed any disturbance in the quiet room. Furthermore, some participants indicated in their questionnaire that they had searched for exam items on the internet in preparation of the exam, even though they had never been instructed to. In addition, in every session, several participants leaned backwards to look at screens of others, a rather useless way of trying to copy answers due to the randomization of the items. However, the nonchalance in this behaviour is what struck; it was as though this was a completely normal thing to be doing. Although it is very difficult to research the prevalence of cheating, it seems to be more "normal" than one might want to suspect. The situations painted by the high prevalency numbers mentioned in the introduction of this paper might not be excessive.

The existential prevalence of cheating calls for advancements and existential research in the relatively small and young field of data forensics for randomized computer based multiple-choice testing. The findings in this study are of great value to data forensics measures for random exams; focusing on display time as a valuable indicator of cheating, and extention of the Guttman model for achievement testing. It is very important to catch up to continuing developments in the overarching field of assessment, as well as that of information technology. Randomization of test items is a strong measure to prevent cheating by copying, but computer-based tests provide many more opportunities to measure behaviour to detect and prevent more methods of cheating and piracy. Technology provides great benefits to test security, and it is time to exploit these to the maximum. If not, cheaters will always be one or two steps ahead, threatening the validity and value of expensive exams, items banks, and credentials. It is time to catch up.

Recommendations for Client

For eX:plain and Xquiry going forward it is recommended to improve the data forensics software in the following ways:

- Have QMP log the true display time for examinees, complying to their actual claim.
- Have QMP log more behaviour than just the display time. It would be very valuable to know whether an examinee changed an answer; opening up options of adding erasure tracking indices, a method that has been prominent in the field of data forensics for decades. Information on the marking of items could be valuable information in determining the perceived difficulty of items. This not only provides options for assembly of more accurate Guttman sequences, but may also hold information that could be useful for item development and maintainance.
- Add more indices to the DFT, focussing on display times and other possibilities available through the new information provided by QMP. Although the current indices are of different quality between themselves, the best results are achieved combining them. Recommended possibilities are:
    - Programming the original display time index, where items in the sequence are selected based on a given number of deviations on the mean display time that is expected for the item. This was the original intention of the designers of the software, however, in programming, the log normal model by Van der Linden (2006) was mixed up, and the current index, although now proven very usefull, was created accidentally. Since the display time has proved to be a valuable behavioural indicator, it is recommended to implement the original theory also.
    - Program the lognormal model by Van der Linden (2006) seperately, in order to evaluate response patterns of cheaters more accurately.
    - Implementing indices for erasure tracking analysis with the model by Van der Linden & Jeon (2012).
    - Exploring options for analysis of visual data.
- Have the software calculate deviation scores for the analyses itself and over larger periods of time, systematically excluding those candidates that are identified as cheaters. This will facilitate the possibility of automated data forensics analyses by Xquiry analists, where the software detects cheaters itself and simply notifies the analist.
- Explore the possibilities of having the software construct multiple Guttman sequences, based on the existing test matrices programmed in QMP. The software would calculate Guttman errors over each seperate sequence and combine those in the Guttman score. It is important to note that the current construction of the sequence should be maintained in the mean time, as they are at the base of the current findings.

Along with the improvements in the software, the following is important in going forward with the test security services using the DFT:

- In future use of the DFT, conduct the established analyses with assigned interpretations.
- Replicate the current study with a larger sample to get a better estimation of the reliability of the analyses. Consider the current findings more as an indication than the true number. Ideally both the experimental and the control group should be over 100 participants. Furthermore the following could possibly be taken into account:
    - How to motivate the participants in the experimental condition better to comply with instructions.
    - Add a question to the questionnaire about the success of the cheating method.
    - Have more realistic settings for the pre-knowledge condition, as well as possibly creating a harvesting condition, as this is a serious threat to the intellectual properties of exam agencies as well.
    - Evaluate the previously set out improvements to the software, including the alternative set of existing analyses that was found in this research.
- On the subject of what the findings of the software should be used for, the recommendation is to have the entire detected sample of examinees retake the concerning exam. It can never be 100% proven who did or did not cheat, but it can be concluded that highly irregular behaviour

was detected, with very high potential of fraud, and the security and thereby the validity of the results of the exam and certification cannot be ensured. All examinees should be notified of this procedure in advance. Offering retakes is a procedure already adapted by several testing agencies and favored by researchers in the field (Van der Linden & Jeon, 2012).

References

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American Statistical Association, 69*, 44-49. doi:10.1002/j.2333-8504.1972.tb00449.x (1972 admission)

Association of Test Publishers (2002). *Guidelines for computer-based testing.* Washington, D.C.: ATP.

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. doi:10.1207/s15328023top1603_15

Belov, D. I. (2015). Comparing the performance of eight item preknowledge detection statistics. *Applied Psychological Measurement, 40*(2), 1-15. doi:10.1177/0146621615603327

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 (2018, January 3)

Bird, C. (1927). The detection of cheating in objective examinations. *School and Society, 25*(635), 261-262.

Bird, C. (1929). An improved method of detecting cheating in objective examinations. *The Journal of Educational Research, 19*(5), 341-348.

Chuang, C. Y., Craig, S. D., & Femiani, J. (2017). Detecting probable cheating during online assessments based on time delay and head pose. *Higher Education Research & Development, 36*(6), 1123-1137. doi:10.1080/07294360.2017.1303456

Cizek, G. (1999). *Cheating on tests: How to do it, detect it and prevent it.* Mahwah, NJ: Lawrence Erlbaum.

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.

Crawford, C. (1930). Dishonesty in objective tests. *The School Review, 38*(10), 776-781.

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.12151

Field, A. (2013). *Discovering statistics using SPSS.* Los Angeles: Sage publications.

Foster, D. (2013). Security issues in technology-based testing. In A. Wollack & J. J. Fremer (Eds.), *Handbook of Test Security* (pp. 39-84). New York, NY: Routledge.

Frary, R. B., Tideman, T. N., & Watts, T. M. (1977). Indices of cheating on multiple-choice tests. *Journal of Educational and Behavioural Statistics, 2*(4), 235-256. doi:10.3102/10769986002004235

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*(2), 139-150. Retrieved from: http://www.jstor.org/stable/2086306 (2018, April 17)

Holland, P. W. (1996). *Assessing unusual agreement between the incorrect answers of two examinees using the K-index: Statistical theory and empirical support* (Technical Report No. 96-4. Princeton, NJ: Educational Testing Service.

Hughes, J. M. C., & McCabe, D. L (2006). Academic misconduct within higher education in Canada. *The Canadian Journal of Higher Education*, 36, 1-21. Retrieved from: https://files.eric.ed.gov/fulltext/EJ771043.pdf (2018, January 8)

Impara, J. C., & Foster, D. (2011). Item and test development strategies to minimize test fraud. In T. M. Haladyna & S. M. Downing (Eds.), *Handbook of test development* (pp. 91-114). Mahwah, NJ: Lawrence Erlbaum.

Impara, J. C., Kingsbury, G., Maynes, D., & Fitzgerald, C. (2005). *Detecting cheating in computer adaptive tests using data forensics.* Paper presented at the Annual Meeting of the National Council on Measurement in Education and the National Association of Test Directors, Montreal, Canada.

Kingston, N. M., & Clark A. K. (2014). Introduction. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 1-3). New York, NY: Routledge.

Lin, C.-H. S., & Wen, L.-Y. M. (2007). Academic dishonesty in higher education: a nationwide study in Taiwan. *Higher Education*, 54, 85-97. doi: 10.1007/s10734-006-9047-z

Marianti, S., Fox, J. P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behaviour in response time modeling. *Journal of Educational and Behavioural Statistics, 39*(6), 426-451. doi:10.3102/1076998614559412

Maynes, D. D. (2017). Detecting potential collusion among individual examinees using similarity analysis. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 47-69). New York, NY: Routledge.

Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement, 18*(4), 311-314. doi:10.1177/014662169401800402

Miller, A. D., Murdock, T. B., Anderman, E. M., & Poindexter, A. L. (2007). Who are all these cheaters? Characteristics of academically dishonest students. In E. M. Anderman & T. B. Murdock (Eds.), *Psychology of academic cheating* (pp. 9-32). Cambridge, MA: Academic Press.

Mroch, A. A., Lu, Y., Huang, C., & Harris, D. J. (2014). In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 137-148). New York, NY: Routledge.

Mueller, L., Zhang, Y., & Ferrara, S. (2017) What have we learned? In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 373-389). New York, NY: Routledge.

Questionmark. What time is recorded when a participant answers a question? Retrieved from: https://www.questionmark.com/content/what-time-recorded-results-when-participant-answers-question (2018, 17 april)

Simon, M. (2014). Local outlier detection in data forensics: Data mining approach to flag unusual schools. In N. M. Kingston & A. K. Clark (Eds.), *Test fraud: Statistical detection and methodology* (pp. 83-100). New York, NY: Routledge.

Sirotnik, K. A. (1987). Toward more sensible achievement measurement: A retrospective. In D. L. McArthur (Ed.), *Alternative approaches to the assessment of achievement* (pp. 21-78). Norwell, MA: Kluwer Academic Publishers.

Van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioural Statistics*, *31*(2), 181-204. doi:10.3102/10769986031002181

Van der Linden, W. J. & Jeon, M. (2012). Modeling answer changes on test items. *Journal of Educational and Behavioral Statistics, 37*(1), 180-199. doi:10.3102/1076998610396899

Wollack, J. A. (1997). A nominal response model a[[roach for detecting answer copying. *Applied Psychological Measurement, 21*(4), 307-320. doi:10.1177/01466216970214002

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.

Zopluoglu, C. (2017). Similarity, answer copying, and aberrance: Understanding the status quo. In G. J. Cizek & J. A. Wollack (Eds.), *Handbook of quantitative methods for detecting cheating on tests* (pp. 25-46). New York, NY: Routledge.

Appendix A
Informed Consent Form

Informed consent – Toestemmingsverklaringsformulier

Detecting cheating in multiple-choice testing
Researcher: Sanette van Noord

In te vullen door deelnemer:

Ik verklaar op een (voor mij) duidelijke wijze te zijn ingelicht over het doel, de methode, en [indien aanwezig] de risico's en belasting van het onderzoek. Ik geef hierbij toestemming aan de voor het onderzoek verantwoordelijke onderzoeker van de Universiteit Twente om de gegevens die zijn verkregen te gebruiken voor het onderzoek. Ik weet dat de gegevens en resultaten van het onderzoek alleen anoniem en vertrouwelijk zullen worden behandeld. Ik ben in de gelegenheid gesteld om vragen te stellen en mijn (eventuele) vragen zijn naar tevredenheid beantwoord.

Ik weet dat meedoen geheel vrijwillig is en dat ik op ieder moment kan beslissen om af te zien van deelname zonder een reden te hoeven geven.

Naam deelnemer: …………………………………………………………

Geslacht: M / V

Leeftijd: ………jaar

Datum: ………………… Handtekening:

In te vullen door onderzoeker:

Ik heb een schriftelijke/mondelinge toelichting gegeven op het onderzoek. Ik zal resterende vragen over het onderzoek naar vermogen beantwoorden. De deelnemer zal van een eventuele voortijdige beëindiging van deelname aan dit onderzoek geen nadelige gevolgen ondervinden.

Naam onderzoeker: …………………………………………………………

Datum: ………………… Handtekening:

Appendix B
Questionnaire for Participants in the Experimental Condition

Naam:…………………………………………

Op welke manier heb je (individueel of samen met de groep) tijdens het examen informatie of antwoorden verkregen?
Vink jouw manier aan hieronder.

○ Door te overleggen met medestudenten
○ Door de surveillant om informatie te vragen
○ Door een spiekbriefje te gebruiken
○ Door mijn telefoon te gebruiken
○ Doordat ik vragen van tevoren kende

Beschrijf, zo gedetailleerd mogelijk, hoe je dit hebt gedaan. Beschrijf bijvoorbeeld met wie je hebt overlegd en waarover, welke vraag je de surveillant hebt gesteld en welke informatie die gaf, wanneer je spiekbriefje van pas kwam, wat je hebt opgezocht op je telefoon, of welke vragen je je kon herinneren.

Schrijf eventueel verder op de achterkant van dit blaadje

Appendix C
Kolmogorov-Smirnov Goodness-of-Fit Tests for All Indices and Settings

		Not cheating		Cheating	
Index	Setting	*D*(36)	*p*	D(79)	*p*
G^*	-	.11	>.200	.09	>.200
$G^*_{dpositions}$	5	.11	>.200	.09	.161
	6	.09	>.200	.10	**.044**
	7	.12	.162	.11	**.018**
	8	.12	>.200	.11	**.031**
	9	.11	>.200	.10	**<.001**
	10	.13	.156	.11	**.014**
	11	.13	.129	.12	**.007**
	12	.13	.107	.12	**.008**
	13	.13	.107	.13	**.003**
	14	.15	**.042**	.12	**.008**
	15	.15	**.026**	.12	**.010**
	16	.13	.111	.12	**.006**
	17	.15	**.045**	.13	**.002**
	18	.14	.086	.12	**.005**
	19	.15	**.036**	.14	**.001**
	20	.17	**.007**	.20	**<.001**
	21	.19	**.002**	.18	**<.001**
	22	.19	**.002**	.18	**<.001**
	23	.21	**<.001**	.20	**<.001**
	24	.17	**.011**	.22	**<.001**
	25	.17	**.007**	.21	**<.001**
$G^*_{dp\text{-}value}$	.01	.12	.199	.08	>.200
	.02	.13	.111	.09	.088
	.03	.10	>.200	.13	**.003**
	.04	.12	>.200	.11	**.022**
	.05	.10	>.200	.12	**.011**
	.06	.11	>.200	.12	**.004**
	.07	.13	.127	.13	**.002**
	.08	.11	>.200	.14	**.001**
	.09	.13	.147	.12	**.004**
	.10	.13	.144	.13	**.001**
	.11	.12	>.200	.12	**.006**
	.12	.13	.138	.13	**.003**
	.13	.15	**.031**	.13	**.002**
	.14	.17	**.009**	.13	**.002**
	.15	.15	**.034**	.15	**<.001**
	.16	.14	.074	.16	**<.001**
	.17	.15	**.044**	.18	**<.001**
	.18	.18	**.003**	.15	**<.001**
	.19	.14	.062	.16	**<.001**
	.20	.21	**<.001**	.16	**<.001**
G^*_{rt}	-	.23	**<.001**	.21	**<.001**
$G^*_{rtdpositions}$	5	.21	**<.001**	.21	**<.001**
	6	.21	**<.001**	.21	**<.001**
	7	.18	**.005**	.23	**<.001**
	8	.18	**.004**	.22	**<.001**
	9	.19	**.002**	.22	**<.001**
	10	.17	**.009**	.21	**<.001**

	11	.17	**.009**	.23	**<.001**
	12	.16	**.014**	.22	**<.001**
	13	.19	**.002**	.25	**<.001**
	14	.15	**.030**	.23	**<.001**
	15	.17	**.011**	.26	**<.001**
	16	.17	**.012**	.24	**<.001**
	17	.20	**.001**	.23	**<.001**
	18	.20	**.001**	.25	**<.001**
	19	.16	**.016**	.25	**<.001**
	20	.19	**.002**	.27	**<.001**
	21	.21	**<.001**	.26	**<.001**
	22	.23	**<.001**	.27	**<.001**
	23	.27	**<.001**	.28	**<.001**
	24	.25	**<.001**	.27	**<.001**
	25	.21	**<.001**	.25	**<.001**
$G^*_{rtdp\text{-}value}$	.01	.20	**<.001**	.19	**<.001**
	.02	.18	**.003**	.21	**<.001**
	.03	.18	**.003**	.20	**<.001**
	.04	.22	**<.001**	.21	**<.001**
	.05	.19	**.001**	.22	**<.001**
	.06	.24	**<.001**	.20	**<.001**
	.07	.20	**.001**	.22	**<.001**
	.08	.18	**.003**	.23	**<.001**
	.09	.16	**.016**	.22	**<.001**
	.10	.17	**.011**	.21	**<.001**
	.11	.16	**.015**	.22	**<.001**
	.12	.17	**.007**	.23	**<.001**
	.13	.23	**<.001**	.24	**<.001**
	.14	.22	**<.001**	.23	**<.001**
	.15	.18	**.005**	.24	**<.001**
	.16	.22	**<.001**	.26	**<.001**
	.17	.20	**.001**	.24	**<.001**
	.18	.23	**<.001**	.26	**<.001**
	.19	.24	**<.001**	.28	**<.001**
	.20	.28	**<.001**	.26	**<.001**
	.21	.26	**<.001**	.28	**<.001**
	.22	.27	**<.001**	.27	**<.001**
	.23	.33	**<.001**	.28	**<.001**
	.24	.25	**<.001**	.32	**<.001**
	.25	.28	**<.001**	.32	**<.001**

Note. Due to values in bold, Mann-Whitney *U* analyses were conducted instead of the intended independent *t*-tests. Results also reported in body.

Appendix D
Shapiro-Wilk Tests for Established Indices in Experimental Groups

Index	Experimental group	*W*	*df*	*p*
$G*_{rt}$	Smart Phone	.85	18	.009
	Internal Collaboration	.81	16	.004
	Proctor Assistance	.84	21	.003
	Cheatsheet	.81	8	.039
	Pre-Knowledge	.84	17	.009
$G*_{rtd12}$	Smart Phone	.80	18	.002
	Internal Collaboration	.84	16	.011
	Proctor Assistance	.83	21	.002
	Cheatsheet	.75	8	.009
	Pre-Knowledge	.81	17	.003
$G*_{rtd0.21}$	Smart Phone	.75	18	<.001
	Internal Collaboration	.90	16	.084
	Proctor Assistance	.76	21	<.001
	Cheatsheet	.70	8	.002
	Pre-Knowledge	.75	17	<.001
$G*_{rtd20}$	Smart Phone	.83	18	.004
	Internal Collaboration	.77	16	.001
	Proctor Assistance	.81	21	.001
	Cheatsheet	.73	8	.005
	Pre-Knowledge	.85	17	.011

Appendix E

Descriptives and Results for Independent *T*- and Mann Whitney *U*-Tests for All Indices and Settings

		Not cheating			Cheating						
		(*N* = 37)			(*N* = 80)						
Index	Setting	*M*	*SD*		*M*	*SD*		*t*(115)		*p*	Cohen's *d*
$G^*_{dpositions}$	5	0.07	1.21		0.89	1.12		-3.58		.001	.76
$G^*_{dp\text{-}value}$	.01	0.00	1.17		0.80	1.09		-3.61		<.001	.76
	.02	-0.01	1.16		0.74	1.09		-3.39		.001	.71
				Mean			*Mean*				
				Rank			*Rank*	*U*	*z*		*r*
$G^*_{dpositions}$	6	0.08	1.22	42.65	0.89	1.16	66.56	875.00	-3.55	<.001	.33
	7	0.08	1.21	42.00	0.90	1.16	66.86	851.00	-3.69	<.001	.34
	8	0.10	1.19	42.39	0.89	1.13	66.68	865.50	-3.61	<.001	.33
	9	0.10	1.20	42.47	0.90	1.14	66.64	868.50	-3.59	<.001	.33
	10	0.08	1.20	42.31	0.89	1.18	66.72	862.50	-3.62	<.001	.33
	11	0.10	1.20	42.64	0.88	1.19	66.57	874.50	-3.55	<.001	.33
	12	0.10	1.19	42.78	0.89	1.21	66.50	880.00	-3.52	<.001	.33
	13	0.09	1.20	42.64	0.90	1.23	66.57	874.50	-3.56	<.001	.33
	14	0.10	1.22	43.01	0.90	1.25	66.39	888.50	-3.48	.001	.32
	15	0.08	1.23	42.58	0.90	1.25	66.59	872.50	-3.57	<.001	.33
	16	0.09	1.24	42.65	0.91	1.27	66.56	875.00	-3.55	<.001	.33
	17	0.08	1.20	42.55	0.90	1.26	66.61	871.50	-3.58	<.001	.33
	18	0.11	1.21	43.72	0.89	1.30	66.07	914.50	-3.32	.001	.31
	19	0.12	1.20	43.59	0.89	1.27	66.13	910.00	-3.35	.001	.31
	20	0.52	1.41	49.07	0.99	1.51	63.59	1112.50	-2.16	.031	.20
	21	0.39	1.23	46.69	1.00	1.45	64.69	1024.50	-2.68	.007	.25
	22	0.57	1.31	46.69	1.23	1.55	64.69	1024.50	-2.68	.007	.25
	23	0.40	1.31	48.16	0.89	1.54	64.01	1079.00	-2.36	.018	.22
	24	0.43	1.36	48.51	1.00	1.57	63.85	1092.00	-2.29	.022	.21
	25	0.43	1.41	50.59	0.93	1.65	62.89	1169.00	-1.84	**.066**	.17
$G^*_{dp\text{-}value}$	.03	-0.04	1.16	44.51	0.67	1.11	65.70	944.00	-3.15	.002	.29
	.04	-0.09	1.14	43.84	0.63	1.09	66.01	919.00	-3.29	.001	.30
	.05	-0.12	1.13	44.61	0.58	1.11	65.66	947.50	-3.13	.002	.29

	.06	-0.17	1.09	44.39	0.53	1.13	65.76	939.50	-3.17	.002	.29
	.07	-0.19	1.09	45.18	0.46	1.14	65.39	968.50	-3.00	.003	.28
	.08	-0.19	1.08	46.12	0.41	1.12	64.96	1003.50	-2.80	.005	.26
	.09	-0.23	1.06	46.32	0.38	1.13	64.86	1011.00	-2.75	.006	.25
	.10	-0.27	1.06	46.18	0.34	1.15	64.93	1005.50	-2.79	.005	.26
	.11	-0.30	1.07	46.54	0.29	1.13	64.76	1019.00	-2.71	.007	.25
	.12	-0.31	1.03	46.41	0.27	1.13	64.83	1014.00	-2.74	.006	.25
	.13	-0.38	1.02	45.47	0.22	1.10	65.26	979.50	-2.94	.003	.27
	.14	-0.40	1.03	45.46	0.20	1.11	65.26	979.00	-2.95	.003	.27
	.15	-0.39	1.03	46.50	0.16	1.10	64.78	1017.50	-2.72	.007	.25
	.16	-0.44	1.05	46.27	0.11	1.11	64.89	1009.00	-2.77	.006	.26
	.17	-0.42	1.00	47.08	0.08	1.10	64.51	1039.00	-2.60	.009	.24
	.18	-0.44	0.97	47.88	0.06	1.10	64.14	1068.50	-2.42	.015	.22
	.19	-0.42	0.93	49.49	0.01	1.09	63.40	1128.00	-2.07	.038	.19
	.20	-0.40	0.91	48.96	0.02	1.10	63.64	1108.50	-2.19	.029	.20
$G*_{rtdpositions}$	5	0.01	1.07	40.15	1.61	2.28	67.72	782.50	-4.10	<.001	.38
	6	0.05	1.07	40.15	1.64	2.29	67.72	782.50	-4.11	<.001	.38
	7	0.00	1.04	39.77	1.64	2.34	67.89	768.50	-4.19	<.001	.39
	8	0.04	1.05	40.82	1.65	2.39	67.41	807.50	-3.96	<.001	.37
	9	0.03	1.06	40.72	1.60	2.31	67.46	803.50	-3.98	<.001	.37
	10	0.00	1.03	39.72	1.62	2.31	67.92	766.50	-4.20	<.001	.39
	11	0.08	1.35	41.31	2.05	3.02	67.18	825.50	-3.86	<.001	.36
	12	0.13	1.07	41.01	1.61	2.39	67.32	814.50	-3.92	<.001	.36
	13	0.04	1.08	41.12	1.63	2.43	67.27	818.50	-3.90	<.001	.36
	14	0.06	1.10	42.18	1.62	2.44	66.78	857.50	-3.67	<.001	.34
	15	0.09	1.10	42.65	1.64	2.49	66.56	875.00	-3.57	<.001	.33
	16	0.00	1.06	40.96	1.65	2.51	67.34	812.50	-3.94	<.001	.36
	17	0.07	1.11	42.26	1.63	2.50	66.74	860.50	-3.66	<.001	.34
	18	0.06	1.00	42.22	1.65	2.54	66.76	859.00	-3.68	<.001	.34
	19	0.10	1.07	43.15	1.66	2.60	66.33	893.50	-3.47	.001	.32
	20	0.41	1.10	46.64	1.87	2.75	64.72	1022.50	-2.71	.007	.25
	21	0.14	0.95	42.82	1.69	2.56	66.48	881.50	-3.57	<.001	.33
	22	0.19	1.04	43.62	1.73	2.57	66.11	911.00	-3.39	.001	.31
	23	0.22	1.02	45.20	1.66	2.59	65.38	969.50	-3.06	.002	.28
	24	0.22	0.94	45.18	1.68	2.61	65.39	968.50	-3.08	.002	.28

	25	0.19	1.04	44.68	1.68	2.57	65.63	950.00	-3.20	.001	.30
$G^*_{rtdp\text{-}value}$	.01	-0.04	1.05	39.46	1.55	2.16	68.04	757.00	-4.25	<.001	.30
	.02	-0.03	1.03	40.14	1.53	2.20	67.73	782.00	-4.11	<.001	.38
	.03	-0.08	1.05	40.12	1.48	2.18	67.73	781.50	-4.11	<.001	.38
	.04	-0.17	1.02	40.77	1.34	2.15	67.43	805.50	-3.97	<.001	.37
	.05	-0.14	1.01	40.53	1.35	2.17	67.54	796.50	-4.02	<.001	.37
	.06	-0.18	0.96	40.14	1.34	2.21	67.73	782.00	-4.11	<.001	.38
	.07	-0.24	0.96	40.39	1.25	2.24	67.61	791.50	-4.05	<.001	.37
	.08	-0.23	0.94	41.27	1.17	2.20	67.20	824.00	-3.86	<.001	.36
	.09	-0.25	0.93	41.54	1.13	2.18	67.08	834.00	-3.81	<.001	.35
	.10	-0.29	0.93	41.72	1.07	2.16	66.99	840.50	-3.77	<.001	.35
	.11	-0.28	0.95	43.32	1.01	2.18	66.25	900.00	-3.42	.001	.32
	.12	-0.31	0.88	42.89	0.98	2.17	66.45	884.00	-3.52	<.001	.33
	.13	-0.35	0.89	42.77	0.90	2.13	66.51	879.50	-3.55	<.001	.33
	.14	-0.34	0.91	43.08	0.88	2.10	66.36	891.00	-3.48	<.001	.32
	.15	-0.39	0.89	43.18	0.78	2.05	66.32	894.50	-3.47	.001	.32
	.16	-0.38	0.92	43.81	0.73	2.09	66.03	918.00	-3.33	.001	.31
	.17	-0.36	0.87	44.42	0.78	2.17	65.74	940.50	-3.20	.001	.30
	.18	-0.40	0.80	45.11	0.68	2.17	65.43	966.00	-3.06	.002	.28
	.19	-0.38	0.79	45.55	0.67	2.19	65.22	982.50	-2.98	.003	.28
	.20	-0.34	0.75	46.22	0.75	2.29	64.91	1007.00	-2.84	.005	.26
	.21	-0.33	0.73	49.31	0.65	2.29	63.48	1121.50	-2.15	.032	.20
	.22	-0.28	0.75	50.05	0.64	2.25	63.14	1149.00	-1.99	.046	.18
	.23	-0.32	0.74	50.88	0.58	2.29	62.76	1179.50	-1.82	**.069**	.17
	.24	-0.44	0.81	47.78	0.55	2.29	64.19	1065.00	-2.51	.012	.23
	.25	-0.46	0.79	46.65	0.48	2.19	64.71	1023.00	-2.77	.006	.26

Note. In this table *p* is reported two-tailed, and *z* is corrected for ties. Values in bold indicate invalid index. Results also reported in body.

Appendix F

All Answers to Questionnaire in Experimental Condition

Table F1

Questionnaire Answers of Those Participants That Were Detected by the DFT sorted by Experimental Group

Smart Phone
I marked one question, to be able to search for it faster on my phone. Furthermore, I searched for answers to some small questions. I remember searching for question 9. Also question 7 but I got that one wrong.
I searched three questions on Google and found the answers.
I searched for answers on the internet, I did not consult anyone else and completed the questions myself.
I searched on my phone for answers to questions I thought were difficult, that’s how I got the answer.
I searched on my phone for the answer to a question I did not know and that’s how I got the answer.
In the time I got to use my phone I searched some difficult words. This helped me answer the question.
Internal Collaboration
Everyone started asking questions outloud and giving answers. I looked on the screens of people next to me what they answered to questions I didn’t know.
We looked at the answers to questions fellow students that had already finished the test got wrong.
We consulted someone that had already finished the test to see what the answers were to questions he got wrong.
Proctor Assistance
I asked what a concept meant and because of the information I could pick the correct answer. We went through all the options together to review them.
I asked him the question what ‘self-defense’ (‘noodweer’) meant and he explained. (*Note.* ‘noodweer’ was synonymous to ‘zelfverdediging’ which both mean ‘self-defence’, however the first generally also means ‘severe weather’.)
I consulted the proctor by reconstructing the question with keywords. I also consulted a peer by just asking the answer to a question.
I consulted the proctor, but it did not help much. I did remember some of the questions and got them right.
I asked for information about concepts in the questions that could help me get the answers. I also asked the answer to the question literally, but then the proctor only gave me information that helped me answer it.
Asking the question but in a different way.
I asked the proctor a question about ‘to appropriate’ (‘toeëigenen’).
I asked the question but in a different way.
I asked about something I did not know and he told me the answers.
Cheatsheet
My sheet only helped me with 7 of the questions. Most of the questions were about topics we hadn’t discussed in class yet.
I used the A4 cheatsheet on my desk.

Note. Answers were translated to English roughly to protect the privacy of the participants, as they were sometime too detailed; for instance names of fellow participants were left out.

Table F2

Questionnaire answers of those participants that were not detected by the DFT sorted by experimental group

Smart Phone
I searched on Google for questions and I found some information about it.

	I searched on my phone on Google. I was able to answer three more questions with the information I found.
	I was very uncomfortable using my phone during the test but I did find some information on Google for questions I hadn’t completed yet.
	I remembered the numbers of the questions I had doubts about. Then I typted the questions into Google and changed my answer.
	I googled concepts that appeared on the test. I used the flag function to be able to search for them quickly.
	I found everything via Google and the dictionary.
	I searched for some questions on the internet.
	I searched a French word and whether an agent is allowed to search a vehicle.
	I searched the legality principle and how long a suspect could be detained.
	I searched certain concepts I didn’t know and typed some questions literally into Google. I also consulted my neighbour real quick.
	I searched questions I wasn’t certain about on my phone and changed the answers.
	I searched some questions on my phone. I did not need my phone for anything else.
Internal Collaboration	
	I marked some questions and consulted the student next to me.
	I answered 46 questions myself and then consulted some peers for things I wasn’t sure about.
	We asked two fellow students we know get good grades and selected the answers they told us to.
	I consulted two fellow students. At least one of them told me several answers.
	We consulted with the entire group and answered the questions together.
	I consulted one students and he told me the answers to two questions.
	I consulted my neighbour for some answers.
	Consulted the student next to me and someone that had already finished.
	I discussed questions with my neighbour real quick and copied some of the answers.
	I asked two fellow students some questions and we compared answers.
	I tried comparing my answers to someone that had already finished and I asked other students whether they knew the answers to my questions.
	I looked at other people’s answers and asked them about mine.
	I had almost finished but another student had finished so I looked at his questions and changed some of my answers.
	One person finished the test and gave the answers to the questions he got wrong because they were displayed on the screen.
Proctor Assistance	
	I asked the proctor what a concept meant, which helped me select the correct answer.
	I asked the proctore what a ‘qualification’ (‘kwalificatie’) meant in the question and he told me exactly what it was. And I asked whether a motorbike certificate was considered a drivers license and it was not.
	While taking the test I asked the proctor content questions about certain concepts to get to the correct answers.
	I consulted the proctor when I didn’t know the answer to the question.
	I asked the proctor about one of the concepts I didn’t understand and the proctor explained it.
	I asked the proctor for help with a question and he explained the question and the correct answer.
	Asking the proctor the question.
	I rephrased the question so he would tell me the correct answer. He also gave me additional information.
	I asked the proctor a question so I knew the answer to the exam question.
	The proctor helped me answer some of the questions.
	I asked the proctor questions about concepts I didn’t understand, so I could answer the question.
	I answered all the questions first and marked some I didn’t fully understand and then asked the proctor for information.

Cheatsheet
I had some papers with information on my desk and consulted them hoping to get the right answers.
There were questions with words I didn’t understand, so I consulted my cheatsheet and was able to then answer the question.
Questions I wasn’t sure of I checked with my cheatsheet I made with my book.
I had comprehensive cheatsheets.
The cheatsheet helped, but most questions were about some of the topics we hadn’t discussed in class yet.
I only checked my cheatsheet when I really didn’t know the answer to the question.

Note. Answers were translated to English roughly to protect the privacy of the participants, as they were sometimes too detailed; for instance names of fellow participants were left out.

Appendix G
Additional Research on Response Times

Reflecting on the results and conclusions, questions arised about how the response times were logged by the QMP system; the response time information that was imported by the software. Due to the importance of this kind of detectable behavioural pattern data for the data forensics analyses assessed in this study, it was decided to conduct additional research to provide a supportive context of the results and conclusions reported in the body.

It is stated by QMP documentation that the recorded response time is the time the item was displayed on screen (Questionmark, 2018). A response time should be considered the time it takes a respondent to select their final answer (Van der Linden, 2006), which is slightly different from the stated recording by QMP since the item is still displayed on the screen after the answer is selected and only ended by opening a new question. This division is not further addressed by consulted publications on response time modeling (i.e., Marianti et al., 2014; Van der Linden, 2006; Van der Linden & Jeon, 2012). It is unclear whether reopening the item after the answer has been submitted, adds to this display time in QMP, and if so, dependent on whether the answer was actually changed. This situation occured for many participants, for instance in the Internal Collaboration group. Participants, namely in the Smart Phone group, also mentioned marking items for review later in their test process, in this case for the planned cheating moment; it is unclear what effect this has on the recorded display time.

A small-scale study was designed and executed by the researcher, in collaboration with the company supervisor, as it would provide information affecting not just the data forensics project, but other projects at eX:plain as well.

Research questions and hypotheses

The central research question reads: How are response times logged by QMP? This translates to several subquestions:

Research Question 1: *Does QMP log the actual response time (selecting an answer) or the entire time the item is displayed on the screen?*

Research Question 2: *Is changing the answer while the item is still displayed on screen reflected by the QMP log?*

Research Question 3: *Does reopening an item add to the time logged?*

Research Question 4: *Is reopening and changing the answer to the item reflected by the QMP log?*

Research Question 5: *How does marking an item by a candidate influence the time log?*

The hypotheses remains that QMP logs the time an item is actually displayed on the computer screen (display time). This means that the time after an answer is selected on the screen is added to the display time, as well as the time after reopening the item and potentially reanswering the question. Changing the answer while the item is still displayed on the screen is not expected to be reflected by the QMP record, as selecting the answer does not influence the recorded display time in the first place. It is not expected that marking an item for review reflects in the time log.

Method

Materials. A new 10 item test was developed with items, from an existing item bank, that had been cancelled for irregularities or design errors previously. There was no previous log information about these items. QMP software randomized the items and the test was administered in the secure browser used for all regular computer-based tests. The simulations with this test were done by hand by the two main researchers, using a digital timer independent from the computer.

Procedure. Eight scenarios were simulated at total. To examine the first research question on whether the time until selecting the answer is logged or the time the item is displayed on screen in total, the test is simulated by displaying each question approximately 20 seconds, selecting an answer after the first 10 seconds. Whether changing the answer during the display is reflected by QMP log, the second research question, is examined by copying the simulation procedure for the first research question, but selecting a different answer after the full 20 seconds, then waiting another 10 seconds to close the item.

In order to adress the third research question on whether time after reopening the item on the screen adds to the recorded response time, the scenario is simulated where each item is opened for 10 seconds, answered and each reopened for another 10 seconds after all items have been answered. This scenario is then replicated but instead of answering the item after the first 10 seconds, the answer is now selected after reopening the item for another 10 seconds, just before it is closed. The first scenario is replicated again for the fourth research question, but after reopening for 10 seconds, the answer is changed, before the item is closed, to see whether this is reflected in the QMP log.

To explore the effects of marking a question, a total of three scenarios are simulated, to take into account the previous research questions. In the first simulation, the items were opened and marked after 10 seconds, closed and, after all items had been marked, reopened and answered after another 10 seconds. In the second simulation, the items are answered and marked after the first 10 seconds. The items are then reopened for another 10 seconds before closing the test. In the third simulation this process is repeated, but after reopening and the following 10 seconds the answer is changed before closing up.

Data analysis. To analyze the registration of display times in QMP, the hypotheses in the previous section are translated to expected display times on the test items. These expected display times were augmented with 1 second, to compensate for time to execute the actions, such as selecting the answer to the question, selecting the next item, or marking the item. Except for the second scenario, the expected display time is 21 seconds per item on average. For the second scenario the expected display time is 31 seconds per item. The average display time on the items in each test was compared to the expected models using one sample *t*-tests. The null hypotheses are rejected if the average display times are significantly lower than the expected display times.

Results

The results of the one sample *t*-tests are summarized in Table G1. The average display time on the simulation was significantly lower than the expected response time model for scenarios 3, 5, 7 and 8.

Table G1

Results on One Sample T-Tests of Eight Different Simulation Scenarios Compared to The Expected Display Time Model

	Model	Simulation				
Scenario	*M*	*M*	*SD*	*t*(9)	*p*	Cohen's *d*
1	21	23.40	0.70	10.85	<.001	3.43
2	31	34.00	1.05	12.68	<.001	2.86
3	21	12.40	1.27	**-21.50**	**<.001**	-6.77
4	21	21.00	2.11	0.00	1.00	0.00
5	21	11.50	1.27	**-23.67**	**<.001**	-7.48
6	21	22.00	1.33	2.37	.042	0.75
7	21	10.50	0.71	**-46.96**	**<.001**	-14.79
8	21	10.20	0.92	**-37.17**	**<.001**	-11.74

Note. Only values in bold are considered for final statements. Positive *t*-values, significant or not, are considered irrelevant to the analysis.

Conclusions

The results for scenario 1 confirm the hypothesis that QMP only logs the time that an item is displayed on the screen, not the time until the actual answer is selected. The second scenario confirms that changing the answer while the item is still displayed, is not reflected in the QMP log. However, scenario 3 reveals that the time an item is reopened and displayed again after an answer has been selected previously is not logged. This is confirmed by scenario 4 as the time after reopening is only recorded if the item has not been answered the first time. Scenario 5 adds that changing the selected answer after reopening the item is also not reflected in the display time. Marking the item does not change any of the outcomes; the display time is recorded until the item is closed after selecting a first answer, whether or not this answer was changed or is changed after reopening and displaying the item again. In conclusion, the information from QMP about its recorded response time is actually incorrect; it is not the time an item is displayed, neither the time it takes a candidate to select an answer. The recorded display time by QMP is the time from first opening the answer until the item is closed after a first answer is selected. Marking, redisplaying or reanswering the item is not reflected in the log.