Psicothema

Psicothema 2020, Vol. 32, No. 4, 549-558
doi: 10.7334/psicothema2020.86

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Detecting Cheating Methods on Unproctored Internet Tests

Susana Sanz[1], Mario Luzardo[2], Carmen García[1], and Francisco José Abad[1]
[1] Universidad Autónoma de Madrid and [2] Universidad de la República de Uruguay

Abstract

Background: Unproctored Internet Tests (UIT) are vulnerable to cheating attempts by candidates to obtain higher scores. To prevent this, subsequent procedures such as a verification test (VT) is carried out. This study compares five statistics used to detect cheating in Computerized Adaptive Tests (CATs): Guo and Drasgow's Z-test, the Adaptive Measure of Change (AMC), Likelihood Ratio Test (LRT), Score Test, and Modified Signed Likelihood Ratio Test (MSLRT). **Method:** We simulated data from honest and cheating candidates to the UIT and the VT. Honest candidates responded to the UIT and the VT with their real ability level, while cheating candidates responded only to the VT, and different levels of cheating were simulated. We applied hypothesis tests, and obtained type I error and power rates. **Results:** Although we found differences in type I error rates between some of the procedures, all procedures reported quite accurate results with the exception of the Score Test. The power rates obtained point to MSLRT's superiority in detecting cheating. **Conclusions:** We consider the MSLRT to be the best test, as it has the highest power rate and a suitable type I error rate.

Keywords: Secure online testing, online assessment, verification tests.

Resumen

***Métodos de Detección del Falseamiento en Test Online*. Antecedentes:** las pruebas de selección en línea sin vigilancia (UIT) son vulnerables a intentos de falseamiento para obtener puntuaciones superiores. Por ello, en ocasiones se utilizan procedimientos de detección, como aplicar posteriormente un test de verificación (VT). El objetivo del estudio es comparar cinco contrastes estadísticos para la detección del falseamiento en Test Adaptativos Informatizados: Z-test de Guo y Drasgow, Medida de Cambio Adaptativa (AMC), Test de Razón de Verosimilitudes (LRT), Score Test y Modified Signed Likelihood Ratio Test (MSLRT). **Método:** se simularon respuestas de participantes honestos y falseadores al UIT y al VT. Para los participantes honestos se simulaban en ambos en función de su nivel de rasgo real; para los falseadores, solo en el VT, y en el UIT se simulaban distintos grados de falseamiento. Después, se obtenían las tasas de error tipo I y potencia. **Resultados**: Se encontraron diferencias en las tasas de error tipo I entre algunos procedimientos, pero todos menos el Score Test se ajustaron al valor nominal. La potencia obtenida era significativamente superior con el MSLRT. **Conclusiones:** consideramos que MSLRT es la mejor alternativa, ya que tiene mejor potencia y una tasa de error tipo I ajustada.

Palabras clave: evaluación en línea segura, evaluación en línea, test de verificación.

The number of companies that have decided to do part of their recruitment processes online is increasing. In addition, due to the current COVID-19 pandemic situation, the number of unproctored evaluations is expected to be higher in many of contexts, including the selection of students. This will allow for increased efficiency and savings as a larger number of candidates can be assessed in a short time (Tippins, 2009). Tests that candidates complete through the Internet are called as *Unproctored Internet Tests (UIT)*. The possibility of conducting tests virtually through the use of technology provides a great deal of flexibility. However, it does also lead to some problems, including the resort to fraudulent means to answer questions. These behaviors are known as *cheating*, and involve the use of unauthorized material or the use of assistance from another individual to fulfil the selection tasks (International Test Commission, 2006; 2016). The possibility of cheating means that UIT scores are deemed to be necessary but not sufficient for hiring individuals since a high score may not correspond to the candidate's real ability (Tippins et al., 2006). Therefore, it is necessary to implement control mechanisms so that we can be certain that the recruited candidate has the required abilities or knowledge. So far, the solutions to this conundrum have focused on developing mechanisms to prevent cheating as well as detecting the incidence of cheating.

Among the solutions to prevent cheating, some of the common mechanisms to be found include general warnings about test content theft, enforcement of copyright laws and the use of web patrols to find if test content has been shared on the Internet (Tippins, 2015; Woods et al., 2020). Some other mechanisms such as the use of usernames and passwords to access test content, the analysis of keystroke analytics, and the use of web cameras, or fingerprint scans to identify candidates are focused on preventing an impersonator from taking a test (Hylton et al., 2016; International Test Commission, 2006; Tippins, 2015). Nevertheless, these measures can be intrusive, and cause negative reactions among the candidates (Guo & Drasgow, 2010). Furthermore, it has been

Received: March 6, 2020 • Accepted: June 5, 2020
Corresponding author: Susana Sanz
Facultad de Psicología
Universidad Autónoma de Madrid
28037 Madrid (Spain)
e-mail: susana.sanzv@uam.es

found that a contract signed by the candidate which stipulates that that test-takers undertake to answer the UIT without any external help as well as reminders about the value of honesty in answering questions is effective in preventing cheating. Moreover, it is highlighted that a candidate who uses fraudulent methods to secure employment may not be competent enough, which would result in an uncomfortable feeling (Dwight & Donovan, 2003; Fan et al., 2012; Tippins, 2015). Finally, the prevention method most related with the current study relates to the warning that selected candidates receive which indicates that they (the ones who reach an establish cut-off on the tests) would have to complete a similar proctored task (Aguado et al., 2018).

Generally, solutions which focus on efforts to detect cheating are more complex. They include an analysis of test responses to determine if the statistics such as the means, standard deviations, pass cut-offs, response patterns, and changes in response latency are in line with expected results. In addition, the use of unauthorized keys, such as cut and paste keys or the 'print screen' key, can be detected (Tippins, 2015). Other sophisticated tools have also been developed. For example, tools can detect if the participants switch from the test page to another window or browser tab (Diedenhofen & Musch, 2017). In some cases, more sophisticated statistical approaches have been proposed, as those based on the detection of aberrant patterns (e.g., where the candidate passes the difficult items but fails easier ones; Tendeiro & Meijer, 2012), or Differential Item Functioning (DIF) analysis (Wright et al., 2014).

In this study, we focus on *psychometric identification,* which is one of the most popular methods used to detect cheating. It consists in the use of a statistic designed to identify if there are changes between the results of the UIT and the execution of a subsequent proctored task, in which candidates complete a shorter version of the test called as a *Verification Test (VT)*.Thus, once we have selected the candidates that reach the level required by the UIT, they are called to complete the VT. After a statistical comparison, a decision about cheating in the UIT is made. If it is determined that the candidate has not used any illegitimate methods, the test results obtained by the candidate on the UIT are taken into consideration as it is understood that a longer test is a better estimation of the candidate's ability (Guo & Drasgow, 2010). Some detection methods have been developed using the Item Response Theory (IRT), as it is necessary to focus on the item's individual properties. And taking takes into account the Standard Error of Estimate (SE) for every candidate, which is not possible with Classical Test Theory. In addition, IRT allows the use of Computerized Adaptive Tests (CAT), in which items are presented according to the previous answers, making the process more flexible, since the candidates do not have to complete the whole test. Thus, the ability level of the candidate is estimated more precisely with less items, and security problems as a result of prior access to the content of the test can be alleviated (Tippins, 2015).

Guo & Drasgow's Z-test

Several statistics have been already proposed. However, some of them make assumptions about the number of cheaters who take a particular test (Cizek, 1999). Tippins et al.'s (2006) revision implies that this number depends, above all, on the perceived profits. If these benefits are perceived as high (like being hired), candidates could take the risk of cheating. Therefore, the proportion of cheaters is unknown in every test, and we consider that it is preferable to use other statistics that does not rely in this information. In this regard, Guo and Drasgow (2010) developed a simulation study in which they proposed two statistics to detect cheating in recruitment contexts: A Likelihood Ratio Test (LRT) and a Z-test. However, their implementation of the LRT was very particular, as they computed the LRT for marginal likelihoods. In the case of the Z-test, their proposal to calculate the z-score is:

$$z = \frac{\hat{\theta}_u - \theta_v}{\sqrt{SE_u^2 + SE_v^2}}, \quad [1]$$

where θ_u and θ_v represent the UIT and VT ability estimates respectively, and SE_u and SE_v represent the standard errors of estimate for the UIT and the VT. The Z-statistic is assumed to be normally distributed, as the maximum likelihood estimations are asymptotically normal (Bock & Mislevy, 1982). The null hypothesis corresponds to no change in the test scores or a higher score on the VT, whereas the alternative hypothesis posits that there is a possibility that the candidate used fraudulent means in the UIT, as the higher score was a result of taking the test in an unproctored environment.

There is some discussion associated with the Z-test. First, some authors say that if it is assumed that the null hypothesis is true, we should calculate the SE of both UIT and VT using the same θ estimate (Finkelman et al., 2010; Lee, 2015), whereas other authors argue that no practical difference between the SE calculated with the same θ estimate or two different ones (Sinharay & Jensen, 2019). Second, when fixed lengths are used for the UIT and the VT, shorter VT's are planned (otherwise, UIT would not be useful), which lead to a higher standard error. The problem worsens when the item bank is small, since the best items would have already been used in the UIT. If the abilities are estimated with maximum likelihood, the results are more likely to return extreme trait and standard error estimates. To avoid standard error overestimations in extreme ability levels, Aguado et al. (201) proposed that the SE take a maximum value of 1, even if it exceeded the limit. This modified method seems to be more effective for individuals with medium to high ability levels, who are likely to take the VT. Despite this, one limitation of this statistic is that the chosen upper limit for the SE is arbitrary. The current study explores several statistical alternatives that do not rely on such dubious decisions.

The aim of this study is to compare the Guo and Drasgow's statistical test with other statistics that may detect cheating on maximum performance tests better, when the item bank is small. In particular, we wanted to study it in the context of a real adaptive test called *eCat listening,* which is a CAT designed to assess English listening comprehension (García et al., 2013; Olea et al., 2011). This CAT was developed due to the success of *eCat grammar* (Abad et al., 2010; Olea et al., 2004) in the selection of candidates for employment and the necessity of considering other skills apart from grammar in recruitment. However, the items bank for the eCat listening is smaller than in the eCat grammar, and the items are easier. In this applied context, the power rates found in Aguado et al. (2018) could be diminished, and the election of the statistical test might be critical. Below, we propose a series of alternatives that we think that are going to demonstrate better power rates.

Adaptive Measurement of Change (AMC)

This statistic was being used largely in clinical psychology and educational measurement, but it can be also used in selection contexts. Finkelman et al. (2010) propose a contrast similar to Guo and Drasgow (2010, which calculates the z-score as follows:

$$|z| = \frac{\left|\hat{\theta}_u - \hat{\theta}_v\right|}{\sqrt{SE_u\left(\hat{\theta}_0\right) + SE_v\left(\hat{\theta}_0\right)}} \quad [2]$$

The difference with equation [1] is that $SE_u(\hat{\theta}_0)$ and $SE_v(\hat{\theta}_0)$ correspond to the same ability level, $\hat{\theta}_0$, which is estimated under the null hypothesis of no change, taking the patterns of UIT and VT as if there were from a unique test; $\hat{\theta}_0$ is assumed to be a better estimate than $\hat{\theta}_u$ or $\hat{\theta}_v$ because it is based on a larger amount of responses. Under the null hypothesis, this statistic follows a normal distribution N~ (0, 1). As we are only interested on the case where UIT is higher than VT, which corresponds to the cheating situation, we deleted the absolute values and transformed it into a one-tailed test, in order to be able to compare the results for different detection methods.

Likelihood Ratio Test (LRT)

As we have said, Guo and Drasgow (2010) compute an alternative LRT, instead of the classic one. We have included the standard contrast (Finkelman et al., 2010; Klauer & Rettig, 2010), which is defined as:

$$\log \Lambda = \frac{\ell\left(\hat{\theta}_0, \hat{\theta}_0\right)}{\ell\left(\hat{\theta}_u, \hat{\theta}_v\right)} \quad [3]$$

where $\ell\left(\hat{\theta}_0, \hat{\theta}_0\right)$ represents the log-likelihood of combined UIT and VT's patterns of responses, under the null hypothesis. This is compared with $\ell\left(\hat{\theta}_u, \hat{\theta}v\right)$, the log-likelihood of separated UIT and VT's patterns of responses, allowing for the possibility of a change on θ. If the obtained value is close to one, we assumed that the null hypothesis is true, as the likelihood is similar. $-2\log\Lambda$ follows χ^2 with one degree of freedom and can be used for two-tailed tests.

For one-tailed tests, $sign\left(\hat{\theta}_u - \hat{\theta}_v\right)\sqrt{-2\log\Lambda}$, which follows a standard normal distribution, can be used (Sinharay, 2017).

Score Test

The Score Test has also been used to compare ability levels estimated in two tests (Rao, 1973). The logic of this contrast is also based on the differences between the results of the UIT and VT when they are joined and separated, and is defined as:

$$R = \left(\frac{\partial \ell(\theta_u, \theta_v)}{\partial \theta_u}\Big|\theta_u = \theta_v = \hat{\theta}_0\right)^2 SE_u^2\left(\hat{\theta}_0\right) + \left(\frac{\partial \ell(\theta_u, \theta_v)}{\partial \theta_v}\Big|\theta_u = \theta_v = \hat{\theta}_0\right)^2 SE_v^2\left(\hat{\theta}_0\right) \quad [4]$$

Where, for example, $\left(\frac{\partial \ell(\theta_u, \theta_v)}{\partial \theta_u}\Big|\theta_u = \theta_v = \hat{\theta}_0\right)$ is the first derivative of the log-likelihood with respect to θ_u evaluated at $\theta_u = \theta_v = \hat{\theta}_0$. SE is calculated as in [2], for the unique test resulting from UIT and VT. If the alternative hypothesis is right tailed $sign\left(\hat{\theta}_u - \hat{\theta}_v\right)\sqrt{R}$, can be used, as it follows a standard normal distribution (Sinharay, 2017). Sinharay and Jensen (2018) detail the mathematical development of the Score test for the two-parameter logistic response model (2PLM), but we have extrapolated the test to a three-parameter logistic response model (3PLM).

Modified Signed Likelihood Ratio Test (MSLRT)

To be useful, VT has to be short, but this usually implies high standard errors. For this reason, the MSLRT (Barndorff-Nielsen, 1986) is based on higher-order asymptotics, where the relative error of the *p* values associated converges to 0 much faster than those obtained with the methods based on first order asymptotics (e.g., the LRT). This is especially useful when the number of observed responses is small (as in the VT). The statistic is based on the calculation of two alternative parameters: ψ, which is the parameter of interest and represents the difference between the two ability levels ($\psi = \theta_v - \theta_u$), and λ, a nuisance parameter that represents the UIT ability estimation ($\lambda = \theta_u$), which needs to be estimated, but does not hold any interest in terms of contrasting the hypothesis. The specific statistic, evaluated under the null hypothesis $\psi = \psi_0 = 0$, that we apply is:

$$r^*(\psi_0) = r(\psi_0) + \frac{1}{r(\psi_0)}\log\frac{q(\psi_0)}{r(\psi_0)} \quad [5]$$

The r* calculation is a LRT's correction based on the second derivative and follows a standard normal asymptotic distribution (e.g. Brazzale et al., 2007). Sinharay and Jensen (2018) explain the algebra of the 2PLM in detail in equations 8-13. In addition, they mention that it cannot be used for the 3PLM because it is necessary that the likelihood of the ability belongs to the exponential family of distributions. However, we wanted to test empirically if the MSLRT works for the 3PLM to detect cheating. The mathematical development of MSLRT for the 3PLM and the corresponding R script with its implementation can be requested from the authors.

Methods

Instruments

As we have stated, the test that we use as a base for the simulations is the *eCat listening*, which contains 95 items fitting the unidimensional factor model (e.g., CFI and TLI >.95, RMSEA <.05), and calibrated with the 3PLM (Olea et al., 2011). The discrimination (slope) parameter signified by *a*, takes values between 0.36 and 2.09 (M = 1.09, SD = 0.38); the parameter *b* which captures the difficulty (location) takes values between -2.75 and 1.16 (M = -0.31, SD = 0.95); and the guessing parameter, *c* takes values between 0.10 and 0.46 (M = 0.28, SD = 0.07). The reliability seems to be very adequate for ability levels between -1.3 and 1.7, as the standard error is lower than 0.3, and it is only above

0.5 for trait levels below -2.2, or above 2.4. In terms of validity, the test seems to predict several criteria related to the level of English reasonably well, such as the number of years spent in academia, and the use of English at home or the level of proficiency achieved. (Olea et al., 2011).

Procedure

The sample size of simulated candidates was 91,000, of which 13,000 were honest and 78,000 attempted to cheat during the test. First, the candidates were programmed to answer UIT items without any systems in place which could check cheating in order to facilitate cheating; subsequently, they also did complete the VT items in a monitored context. We assumed that all candidates answered the questions according to their true trait level in this second test. The honest candidates would answer all the items according to the same true trait level, and we established θ levels from -3 to 3 by 0.5, with 1,000 simulations on each level. In contrast, the cheaters would complete the UIT as they had a higher trait level, so there were 1,000 cases for every superior θ than cheaters could score. Thus, if the real θ was -3 (which corresponds to the VT level), 1,000 simulations were supposed to answer the UIT as if their θ level was -2.5. Likewise, 1,000 simulations were supposed to answer the UIT as if it was -2 in order to reach all the possible combinations. We took 0.5 and 6 (i.e., from -3 to 3) as the lower increment and higher increment on the ability level respectively.

We applied the adaptive algorithm described by Olea et al. (2004) for the eCat grammar test. The test starts with the simulator choosing any ability level between -1 and 1. For the first five questions, the search of the best item was constrained to those with discriminations not exceeding 1. Subsequently, the 5-4-3-2-1 method was applied (the first item is selected among the five more informative items, the second one among the four more informative items, etc.). For the first five items we used a Dodd's method variation which uses the mean between the last estimate ability level and ±2 (2 if there is a correct answer and -2 if there is a fail) when the estimation level was extreme. The = other items were selected based on Maximum Information (MI), and the MLE combined with Herrando's method was used to estimate the ability level. The Herrando's method included the answers to two made-up items: one correct answer for a very easy item and one wrong answer for a very difficult item. This process was aborted when 20 questions were answered. The same process was followed with regard to the VT, but the questions what were asked in the UIT were not repeated, and the test was aborted once 10 of the questions were answered.

Finally, the following statistical tests were computed: Guo & Drasgow's Z-test, AMC, LRT, Score Test and MSLRT. In order to replicate the current conditions of eCat listening, we truncated SE for Guo and Drasgow's Z-test, as Aguado et al. (2018) had suggested. The simulations were conducted with version 3.6.3 of R software (2020), using the IRT routines in the *irtoys* package (*v0.2.1*, Partchev, 2017).

Data Analysis

We calculated type I error and power rates for all the statistic contrasts. A one-tailed test with a significance level of $\alpha = .01$ was conducted, so the rejection took place when the z-statistic was larger than 2.326. For each dependent variable, we tested for significant differences between the detection methods with a repeated measures ANOVA.

Finally, we compared the ROC curves for all the statistics, to capture the overall performance of the test in terms of sensitivity and specificity. The data analyses were conducted with version 3.6.3 of R software (2020).

Results

Type I error rates

First, we examined the type I error rate obtained from every contrast, to determine if these were close to the nominal level of .01. Table 1 shows the type I error rate for every ability level, and the average of them.

As we can see, type I error rates were close to the nominal level, except for the most liberal Score Test. Moreover, we performed repeated measures ANOVA, and the differences between the type I error rates were significant ($\alpha = .05$; $F_{2.257, 27.087} = 15.898$; $p < .0001$; $\eta^2 = .999$). When we tested the multiple comparisons with the Bonferroni correction, significant differences were found mainly for the Score Test (against the Guo & Drasgow's Z-test, the AMC and LRT).

Power rates

Figure 1 illustrates the information regarding the average power rate for detecting cheating, depending on the true ability level of the cheater candidates.

On the whole, power seems poor, as we are averaging all the differences between the ability level estimates. However, we have to take into account that it is difficult to detect low levels of cheating (differences of .5 between UIT and VT's ability levels) and the probability of rejecting the null hypothesis grows as differences increase. This is illustrated in Tables 2-6, which contains all the power rates for every statistic.

Figures 2-4 exemplify conditions when the θ's increments are 0.5, 1 and 3, to demonstrate the working of each of the statistics as a result of the amount of cheating.

Table 1
Type I error rates for all the statistics

Real θ	Guo & Drasgow's Z-test	AMC	LRT	Score Test	MSLRT
-3	.017	.003	.011	.017	.028
-2.5	.011	.005	.007	.013	.015
-2	.013	.004	.009	.020	.012
-1.5	.026	.012	.012	.022	.015
-1	.011	.011	.009	.014	.013
-.5	.015	.012	.016	.031	.019
0	.007	.006	.011	.016	.012
.5	.011	.006	.016	.029	.019
1	.016	.007	.014	.029	.016
1.5	.012	.010	.013	.037	.012
2	.008	.009	.010	.031	.008
2.5	.012	.007	.007	.017	.007
3	.000	.000	.000	.001	.000
Average	.012	.007	.010	.021	.014

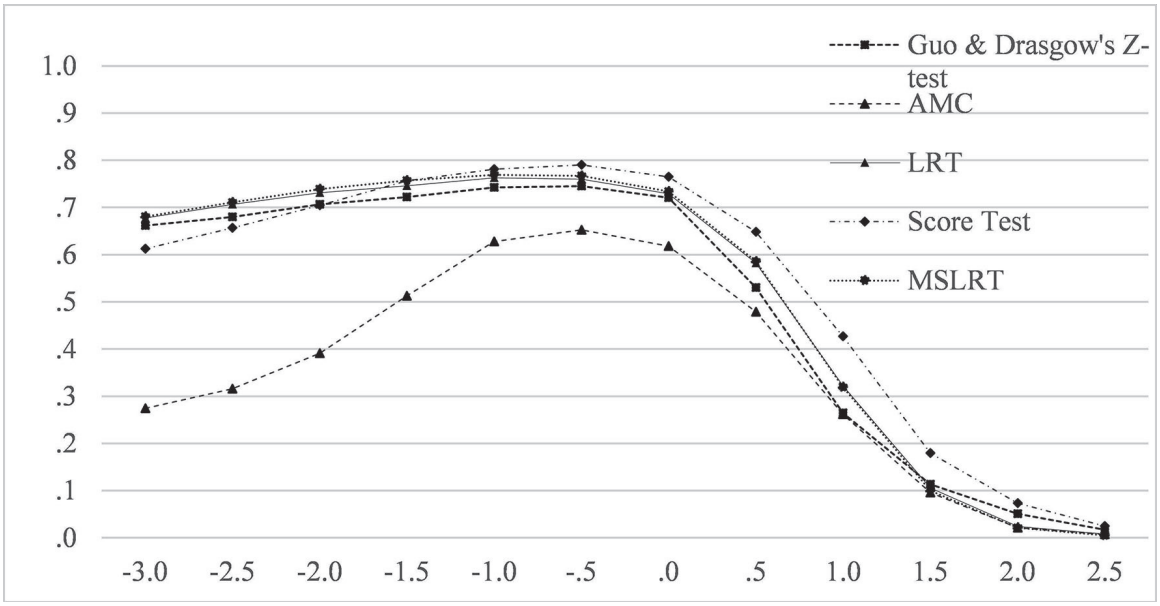

Figure 1. Average power rates for all the statistics

Table 2
Guo & Drasgow's Z-test power

VT	UIT											
	-2.5	**-2**	**-1.5**	**-1**	**-.5**	**0**	**.5**	**1**	**1.5**	**2**	**2.5**	**3**
-3	.024	.042	.175	.340	.617	.840	.925	.981	.998	1.000	1.000	1.000
-2.5		.038	.092	.197	.456	.794	.922	.985	.997	1.000	1.000	1.000
-2			.045	.096	.314	.726	.909	.976	.999	1.000	1.000	.999
-1.5				.040	.172	.522	.819	.957	.988	1.000	.999	1.000
-1					.054	.271	.692	.933	.994	.999	.999	1.000
-.5						.077	.376	.806	.970	.993	.998	.999
0							.109	.458	.857	.951	.962	.986
.5								.112	.381	.620	.723	.815
1									.066	.204	.332	.457
1.5										.040	.106	.194
2											.035	.066
2.5												.017

Table 3
AMC power

VT	UIT											
	-2.5	**-2**	**-1.5**	**-1**	**-.5**	**0**	**.5**	**1**	**1.5**	**2**	**2.5**	**3**
-3	.007	.027	.108	.230	.280	.252	.234	.236	.273	.392	.564	.692
-2.5		.017	.060	.164	.254	.294	.277	.256	.329	.447	.635	.738
-2			.027	.091	.173	.342	.373	.367	.444	.565	.728	.800
-1.5				.026	.100	.276	.535	.591	.645	.746	.821	.876
-1					.028	.186	.534	.759	.815	.877	.892	.934
-.5						.053	.268	.626	.840	.899	.926	.955
0							.052	.262	.667	.862	.917	.947
.5								.044	.250	.546	.725	.828
1									.043	.192	.341	.471
1.5										.041	.092	.154
2											.014	.029
2.5												.007

Table 4
LRT power

VT	UIT											
	-2.5	**-2**	**-1.5**	**-1**	**-.5**	**0**	**.5**	**1**	**1.5**	**2**	**2.5**	**3**
-3	.016	.049	.159	.393	.701	.878	.948	.992	.998	1.000	1.000	1.000
-2.5		.026	.092	.266	.609	.841	.954	.990	.998	1.000	1.000	1.000
-2			.043	.155	.408	.793	.933	.981	.999	1.000	1.000	.999
-1.5				.050	.224	.601	.879	.970	.989	1.000	.999	1.000
-1					.070	.339	.752	.950	.995	.998	1.000	1.000
-.5						.108	.432	.821	.970	.993	.998	1.000
0							.116	.470	.857	.958	.977	.995
.5								.112	.402	.689	.816	.894
1									.085	.270	.420	.515
1.5										.049	.105	.163
2											.016	.032
2.5												.008

Table 5
Score Test power

VT	UIT											
	-2.5	**-2**	**-1.5**	**-1**	**-.5**	**0**	**.5**	**1**	**1.5**	**2**	**2.5**	**3**
-3	.036	.077	.197	.437	.679	.716	.709	.786	.858	.914	.961	.977
-2.5		.038	.133	.303	.609	.746	.790	.800	.899	.943	.977	.984
-2			.054	.194	.456	.752	.824	.892	.934	.963	.985	.991
-1.5				.066	.296	.659	.872	.956	.979	.990	.996	.998
-1					.104	.411	.792	.958	.988	.999	1.000	1.000
-.5						.152	.524	.877	.983	.996	.999	1.000
0							.162	.563	.899	.981	.986	.997
.5								.179	.489	.780	.873	.922
1									.159	.384	.542	.623
1.5										.096	.181	.261
2											.056	.090
2.5												.025

Table 6
MSLRT's power

VT	UIT											
	-2.5	**-2**	**-1.5**	**-1**	**-.5**	**0**	**.5**	**1**	**1.5**	**2**	**2.5**	**3**
-3	.027	.051	.159	.397	.724	.877	.942	.991	.998	1.000	1.000	1.000
-2.5		.031	.084	.290	.632	.847	.949	.989	.998	1.000	1.000	1.000
-2			.048	.180	.449	.804	.929	.982	.999	1.000	1.000	.999
-1.5				.061	.267	.642	.883	.970	.991	1.000	.999	1.000
-1					.086	.363	.760	.953	.996	.998	1.000	1.000
-.5						.124	.452	.832	.972	.993	.998	1.000
0							.123	.482	.866	.960	.977	.995
.5								.120	.411	.688	.817	.894
1									.088	.265	.417	.508
1.5										.046	.102	.153
2											.012	.031
2.5												.005

The AMC test had the worst performance in every situation. The Score Test seemed to show the best results, but since the type I error was not well controlled, the power's rate could be overestimated. Guo & Drasgow's Z-test performed well overall but it was not as good as LRT and MSLRT which were very similar.

When we performed a repeated measures ANOVA, the effect of the method on the power rate was significant ($\alpha = .05$; $F_{1.134, 87.354} = 83.413$; $p < .0001$; $\eta^2 = 1$). Table 7 shows that almost all the pairwise differences between the statistics were significant.

As we can see, the only non-significant comparisons were found for the Score test against the Guo and Drasgow's Z-test, the LRT and the MSLRT. Therefore, even with the highest type I error, the Score test did not do better than the other statistics with regard to the power rates. Finally, we obtained the ROC curve for each statistic, as shown in Figure 5.

ROC curve results point at what we have stated: LRT and MSLRT seemed the best methods in terms of overall performance, and MSLRT was a little superior with regard to the power rate.

Discussion

Due to the global situation caused by COVID-19, online assessments have increased in both employment and educational contexts, creating extensive opportunities for cheating. Thus, it is

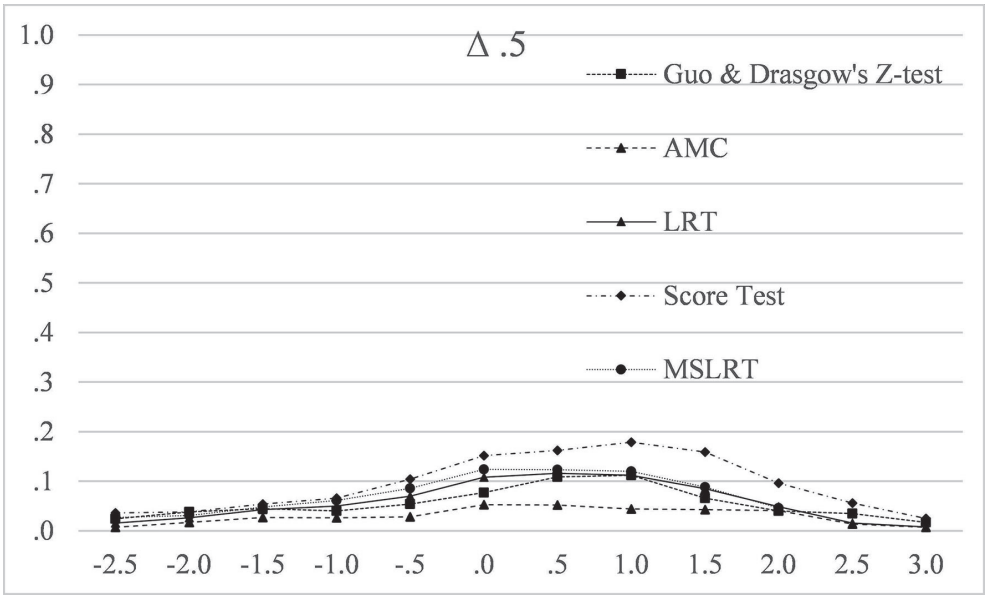

Figure 2. Power with Δ .5 in θ

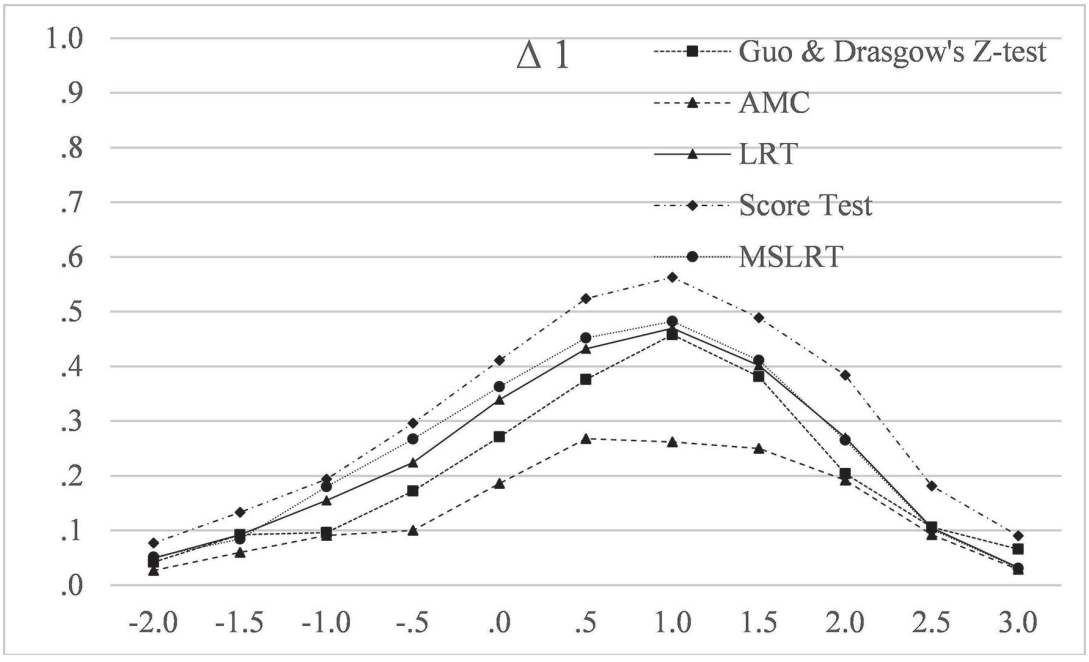

Figure 3. Power with Δ 1 in θ

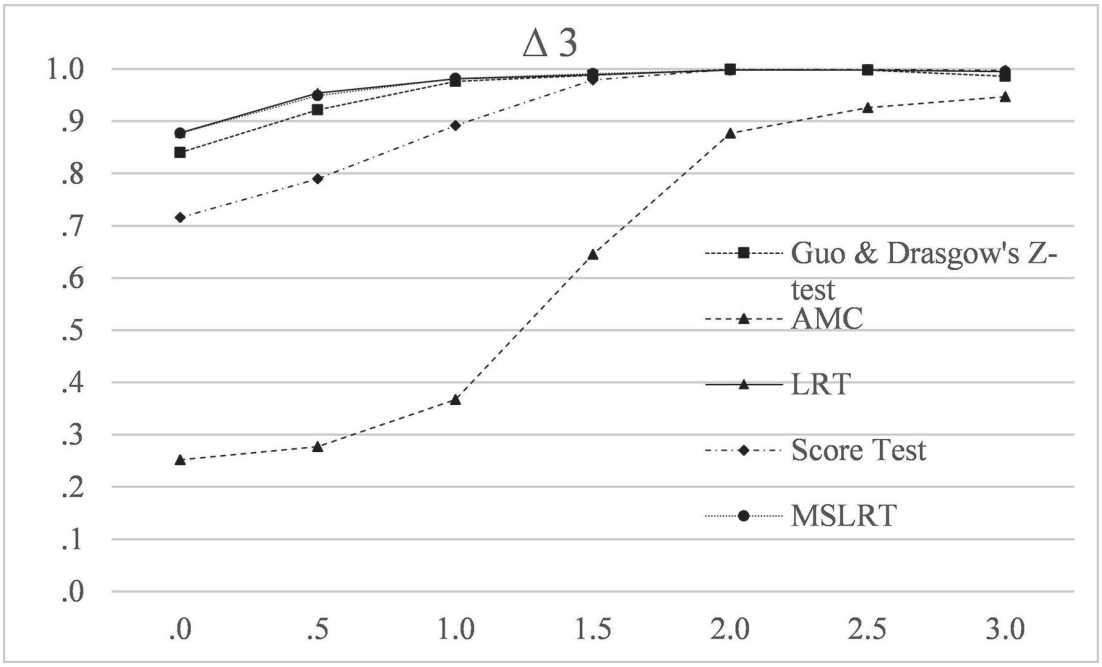

Figure 4. Power with Δ 3 in θ

Table 7
Pairwise comparisons between power rates

Method 1	Method 2	Mean differences	Standard error	p-value
Guo & Drasgow's Z-test	AMC	0.209	0.025	<.001
	LRT	-0.021	0.004	<.001
	Score Test	-0.023	0.010	.200
	MSLRT	-0.026	0.005	<.001
AMC	LRT	-0.230	0.025	<.001
	Score Test	-0.232	0.018	<.001
	MSLRT	-0.235	0.025	<.001
LRT	Score Test	-0.002	0.008	1.000
	MSLRT	-0.005	0.001	.002
Score Test	MSLRT	-0.003	0.008	1.000

especially important to improve cheating detection methods as much as possible. In this study, we focused on a situation in which a short verification test was available. In this particular case, we think that optimizing the cheating detection method is particularly relevant as experience has shown that adaptive VT scores are expected to be less accurate when the item bank is small. Therefore, an adequate balance between false positives and negatives is especially advisable, in order to be fair to the candidates while endeavoring to provide the greatest benefit to the company. The adverse effects of having low power rates are evident, as we selected candidates who did not have the required knowledge or abilities. On the other hand, those who were deemed to be "cheaters" did not go further in the selection process. Therefore, the company could have been losing great workers if the type I error was high, which will lead to harmful long term effects. The effects of our errors can depend

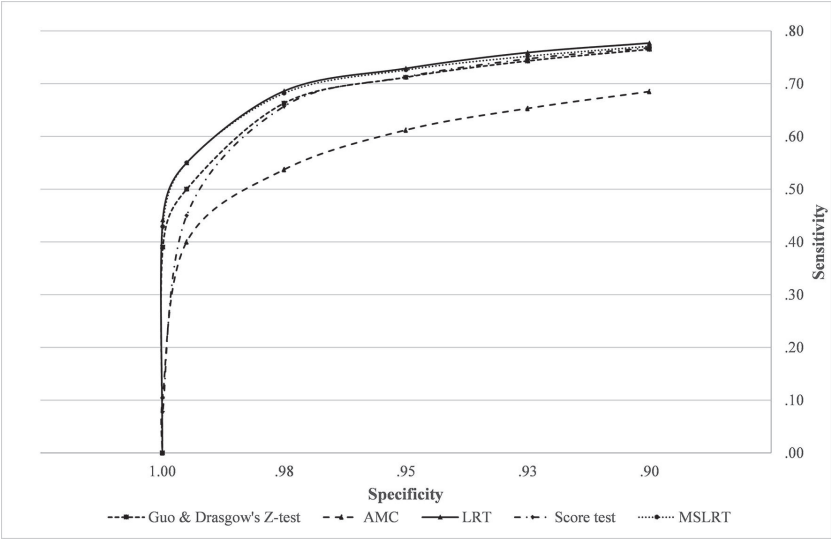

Figure 5. ROC curve for all the statistics until a 0.8 sensitivity level and 0.9 specificity level

on many factors including the cut-off score of the selection process and the frequency and the intensity of cheating (Guo & Drasgow, 2010).

Attending to the results, the best cheating detection method is the proposed MSLRT. Overall it has the most accurate type I error and the best power. Furthermore, it was significantly better than all the other methods, except for the Score Test. However, the Score Test type I error rate is not well controlled. As the ROC curve demonstrates, the MSLRT comes out as the best performing statistic. Thus, we recommend the use of MSLRT to detect cheating in unproctored exams. The R routine that performs MSLRT and its theoretical development can be requested from the authors, in order to simplify the application. In case that the researcher still prefers other methods, LRT could be a good option too, since it is also significantly better than the Guo and Drasgow's Z-test. Moreover, it has the great advantage that it can be used not only in selecting potential employees but also in other relevant contexts. Clinical and educational psychology professionals sometimes need to compare differences between two test applications, when they want to understand if the patients are improving when a particular treatment is administered, or whether students were learning the content of the subject.

Some limitations of the current study should be taken into account. Regarding the MSLRT, Sinharay and Jensen (2018) pointed out that MSLRT can't be applied to 3PLM. Our empirical results prove that its performance is adequate, so we try to demonstrate the possibility of its application mathematically. We intend to address the applicability of the MSLRT to 3PLM in the future. Another important issue is that we have worked with eCat listening test, which includes specific questions as well as a specific CAT algorithm. The relative advantage of the MSLRT statistic might be diminished with different CAT specifications (e.g., with larger VT tests). In this regard, there is some evidence that Bayesian estimation methods perform better in this regard (Wang & Hanson, 1999). Thus, cheating detection statistics can be improved substantially.

Additionally, although the implementation of the MSLRT instead of a Z-test is easy and cheap (e.g., does not require changing the adaptive algorithm), it should be not used without prior consideration. As Tippins argues, there are a lot of preventive measures that the employers could implement, as new challenges emerge from the use of varied digital selection procedures (Woods et al., 2020). A strategy which combines preventive methods with a strong detection strategy with the help of robust statistical analysis is desirable, as the current processes are not sufficient to select the best employees.

Acknowledgements

This research was partially supported by Ministerio de Ciencia, Innovación y Universidades, Spain (Grant PSI2017-85022-P), European Social Fund, and Cátedra de Modelos y Aplicaciones Psicométricos (Instituto de Ingeniería del Conocimiento and Autonomous University of Madrid).

We want to thank Vicente Ponsoda for all his work, help, and good advices in the development of this study.

References

Abad, F.J., Olea, J., Aguado, D., Ponsoda, V, & Barrada, J.R. (2010). Item parameter drift in computerized adaptive testing: Study with eCAT. *Psicothema, 22*(2), 340-347.

Aguado, D., Vidal, A., Olea, J., Ponsoda, V., Barrada, J.R., & Abad, F.J. (2018). Cheating on Unproctored Internet Test Applications: Analysis of a verification test in a real personnel selection context. *The Spanish Journal of Psychology, 21*(62), 1-10. https://doi.org/10.1017/sjp.2018.50

Barndorff-Nielsen, O.E. (1986). Inference on full partial parameters based on the standardized signed log likelihood ratio. *Biometrika, 73,* 307-322. https://doi.org/10.1093/biomet/73.2.307

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431-444. https://doi.org/10.1177%2F014662168200600405

Brazzale, A.R., Davison, A.C., & Reid, N. (2007). *Applied asymptotics. Case studies in small-sample statistics*. Cambridge University Press.

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

Diedenhofen, B., & Musch, J. (2017). Page Focus: Using paradata to detect and prevent cheating on online achievement tests. *Behavioral Research, 49,* 1444-1459. https://doi.org/10.3758/s13428-016-0800-7

Dwight S.A., & Donovan J.J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*(1), 1-23. https://doi.org/10.1207/S15327043HUP1601_1

Fan J., Gao D., Carroll S.A., López F.J., Tian T.S., & Meng, H. (2012). Testing the efficacy of a new procedure for reducing faking on personality tests within selection contexts. *Journal of Applied Psychology*, *97*(4), 866-880. https://psycnet.apa.org/doi/10.1037/a0026655

Finkelman, M.D., Weiss, D.J., & Kim-Kang, G. (2010). Item selection and hypothesis testing for the Adaptive Measurement of Change. *Applied Psychological Measurement, 34*(4), 238-254. https://doi.org/10.1177/0146621609344844

García, P.E., Abad, F.J., Olea, J., & Aguado, D. (2013). A new IRT-based standard setting method: application to eCat-Listening. *Psicothema, 25*(2), 238-244. http://dx.doi.org/10.7334/psicothema2012.147

Guo, J., & Drasgow, F. (2010). Identifying cheating on Unproctored Internet Tests: The Z-test and the likelihood ratio test. International. *Journal of Selection and Assessment, 18*(4), 351-364. https://doi.org/10.1111/j.1468-2389.2010.00518.x

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Dow Jones-Irwin.

Hylton, K., Levy, Y., & Dringus, L.P. (2016). Utilizing webcam-based proctoring to deter misconduct in online exams. *Computers & Education, 92*, 53-63. https://doi.org/10.1016/j.compedu.2015.10.002

The International Test Commission (2006). International guidelines on computer-based and Internet-delivered testing. *International Journal of Testing, 6*(2), 143-171. https://doi.org/10.1207/s15327574ijt0602_4

The International Test Commission (2016). International guidelines on the security of tests, examinations, and other assessments. *International Journal of Testing, 16*(3), 181-204. https://doi.org/10.1080/15305058.2015.1111221

Klauer, K.C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with latent trait model. *British Journal of Mathematical and Statistical Psychology, 43*(2), 193-206. https://doi.org/10.1111/j.2044-8317.1990.tb00935.x

Lee, J.E. (2015). *Hypothesis testing for adaptive measurement of individual change* [Unpublished Thesis Dissertation]. University of Minnesota.

Olea, J., Abad, F., Ponsoda, V., Barrada, J., & Aguado, D. (2011). eCAT-Listening: Design and psychometric properties of a computerized adaptive test on English Listening. *Psicothema, 23*(4), 802-807. https://www.redalyc.org/articulo.oa?id=72722232042

Olea, J., Abad, F., Ponsoda, V., & Ximénez, C. (2004). Un test adaptativo informatizado para evaluar el conocimiento de inglés escrito: diseño y

comprobaciones psicométricas [A computerized adaptive test for the assessment of written English: Design and psychometric properties]. *Psicothema, 16*(3), 519-525. https://www.redalyc.org/pdf/727/72716327.pdf

Partchev, I., Maris, G., & Hattori, T. (2017). Irtoys: A collection of functions related to item response theory (IRT). R package version 0.2.1. https://CRAN. R-project. org/package= irtoys

R Core Team (2020). *R: A language and environment for statistical computing* (Version 3.6) [Computer Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, C.R. (1973). *Linear statistical inference and its applications* (2nd ed.). Wiley.

Sinharay, S. (2017). Detection of item pre-knowledge using likelihood ratio test and score test. *Journal of Educational and Behavioral Statistics, 42*(1), 46-68. https://doi.org/10.3102/1076998616673872

Sinharay, S., & Jensen, J. L. (2018). Higher-Order Asymptotics and Its Application to Testing the Equality of the Examinee Ability Over Two Sets of Items. *Psychometrika*, *84*(2), 484-510. https://doi.org/10.1007/s11336-018-9627-8

Tendeiro, J.N., & Meijer, R.R. (2012). A CUSUM to detect person misfit: a discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36*(5), 420-442. https://doi.org/10.1177/0146621612446305

Tippins, N.T. (2009). Internet alternatives to traditional proctored testing: Where are we now? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *2*, 2-10. https://doi.org/10.1111/j.1754-9434.2008.01097.x

Tippins, N.T. (2015). Technology and assessment in selection. *Annual Review of Organizational Psychology and Organizational Behavior, 2*(1), 551-582. https://doi.org/10.1146/annurev-orgpsych-031413-091317

Tippins, N.T., Beaty, J., Drasgow, F., Gibson, W.M., Pearlman, K. Segall, D.O., & Shepherd, W. (2006). Unproctored Internet testing in employment settings. *Personnel Psychology, 59*, 189-225. https://doi.org/10.1111/j.1744-6570.2006.00909.x

Wang, T., & Hanson, B.A. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23*(3), 263-278. https://doi.org/10.1177/01466219922031383

Woods, S.A., Ahmed, S., Nikolaou, I., Costa, A.C., & Anderson, N.R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology, 29*(1), 64-77. https://doi.org/10.1080/1359432X.2019.1681401

Wright, N.A., Meade, A.W., & Gutiérrez, S.L. (2014). Using invariance to examine cheating in unproctored ability test. *International Journal of Selection and Assessment, 22*(1), 12-22. https://doi.org/10.1111/ijsa.12053