Lingue e Linguaggi
Lingue Linguaggi 59 (2023), 39-63
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v59p39
http://siba-ese.unisalento.it, © 2023 Università del Salento

MALPRACTICE IN ONLINE VERSUS ONSITE COMPUTER-BASED LANGUAGE TESTS
Reflections from the COVID lockdown experience

THOMAS WULSTAN CHRISTIANSEN
UNIVERSITY OF SALENTO

Abstract – Computer-based tests provide a fast, efficient and cost effective means of assessing students, especially if they can be administered online (Chalhoub-Deville 2002; Suvorov, Hegelheimer 2013). Traditionally, they are administered in controlled conditions, where supervision and invigilation (or proctoring) can be conducted in much the same way as it is with paper-based tests. Around the world, the COVID emergency (2020-21) led to the sudden closure of educational institutions and a shift online of all teaching and testing. In this paper, we will examine the effects of such a change in the context of two specific university courses where a set of bespoke Moodle-hosted computer-based (CB) tests were adapted from a controlled and supervised onsite context to a much freer online context using a general business communication platform (Microsoft Team). We compare the results of over 4,500 tests (Lexis and Use of English, Translation, Reading, Listening and Writing Tasks: the first four marked automatically, the last manually) taken in both onsite and online contexts to identify differences in results between both testing environments. The aim is to provide some hard data to ascertain how far online distance-invigilated tests taken under only the most basic supervision can be adequate substitutes for more traditional formats especially in view of both their positive aspects (e.g. greater accessibility and inclusivity for students) and their negative ones (e.g. the greater opportunity for malpractice).

Keywords: online testing; distance invigilation; proctoring; malpractice; cheating.

1. Introduction

During the COVID “lockdown” in Italy (2020-21), educational institutions, like virtually everything else, were suddenly shut down by emergency decree. Remarkably, as it seemed at the time, all teaching and testing was, after only a short delay of about a month, rapidly moved online. In this paper,[1] we will examine the effects of such a change in the context of two specific university courses where a set of bespoke Moodle-hosted CB tests were adapted from a controlled and supervised onsite context to a much freer online context using a general business communication platform (Microsoft Teams).

The tests in question consisted of five custom-designed computer-based tests which served to as an obligatory preliminary to an oral exam[2] at the conclusion of a 1st year English module on two Bachelor’s degree courses.[3] The tests were:

[1] This paper was originally given as a presentation at the ALTE (Association of Language Testers in Europe) 8th International Conference “Language Assessment Fit for the Future”, Madrid, Spain, 26-28 April 2023.

[2] In the Italian system, most university exams are taken orally, although written papers (*prove scritte*, *idonei* or *verifiche* the different terms reflect their different legal status) may contribute to final marks or restrict access to the final course exam proper. On language courses the written test plus final oral test format is conventional.

1. B-level Lexis and Use of English (48 multiple choice questions and 12 one word answers)
2. Translation Italian – English (30 multiple choice questions)
3. Reading Tasks (18 true / false questions, 6 extracts, 3 questions each)
4. Listening Tasks (6 short extracts with 1 multiple choice question each and a longer extract for note-taking, 9 questions)
5. Writing Task (1 composition of 150-200 words, out of a choice of 3 or 4 questions, depending on course, marked by a human assessor)

This CB test had fortunately been designed, created and introduced long before the COVID emergency and was already up and running in December 2018, initially as an option instead of the more traditional paper-based tests (which could only be attempted four or five times a year). This new system was administered in computer laboratories, which, for convenience, we will henceforth refer to as onsite (OS). It rapidly proved very popular with students not least because grading for four of the tests was immediate, and because of the fact that we ran the tests once a week, allowing students to attempt the tests as often as they liked (availability of computers in the room permitting). The one condition was that the final mark was calculated on that of the most recent test, not the highest score overall, which discouraged people repeatedly doing the test on the off-chance of eventually getting a high grade – as if the test were some kind of lottery scratch card.[4]

Once the lockdown was imposed (March 2020), these tests were delivered online via Microsoft Teams, the platform officially selected at a national level to permit the continuation of teaching, testing and examining at universities. The CB tests that we already had were not too difficult to administer this way, seeing that we were using a Moodle platform that students could easily access via internet, and that they were already used to using both for other courses and for private study on this particular course.

Given the suddenness of the emergency and the lack of preparation for the switch to online delivery (henceforth OL), measures that could be taken to ensure security and supervise students while they took the tests were rudimentary, improvised, and far from ideal. For one thing, many students did not have adequate computer equipment or internet connections for us to adopt other than the most basic methods of supervision – for example, efforts to use Safe Exam Browsers (SEB) were thwarted by the complexity of the download and installation process, which many students professed (sincerely or not: we had no way of telling) that they were unable to complete. Furthermore, the SEB that was made available to us did not appear fool proof. The simple method we used, the same recommended by the university, was to have students connect via Teams on their smartphones and place these in a position where we could see them, their computer keyboard, and their screen (Figure 1).

[3] Officially, they were known as: "Computed-Based PORT / *Esonero Scritto*": PORT meaning *Prova Orale* (Oral exam) Readiness Test; *Esonero Scritto*, Exemption from a Written Exam (i.e. the passing of this test meant one did not need to the written exam).

[4] Of course, part of the psychology behind this system was that it encouraged students to attempt tests many times and thus, presumably, to learn and improve by force of practice and repetition alone. This is part of the methodology behind many computer apps like Duolingo, where learners are given such things as daily, weekly or monthly challenges, and encouraged to join leader-boards, all in an attempt to gamify their learning and make them practice as much as possible.

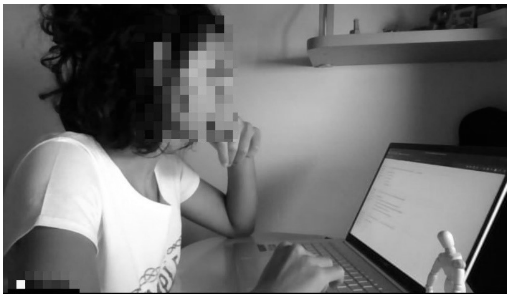

Figure 1
Obligatory camera position during OL tests.

Lastly, given the climate of fear and confusion, especially in the first month or so of the lockdown, we decided to take a relaxed approach and ask for collaboration from students, which most seemed to appreciate, their having understood the gravity of the situation, and because we impressed upon them the idea that, to quote Berkey and Halfond (2015), "Student honesty is a prerequisite for the credibility of online programs". Thanks to the efforts of the university didactic managers and IT staff, who were able to set up a working university-wide system of Microsoft Teams at short notice, we were able, as soon as April (2020), to start weekly tests again. For many students, this constituted, in a small but significant way, a return to normality in otherwise uncertain (and for some distinctly dark) times.

The tests were administered OL but under supervision (by distance invigilation[5]), with, due to a shortage of personnel, between only 2 and 4 invigilators,[6] sometimes with numbers as high as almost 120 test-takers in peak periods. Our hope was that most students would act responsibly and be cooperative and that at least some of the rest would be put off by something akin to what in airports is dubbed "security theatre" – the impression we tried to give that the means that we had at our disposal to invigilate were far more reliable and sophisticated than they actually were.

2. Research Questions

The present research is conducted on data pertaining to the two examining contexts as outlined in §1. It attempts to answer two basic questions, both of which are relevant to the question of OL testing in general, an area which was already in development before the period of COVID[7] but whose adoption was no doubt accelerated greatly by the necessity of

[5] *Invigilator* is the term adopted by ALTE, the Association of Language Testers in Europe, see ALTE (2011, p. 83)

[6] Contrast this with the LanguageCert International ESOL online, where no more than eight candidates are supervised remotely in real time by two trained proctors using multiple screens, and working separately to increase reliability (the ratio may be even tighter for high stakes exams such as the Secure English Language Test valid for UK Visa applications) – see Johnson and Molfetas (2018). Compare this with the Cambridge Assessment English regulations for traditional OS tests, where there is one invigilator for every 25 candidates (and such a ratio is widely used among other members of ALTE: *Association of Language Testers in Europe*).

[7] See for example the tests administered by the EU's Online Language Support (OLS) for the Erasmus Plus programme, European Solidarity Corp volunteers or those going on Youth Exchanges, and the ICS (International Convention on Standards of Training, Certification, and Watchkeeping for Seafarers) Marlins English Language Test for the maritime industry. Both of these tests can be taken online without

stopping all OS tests. Even though this ban has since been lifted, OL tests are set to become a permanent feature of assessment, despite reservations about reliability and security. This is because they undoubtedly offer a cost effective, convenient, and accessible alternative to traditional OS methods.

The two basic questions we intend to address in this study are those which to us seem the most pressing from the point of view of supervision and invigilation (proctoring), namely:

- Do students cheat or indulge in malpractice more in OL tests than in OS ones?
- If they do, does this have a *significant* effect on their marks? I.e. Is it a cause for concern?

It is, of course, difficult to prove malpractice after the fact, especially with the "reasonable degree of certainty" usually required in court by an expert witness: a fluid term that often depends on the precise field and context, but commonly is defined as 90% (see Dias *et al.* 2015). What we hope to do in this research is to find indications that allow us to weigh up the probabilities of malpractice having occurred, and, if it did, to try to estimate how much the reliability of final marks was compromised.

3. Student Survey

To investigate malpractice in these tests, a request was sent to all 1197 students who were registered on the Moodle Course associated with the tests, inviting them to complete an anonymous online survey. In the request, is was underlined that no measures would be taken to identify respondents or punish those who admitted to doing something that they should not have done. The survey contained 24 principal questions, mostly multiple choice but some invited respondents to give short answers. Because responses to such surveys cannot be taken at face value,[8] we added detractors designed to reduce such factors as confirmation bias or social desirability bias.[9] Some questions were intended to verify information given in other replies and therefore identify inconsistences in the respondents' answers.

3.1. Representativeness of the student survey

We received answers from 151 (a reply rate of *only* 12.61%). Technically, this sample is small and there is consequently a real risk of sampling or non-response bias. For a 95% confidence level, around 292 respondents would constitute a representative sample.[10] Furthermore it is not truly random: the respondents selected themselves by accepting the invitation to do the survey. Beyond purely statistical means, it is difficult to gauge how representative the sample is, but one indicator we can measure is gender.[11] By comparing self-declared gender of respondents with the gender conventionally denoted by the given names of all the 1197 students registered on the course, we have some indication, at least, of

supervision. By contrast, the LanguageCert International ESOL online entails remote live proctoring (see previous footnote).

[8] As Stephen J. Dubner, one of the authors of the *Freakonomics* book series and host of the eponymous podcast, often says, "survey data can be the lowest form of data."

[9] See, for example, Groves *et al.* (2009).

[10] www.calculator.net/sample-size-calculator.html?type=1&cl=95&ci=5&pp=50&ps=1197&x=78&y=14.

[11] We are aware of the complex issues behind the concept of gender. Here we make our judgements purely on the gender that the respondents identify themselves as, and the given name that their parents or guardians chose for them, based presumably on their biological sex at birth.

how representative the survey respondents are (Table 1) - here as elsewhere, we give results to two decimal places.

"Which gender are you?" Survey respondents		Gender as indicated by given name Total students registered on course	
"Female"	122 (80.79%)	Conventionally female names (e.g. Maura)	942 (80.72%)
"Male"	22 (14.57%)	Conventionally male names (e.g. Mauro)	225 (19.28%)
"Non-binary"	6 (3.97%)	Gender-neutral names (e.g. Andrea)	31 (2.66%)
"I prefer not to say"	1 (0.66%)		

Table 1
Comparison of declared gender of respondents with gender of all students registered on course as indicated by their given names.

As can be seen from Table 1, the proportion of female respondents corresponds very closely to the proportion of students with conventionally female names registered on the course (80.79% against 80.72%), and for males, the difference was slightly larger but still close (14.47% against 19.28%). This indicates that for one parameter at least the sample is fairly representative.

3.2. Indications of the overall sincerity of respondents in their replies

Below in Table 2, we show the respondents' replies to questions directly regarding their general honesty, and indirectly regarding their propensity to give truthful answers and not to try to hide less commendable aspects of their characters (i.e. social desirability bias). Here and all tables below, we give in brackets the figures calculated as a percentage of the total number of respondents (i.e. 151):

	Yes	Maybe, I'm not sure	No, never
1. Have you ever deliberately done anything that your parents or guardian told you not to do? (Choose one answer)	84 (55.63%)	26 (17.22%)	41 (27.15%)
2. Have you ever deliberately broken the law? (Choose one answer)	21 (13.91%)	100 (66.23%)	30 (19.87%)
3. Have you ever deliberately broken any rules or regulations at school / university? (Choose one answer)	22 (14.57%)	86 (56.95%)	43 (28.48%)
4. Have you ever copied material from a book or internet site and presented it as your own? (Choose one answer)	17 (11.26%)	109 (72.19%)	25 (16.56%)
5. Have you ever copied another student's work and presented it as your own? (Choose one answer)	18 (11.92%)	122 (80.79%)	11 (7.28%)
6. Have you ever helped another student by letting them copy your work when they were not supposed to? (Choose one answer)	81 (53.64%)	32 (21.19%)	38 (25.17%)

7. Have you ever told a teacher or supervisor about another student who was cheating at school / university? (Choose one answer)	6 (3.97%)	138 (91.39%)	7 (4.64%)

Table 2
Respondents' replies to questions that indirectly measure their degree of sincerity.

Questions 1-3 relate to bad things that each and every one of us can be assumed to have done at some time. Giving a negative or even hedged answer (i.e. "Maybe, I'm not sure") to a question like "Have you ever deliberately done anything that your parents or guardian told you not to do?" would indicate strongly that the respondent is not being sincere, as virtually no child obeys their parents or guardians *all the time*. Taking an average of the negative replies to these 3 questions gives a figure of 38. That is to say, 25.17% of respondents show a tendency to say things that they must know are untrue, i.e. to lie; taking an average of the hedged replies (i.e. "Maybe, I'm not sure") to the same questions gives us a figure of 70.67 respondents (or 46.80%) who show a tendency not always to be sincere, even with themselves. Both these forms of behaviour are, after all, perfectly human; they highlight the limits of surveys especially when not backed up by other forms of data.

Questions 4-6 relate to activities that any teacher will confirm are typical of students, at least in contexts like the one in which these respondents come from. We can be less sure that absolutely every student indulges in them but we can at least be sceptical of those who reply with an outright "no" (on average 48.67 or 32.23%). The number of hedges is also high here: on average 87.67 or 58.06%.

Question 6 is interesting because, together with no. 1, this is where respondents seem to be prepared to be most sincere. Indeed, these are the only two questions in which the number of hedgers is not superior to those giving an unambiguous 'yes' or 'no'. The reason may be because both these questions relate to issues that, though technically "wrong", in some senses, can be viewed as less reprehensive: in the case of Question 1, the natural and often legitimate desire of young adults to rebel against their perhaps misguided or unreasonable elders; in that of 6, the act of putting the welfare of a colleague in difficulty before the edicts issued by some authority.

Question 7 is the perhaps the least black and white of those reported on Table 2. It relates to the taboo of "snitching" or "grassing" on a classmate. For close-knit peer groups like young adult students, this is something that would be considered worse than whatever the person being reported on was doing, except in extreme cases. It is interesting that, with this question, the hedged answer (i.e. "Maybe, I'm not sure") accounts for by far the majority of replies, 138 or 91.39%, with relatively only a few respondents (6 for yes, 7 for no, respectively 3.97% and 4.63%) choosing to come down off the fence. Once again, this underlines how replies in surveys such as this cannot always be taken at face value.

3.3. Attitudes of respondents towards cheating

Before looking at students' attitudes towards cheating, it is of course necessary to examine what they understand by the term: what *they* believe constitutes "malpractice". In Table 3, we show respondents replies when asked to identify different forms of malpractice in tests, listing the answers by order of popularity:

What do you consider cheating? (You may choose multiple answers)	
1. Asking someone else to pretend to be you and to do the test in your name	124 (82.12%)
2. Copying the answers from another student	123 (81.46%)
3. Using the internet or books to look up answers during a test	122 (80.79%)
4. Bringing notes into the test to help you remember what you have studied	69 (45.70%)
5. Asking another student for help with an answer	46 (30.46%)
6. Using a dictionary or online translator during a test	43 (28.48%)
7. Other	29 (19.21%)

Table 3
Activities which respondents recognise as cheating.

As can be seen, none of the options given was identified by all the respondents (151) as cheating, not even something as obvious as one would think no. 1 would be. Indeed, only the first three were considered cheating by a clear majority of respondents (over 80% in each case).[12] This indicates that respondents have a notably different idea of what malpractice is from what we did when we drafted the survey. The implication is that they may be ambivalent about the details of cheating whatever their views are about the general principle (see Table 5 below). It is also interesting that although 29 respondents (19.21%) identified other forms of cheating, no one at all provided any examples when asked to do so in the next question ("If you put 'Other' for the above question, give details").

To understand better respondents' attitudes to what constitutes cheating, it is interesting to consider their answer to another question from elsewhere in the survey, namely "If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, how did you do it?" – Table 4:

If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, which of the things below did you do? (You may choose multiple answers)	
1. I used an electronic / online translator (e.g. Google Translate)	7 (4.64%)
2. I copied from a file on the computer or from the internet.	6 (3.97%)
3. I communicated with another person who helped me	5 (3.31%)
4. I used material that I had hidden (e.g. a dictionary, notes, the internet)	5 (3.31%)
5. I had someone else impersonate me and do the test in my place	3 (1.99%

[12] It is interesting that the form of cheating that is identified by most ("Asking someone else to pretend to be you and to do the test in your name") involves identification of test takers and making sure they are who they say they are. As D'Souza and Siegfeldt (2017) state "One major technical issue with OL [online] exams is authentication regarding who is actually taking the OL exam." Regarding this issue, increasingly sophisticated technology may be able to interpret the distinct biometric signals of test-taker doing a test online and may eventually constitute an affordable and effective way to verify the identity of the latter during the whole test, and not just at the beginning - see Ketab *et al.* (2015).

	)
6. I was able to see another student’s work and I copied from them	1 (0,66%)

Table 4
Methods of cheating used by respondents who admitted to cheating.

In Table 5, we give the answers, again listed in order of popularity, given by the respondents to the question “How serious is the problem of cheating?”

How serious is the problem of cheating? (You may choose multiple answers)	
1. It’s a serious problem because cheating is widespread and mostly goes unpunished	70 (46.36%)
2. It’s not a problem because most cheaters do not get as good marks as they would have if they had studied.	46 (30.46%)
3. It’s not a problem because few people cheat	15 (9.93%)
4. It’s not a problem because most cheaters get caught	11 (7.28%)
5. Other	9 (5.96%)

Table 5
Respondents’ estimation of the seriousness of the problem of cheating.

Although 46.36% of respondents choose the first option, indicating that they consider cheating to be a serious problem, a slightly larger number (together 47.68%), choose options that indicate it is not a problem (Options 2-4). The respondents are then split down the middle on the issue of whether or not cheating is a problem. Almost a third of the respondents choose Option 2, namely that cheating is futile. Two smaller sections of the sample choose Options 3 and 4, which assume, rather naively to our mind, that cheaters are few and that most cheaters are found out. The option “Other” (number 5) is chosen by the smallest number.

Those who chose the “Other” option, were asked in the next question to give reasons for their answer. More people (16) made comments here than actually chose Option 5 (only 9 respondents). What they wrote can be divided into four general attitudes to cheating: seeing cheating as inevitable (1 comment); seeing some justification for cheating and thereby showing some sympathy towards its perpetrators (5 comments); seeing cheating as counterproductive (8 comments); and those who worry about the effects of cheating on honest students who do not cheat (2 comments). We can show these in Table 6 below (reporting the respondents’ actual words):

Sees cheating as inevitable	1. The issue is mid-serious as many people cheat and never get caught, altering the results in the future in case of a job (not fair). On the other end it's a problem that can't be stopped as if someone wants to it will find its way, so there is no actual solution.
Sees some justification for cheating	2. Sometimes if someone cheats is not because they don't study but because either the load of material to study is too heavy for a single test or/and the way the test is presented and the questions made/received by the computer is not correct. Sometimes there are bugs in the computers so even a correct answer is considered wrong, or a question may accept different answers (sometimes it is even bad formulated) but the computer code just accepts one. That has happened to me in your in lab tests (but I pointed that out and you solved the problem changing the source code) and even in another subject, "Lingua tedesca 2", but they don't accept questioning their evaluations.
	3. I believe sometimes students cheat because they have personale problems or it's what they have been doing for a long time, in some cases students do not need to be punished but sometimes they do need to be punished because they cheat on tests they haven't studied for.
	4. I think that while it is a serious problem, students aren't the only ones to blame. They are expected to get perfect scores (both from parents and teachers) or else they get punished, made fun of and they can't get into a good university or even get a job.
	5. I think that everybody sometimes need to cheat, because if you cheat it doesn't necessarely mean that you haven't studied, maybe your mind went blank in that moment and you might have forgotten everything you studied due to the anxiety...also, even if you cheat and you didn't study, in the future you are the one who's gonna be incompetent and none is gonna hire you so...cheating isn't actually that much of a deal in my opinion
	6. It Is a serious problem but more serious Is the reason why many people cheat: they don't find any pleasure nor motivation in studying
Sees cheating as Counterproductive	7. It gets hard to fake skills you don't have. If one scored 30/30 in their grammar test but then during the spoken part of the exam the same person makes far too many mistakes that's a tell-tale and the teacher would know. Therefore the score suffers from it and the cheater still gets a lower mark than they would have if they had studied in the first place.
	8. It's not a serious problem because cheaters demage only themselves and they won't achieve big goals if they don't study
	9. Cheating in a test can be a personal problem because if someone does it, he/she simply doesn't learn anything.
	10. I don't think that cheating is a problem because School or University are basically like gyms for the future job. If we cheat we don't damage anyone alse than ourselves.
	11. I strongly think that cheating is counterproductive.
	12. I think it's serious but at the same time an attentiv professor could notice if the students isn't prepared for the exam oral.The study of languages Is very different from other subjects.A language exam isn't prepared in a week or a month,but Is the study of a path.
	13. I totally disagree with cheating because the acrion goes against the tester and he never prooves his qualities and skills.
	14. It could be viewed as a serious problem, but in the end if people don't learn things it's their problem, not professors' or mine.
Worries about the effects of cheating on others	15. In my opinion the real problem is when people who cheat get higher marks than someone who has really studied and not cheated. The point is that who has studied might feel as all their efforts were in vain, and this may cause a sense of disappointment towards the academic environment. Even though I'm firmly convinced that the main reason for studying should be purely an interest in the chosen topic, I'm also aware of the fact that a lot of people only feel validated when they see their work appreciated by

	others.
	16. E' gravissima perché oltre a non essere leali con i compagni, i quali hanno probabilmente speso molto tempo nello studio di una determinata materia, si è sleali con docenti, i quali spendono molto del loro tempo a preparare i test e a correggerli. [*It is very serious because in addition to being disloyal to fellow students, who have probably spent a lot of time studying a particular subject, you are being disloyal to teachers, who spend a lot of their time preparing tests and correcting them*]

Table 6
Respondents' reasons for choosing the "Other" option in response to the question shown in Table 5.

Broadly speaking, Table 6 shows that the largest number of replies (8 out of 16) falls into the category of considering cheating as counterproductive (Comments 7– 14), and espouses the view that cheaters are essentially harming themselves. The next biggest category (5/16) expresses tolerance of the practice due to an awareness of some of the factors that may lead a student to cheat (Comments 2-6). The smallest category (1/16) is that of seeing cheating as unavoidable, a fact of life so-to-speak (Comment 1). Revealingly, only the last two comments on Table 6 (15 and 16) can be considered to be unequivocally critical of cheating, focusing on the general damage it may cause to the wider community at large. This means that while only a minority of comments on Table 6 may be seen, to a degree, to condone or justify cheating (5/16), i.e. Comments 2-6, only a small proportion (2/16), Comments 15 and 16, actually condemn the practice. This shows that overall, respondents seem to accept the fact that some students will always feel the need (or, in any case, find some excuse) to cheat, and do not regard the practice either as outrageous or unusual.

In a different question, respondents were also asked why they had *not* cheated. For this, they could choose more than one option.

Why didn't you cheat? (You may choose multiple answers)	
1. Because I did not need to	74 (49.01%)
2. Because I liked the challenge of seeing what mark I could get without cheating	60 (39.74%)
3. Because I am against the idea of cheating	45 (29.80%)
4. Because I was scared of getting caught	39 (25.83%)
5. Because it was too difficult to cheat	13 (8.61%)

Table 7
Respondents' stated reasons for not cheating.

As can be seen, the most popular option, Number 1, chosen by very nearly one half (49.01%) of the respondents (74) related to the fact that the respondent did not cheat because they did not need to (which a cynic may say leads open the possibility that had they needed to, they might well have done so). Option 1 also shows a certain pride in their own ability on the part of those respondents who chose it, the subtext being that cheaters are inadequate in some way. The second most popular was Option 2 which was chosen by 39.74% of the respondents (60). This concerns the fact that respondents enjoyed the challenge of doing the test without resorting to malpractice. In third place, Option 3, and most relevant to our discussion of respondents' attitudes to cheating as being discussed in this section, is the fact that, of the reasons for not cheating given on Table 7, only 29.80% of the replies related to the idea of the respondent being against the idea of cheating in principle, which confirms our analysis of Table 6.

Option 4 relates to the security measures adopted to counter malpractice and the students’ estimation of the likelihood of being detected (a topic we will turn to in the next section), although here we have made no distinction between OS and OL Tests. The figure of 25.83% is relatively low, and indicates that most respondents did not consider the anti-cheating measures an adequate deterrent (and so our “security theatre” was ineffective). This view is no doubt influenced by the general lenience shown by those in authority in the local context towards those who cheat, as typically, in our experience at least, if detected, it will lead to a mild rebuke, not even a severe reprimand, and only very rarely indeed to formal disciplinary action of any kind. Finally, Option 5, proved the least popular, and shows that only a small minority of respondents’ reason for not cheating was because they could not think of a practical way to do so (even though again here we do not distinguish between OS and OL tests).

3.4. Respondents’ estimations of how difficult it would be to cheat

In the survey, we also asked respondents about how difficult they considered it to cheat in tests. This gave us an idea not only of their assessment of the security measures adopted, but also indirectly of whether they had a disposition to cheat or not by inclusion of the option “I have never thought about it”.

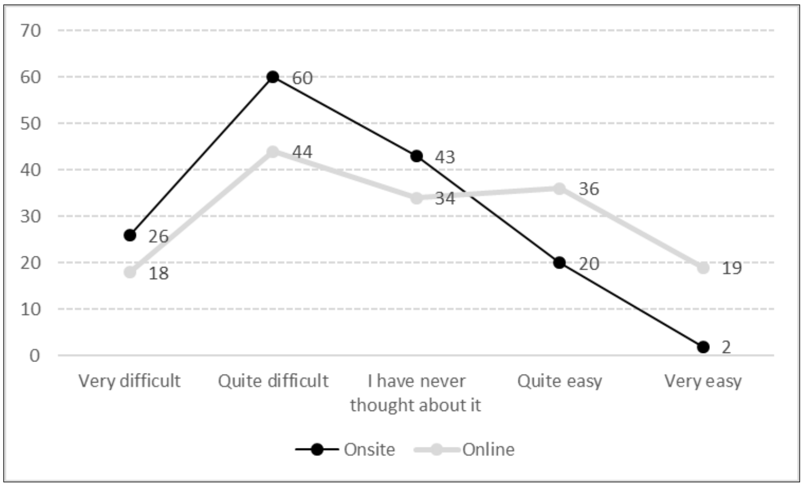

Figure 2
Respondents’ estimations of how difficult it would be to cheat in “a Computer-Based PORT / *Esonero Scritto* under supervision in a computer laboratory or during a lesson” (OS) as opposed to “in an ONLINE Computer-Based PORT / *Esonero Scritto* under supervision via Microsoft Teams.” (OL).

Figure 2 shows that in both testing contexts, OS and OL, the most commonly chosen option was “Quite difficult” (i.e. 60 or 39.74% for OS; 44 or 29.14% for OL) implying that respondents did think that cheating would involve some risk. The option “I have never thought about it”, implying a respondent who is unlikely to have ever cheated or to do so in the future, was chosen by 28.48% (43) in regard to OS contexts and 22.52% (34) for tests taken OL. Most obvious was that respondents regarded it easier to cheat in OL tests than in OS ones, which given the improvised methods we were forced to adopt, by the sudden move online (see §1.0), reflects the reality of the differences in degree and type of supervision that we could enact OS as opposed to OL.

In a different question, respondents were asked why they had cheated, if they had done so, Table 8:

If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, why did you do it? (You may choose multiple answers)	
1. Because it was so easy to cheat, I couldn't resist it	1 (0.66%)
2. Because everyone else was cheating	2 (1.32%)
3. I don't know	3 (1.99%)
4. Because I wanted to get a better mark	12 (7.95%)

Table 8
Reasons respondents who had cheated chose to explain why they had done so.

In relation to Figure 2, it is interesting to see that only one respondent chose Option 1. Similarly, Option 2, was chosen by only two respondents (1.32% of the total who did the survey). The former would indicate that that fact that, even if respondents thought it would be easy to cheat, this would not have led many to do so. The latter also indicates that cheating, even though it could in theory have been widespread, had not become a trend: something that respondents did merely to follow the herd, and to fit in. Option 3, is of course a hedge, and indicates that the respondent, for whatever reason, does not wish to either examine, or at least air their reasons for something which they obviously feel uncomfortable about. Finally, Option 4, which was unsurprisingly the most popular, fed into the belief of by far the most cheaters (two thirds of the respondents who answered this question) that cheating would pay off. It is interesting to remind ourselves of the contrasting attitudes to those expressed on Table 5, where we saw a contrast between the 46.46% who saw cheating as a serious problem (Option 1) and the 30.36% who thought that cheating was in essence a waste of time (Option 2). Obviously any actual cheaters are more likely to fall into the first, larger, group than the second.

3.4. Estimating the number of respondents who cheated in the tests

In the survey, respondents were asked to state whether or not they had cheated. For this question we provided three options, "Yes", "No, never" and the hedge "Maybe, I'm not sure". In Table 9, we can see the answers to this direct question:

Have you ever cheated in a Computed-Based PORT / *Esonero Scritto* either in a classroom / computer lab or online? (Choose one answer)	
Yes	8 (5.30%)
Maybe, I'm not sure	6 (3.97%)
No, never	144 (95.36%)
Blank	1 (0.66%)

Table 9
Number of respondents prepared to admit to having cheated.

As can be seen, 5.3% of the respondents were prepared to admit to having cheated when asked a direct question. By contrast, 95.36% stated that they never had. In between these two groups are the hedgers (3.97%), who, one may presume, are mostly cheaters who are reluctant to say as much, categorically at least. On this basis, a preliminary estimate of the rate of cheating among our respondents could be put at somewhere just between the figures of 8 and 14 (8+6) so at a maximum of around 8-9%.

As we detail in §3.2, it would be naive to take these results at face value and believe that all our respondents are 100% sincere in 100% of their replies. In this section, we shall examine some replies to different but related questions in an attempt to gain some further insight to the actual level of cheating among our respondents.

Respondents were also asked to identify the context (OS or OL) in which they had cheated in these tests, Table 10.

If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, where was it? (Choose one answer)	
OS	5 (3.31%)
OL	10 (6.62%)
Both	4 (2.65%)

Table 10
Test contexts where respondents were prepared to admit to having cheated.

Not surprisingly given our discussion in §3.3, it transpired that 14 out of 19 (i.e. "OL" plus "Both"), or 73.68% of those who replied to this question, had cheated in the OL format, against, in the OS one, 9 out of 19 ("OS" plus "Both") or 47.36% of those answering this question.[13]

It is also relevant to look at the specific parts of the test (see §1.0) where respondents are prepared to admit to having cheated. Table 11 lists the parts in order of popularity, so to speak:

If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, which part/parts was it in? (You may choose multiple answers)	
Writing Task	7 (4.64%)
BLUE Test	6 (3.97%)
Translation	4 (2.65%)
Reading Tasks	3 (1.99%)
Listening Tasks	1 (0.66%)

Table 11
Test types where respondents were prepared to admit to having cheated.

Perhaps unsurprising, the Writing Task paper was that where most respondents stated that they had cheated. This is the one paper where students might feel the need for extra illicit help, and, on a practical level, where they would have time to use it. The BLUE Test comes second (this consists of 48 multiple choice questions and 12 one word answers), and should not cause too many difficulties, but it is in a sense relatively high stakes because the system is set up so that, unless students have passed this (i.e. achieved a score of 60% or higher), they cannot access the other tests. Translation and Reading Tasks are obviously papers where the temptation to use something like Google Translate (see Table 4) may be strong. The Listening Tasks paper comes last, perhaps because it is timed in such a way that students have just enough time to read the questions and listen twice to each question.

Another interesting point to observe in Table 11, is that those who admit to cheating seem to have resorted to malpractice inconsistently, that is, they did not all cheat in all papers (seven respondents said they cheated in the Writing Task but only one did in the

[13] The fact that 73.68 and 47.36 do not add up to 100 (as percentages normally do) is due to the fact that the option "Both" constitutes an overlap category not a distinct one.

Listening Tasks). Furthermore, we cannot assume that, even if they do admit to cheating, they *always* cheated. A plausible scenario is that at least some of the cheaters attempted the tests in a correct fashion and only resorted to malpractice when they deemed it necessary to achieve the mark that they were hoping for.

To return to the question of sincerity of the respondents (see §3.2), there is an obvious inconsistency between Table 10 and Table 9. On the latter, a total of 14 respondents answered either "Yes" or "Maybe, I'm not sure" when asked if they had ever cheated in the tests in questions. On the former we have a total of 19 respondents (i.e. 5 more) giving information about the test context they had cheated in. This gives us good reason to suspect that the figures given on Table 9 underestimate the true numbers of respondents who cheated.

Indeed, we never expected that such a direct question as the one that gave us the answers reported in Table 9 would give us a true picture. In view of this, it is interesting to look at the number of respondents who replied to those other questions in the survey which each, like the question in Table 10 (given here as Question 1), presupposed that anyone who answered had cheated. These we list on Table 12:

	No. of respondents who answered this question
1. If you have ever cheated in a Computer-Based Test, where was it?	19
2. If you have ever cheated in a Computer-Based Test, do you think that your mark improved as a result?	24
3. If you have ever cheated in a Computer-Based Test, which part/parts was it in?	16
4. If you have ever cheated in a Computer-Based Test, which of the things below did you do?	17
5. If you have ever cheated in a Computer-Based Test, why did you do it?	17

Table 12
Number of respondents who answered questions which presupposed that they had cheated.

As we can see, when respondents are not asked directly about cheating but rather, indirectly, things that *presuppose* that one has cheated, many more are prepared to reveal that they have cheated albeit indirectly, or do so unwittingly (and such questions are a favourite trick of interrogators, or barristers in direct or cross examination). Numbers range from 16 (Question 3) to 24 (Question 2). We could take the latter as our final figure, but one has to take into account a margin of error for all these figures due to the fact that, being L2 speakers, some of the respondents may have misinterpreted some of the questions. Question 2 in particular, as a conditional, may have caused confusion with less proficient or more careless respondents. These may have mistaken it for a third conditional: "If you *had* ever cheated in a Computer-Based Test, do you think that your mark *would have* improved as a result", i.e. a question about a hypothesis about what might have happened if circumstances had been different to what in reality they were. For this reason, it is more prudent to take an average for all the questions of Table 12, namely 18.6 (or 12.32% of the total number of respondents to this survey), still about 3% higher than the figure we estimated from Table 9 (i.e. 8-9%).

3.5. Cheaters' perceptions of whether malpractice had improved their mark

The respondents' answers to Question 2 in Table 12 are worthy of special note because this forms the second of our research questions (see §2) namely, if students do cheat, does this have a *significant* effect on their marks? This point will be discussed at length in §4.0 where we will conduct an analysis of test results. In Table 13, we give the estimations by respondents who admitted to cheating as to whether their marks had improved as a result:

If you have ever cheated in a Computer-Based PORT / *Esonero Scritto*, do you think that your mark improved as a result? (Choose one answer)	
No	8 (5.30%)
I don't know	4 (2.65%)
Yes, by a little	10 (6.62%)
Yes, by a lot	2 (1.32%)

Table 13
Estimation of how cheating had affected their marks by respondents who had admitted that they had cheated.

What is noteworthy is that only 2 of the respondents who answered this question (only 1.32% of the total number of respondents) thought that their marks had increased by "a lot" because of their cheating. Just under half of the people who answered this question (10 out of 24) thought marks had increased "by a little", and eight of the respondents (one third of those who replied to this question) thought that cheating had not improved marks – in line with Option 2 on Table 5 ("It's not a problem because most cheaters do not get as good marks as they would have if they had studied"). The general picture from this table is that the majority of cheaters themselves believe that malpractice has little (10/24) or no (8/24) effect on their marks. It can be assumed, that to a degree, they know what they are talking about as they are in a position to compare their own marks with and without cheating. Indeed, their decision to cheat will no doubt most likely have been influenced by the realisation that cheating would help them pass a test or achieve a higher score (Option 4, Table 8), which would probably also go towards their mark in the final oral exam for the course.

4. Analysis of test results

As a counterbalance to the information that we have extracted from the replies to the survey discussed in §3, we can use data from analysis of the actual scores that our students as a mass got in the tests in question.

From §3, we can conclude that students do cheat, by our estimation around 12% (see §3.3). Admitting one has cheated at least once does not mean that one has cheated in every test that one has ever taken. Consequently, we cannot just assume that 12% of the test results that we analyse will have been obtained through malpractice or that, say, one in eleven of tests taken in any single test session has been compromised by malpractice. A more accurate conclusion would be that 12% of the test takers in any single test session have shown in the past the capacity to cheat.

We have also seen in §3 that respondents admit to cheating more in OL tests than in OS ones (Table 10) but most of those who admit to having cheated do not believe that it had increased their mark by more than "a little", if at all (Table 13). Unfortunately, we can neither confirm nor deny whether they are correct in this view (but as educators, we would like to think that they are), precisely because we do not know the identity of those who cheated or in which precise tests they cheated. We can, however, look at how marks differed generally between tests taken OS and OL. If we can find big notable differences in average marks, we will have grounds for suspicion that the greater degree of malpractice, which experience has led us to suspect and which some respondents have confirmed (see Table 10), has led to higher average scores.

In the subsections below we will look at test results, trying to extract data from these which may clarify this issue.

4.1. Number of test results to be analysed

In Table 14, we show the numbers of the various tests divided into OS and OL:

No. Tests	OS	OL	Total
BLUE	270	971	1241
Translation	199	396	595
Reading Tasks	212	590	802
Listening Tasks	309	701	1010
Writing Task	266	597	863
Total	1256	3255	4511

Table 14
Number of tests taken.

As can be seen by looking at the bottom line, a total of 1256 OS tests were taken and 3255 OL ones: 4511 tests overall. In Table 14, we see a notable difference between the two test contexts, while the order of most taken tests is exactly the same for three of the papers (Writing Task, Reading Task and then Translation), in OS contexts the Listening Tasks are taken the most times, while for OL, it is the BLUE test. The reason for this may be that in older versions of these tests, those prior up to the academic year of 2018-2019, the BLUE test was just an *idoneità* (eligibility test); it was necessary to pass it to access the rest of the tests, but the actual mark obtained did not contribute to the overall average of the whole. From 2019-2020, it was treated as a normal test and thus it was worth students' while to retake it to get a better mark as that way they could raise their final average mark of all tests.

4.2. Comparison of marks between Onsite and Online tests

In Table 15, we compare the average marks for the various tests taken in either OS or OL contexts. The figures given were originally out of 30, as is the norm in Italy. Here, we have converted them into percentages (i.e. 30/30 = 100%) for ease of reference:

	BLUE	Trans	Reading	List	Writ	Average
OS	76.63%	84.49%	79.39%	73.59%	63.14%	75.45%

OL	77.21%	87.33%	83.43%	78.74%	67.39%	78.82%
Difference	0.58%	2.85%	4.04%	5.16%	4.25%	3.38%

Table 15
Comparison of marks between OS and OL contexts.

First of all, it can be seen that for each test, marks are higher in OL than they are OS, by as much as 5.16% in the Listening Tasks, but only by 0.58% in the BLUE Test, with an overall average of 3.38%.

We are no closer to proving whether students were cheating, and here data relates to all test takers, not specific individuals, but we can state that if more cheating is going on in the OL tests, as we are led to believe by the survey results (see §3.3), then it can be seen that it is not causing a major rise in marks overall, which is also in line with our conclusions from Table 13.

Indeed, it is illuminating to compare the types of tests where the increase in scores is most marked with the test types that respondents to the survey said that they had cheated in (Table 11), this we do in Table 16:

Tests parts in which respondents said that they had cheated (Table 11), descending order:	Tests types ordered according to mark difference between OS and OL contexts (Table 15), descending order:
Writing Task	Listening Tasks
BLUE Test	Writing Task
Translation	Reading Tasks
Reading Tasks	Translation
Listening Tasks	BLUE Test

Table 16
Comparison of ranking of various test types according to in which respondents said they had cheated (Table 11) and according to mark difference between OS and OL contexts (Table 15).

As can be seen, there is no obvious similarity between the two lists. Indeed, giving each test type a numeral value (i.e. Writing Task = 1, BLUE Test = 2 etc.), we can calculate a correlation coefficient *r* (ratio) of -0.4 between the two sets of data, which counts statistically only as a *weak correlation*.[14] That is to say, there is no obvious link between the test types where respondents said they had copied and the tests types where scores were higher in OL as opposed to OS contexts.

As a case in point, according to Table 11, respondents who admitted cheating did so least of all in the Listening Tasks, yet, looking at results (Table 15), this was the test where the increase in marks in OL contexts was more marked. One explanation for this might be that doing Listening Tasks at home involved fewer distractions, fewer practical difficulties, and less stress all round for the student, because they were in a familiar environment using hardware and equipment of their own. The difference in marks may thus have plausible causes that do not involve malpractice.

[14] The correlation coefficient *r* is a value between -1 and +1 measuring the strength of the relationship between two variables. A value of ±1 = a perfect correlation; above ±0.70, a strong correlation; above ±0.50 a moderate correlation; and above ±0.30 a weak correlation. Below ±0.30 no correlation is said to exist at all. Calculating the correlation coefficient involves a set of complex calculations. Fortunately, there is a special tool in Microsoft Excel (using the classic Pearson formula) that will do it automatically.

Conversely, respondents who had admitted cheating (Table 11) said that they had done so second most in the BLUE Test, which turns out to be the test where the difference in marks between the OS and OL contexts was smallest (Table 15). This would seem to support the opinions expressed by respondents who had admitted that they had cheated (see Table 13), namely that their marks did not increase by more than "a little" as a result.

Another explanation worth exploring is that the students taking the test in OL contexts were indeed cheating and their marks did improve notably, but this fact was obscured by the fact that, compared to the students of previous years (i.e. those using doing the tests OS), they were of a lower linguistic level. Such a hypothesis is possible because it is widely accepted that the COVID emergency and the lockdown, with the resultant shift to teaching online, did have a detrimental effect on teaching, on student's levels of interest and motivation, and consequently on students' performance (see Usher *et al*. 2021). One way that we can test this hypothesis is to compare the marks in the final oral exam between the students who took the tests OS and those who took them OL. If the OL test takers were of a lower linguistic level, one could expect it to be reflected in their oral exam mark.

The oral exam is a conversation in English of about 15 minutes covering both general topics and the specific topics studied in the main course (the lectures given by the course professor). The marks of the various CB tests did generally contribute to the final mark, but are treated merely as a guide, not formally an integral part of the final mark. For example, a student who performed badly in the oral, that is to say, one who could not communicate adequately in English and / or showed no knowledge of the matters covered in lectures, may get a lower mark than their average for the various CB tests, or even fail in extreme cases. Conversely, a student who performed particularly well in the oral but only mediocrely in the CB tests may see their final mark rise significantly above the average for the CB tests. Students in the oral are expected to answer specific questions in a spontaneous and focused manner and not to rely on memory, rote learning, or digression. They are also expected to give their own opinion about a range of diverse topics related to the course and not, all of which makes cheating difficult, even OL.

In Table 17, we show the average oral exam marks achieved by students who had taken the CB tests respectively in OS and OL contexts:

Average oral exam mark for students who has taken CB tests OS	86.28%
Average oral exam mark for students who had taken CB tests OL	90.43%
Difference	4.15%

Table 17
Average marks in final oral exam for students who had taken CB tests OS or OL.

As can be seen, judging by the marks in the oral, there is no evidence that the linguistic level of students who did their CB tests OL during the COVID lockdown was notably lower than those who had done them OS before COVID: if anything, the opposite.

Another calculation that we use to try to discern from the general data whether unidentified individuals were cheating is to look at standard deviation – that is a measure of the dispersal of data around the mean or average: whether individual values are clustered close to the mean or differ more widely from each other.[15] In the former case, it will indicate

[15] E.g. the students in two separate classes may have the same average age (say 14). In one class, where there is a low standard deviation, the students may be 13, 14, 14, and 15 years old. In the other, where there is a high standard deviation, they may be aged 10, 18, 12 and 16. The former group is more homogenous as regards age range.

that there is less difference between marks. This could be taken as indication that fewer students would seem to be benefiting from getting notably higher marks through malpractice. In the opposite case, that of a high standard deviation, it would mean less homogeneity, something that might be explainable in terms of a proportion of students (say about 12% – see §3.3) using malpractice and thereby gaining an unfair advantage over their peers. In Table 18 we give the standard deviation for the average marks listed on Table 15:[16]

	BLUE	Translation	Reading Tasks	Listening Tasks	Writing Task	Average
OS	3.77	4.90	3.69	6.12	5.60	4.82
OL	4.03	3.63	4.09	4.78	5.07	4.32
Difference	0.26	-1.27	0.39	-1.34	-0.53	-0.50

Table 17
Standard deviations for average marks as given on Table 15.

Looking at the overall average (left-most column), it can be seen that overall the difference in standard deviation for the various tests types is small only, 0.50, with the marks in OS contexts being less tightly clustered around the mean than they are in OL tests – showing a greater homogeneity in marks in the latter. In only two of the papers (the BLUE test and Reading Tasks) is the standard deviation higher in OL contexts than in OS contexts, but in both of these the actual difference (0.26 and 0.39) is relatively small.

4.3. Evidence of copying from internet in written tests

Another way that we can use hard data to check the findings of the survey as reported in §3 is to look at student's answers themselves. The obvious place to do this is in the Writing Task where students had to produce between 150-200 of their own words in the accomplishment of a particular task. This output can be checked for plagiarism from the internet by using dedicated commercially available software. For this purpose, we used *Ithenticate*™. When marking Writing Tasks, assessors had been told to look out for plagiarism, i.e. texts or portions of them that looked too good (i.e. were evidence of a highly proficient language user, especially when compared to that student's scores in other papers, which can be easily accessed on Moodle) or answers which, in addition to this, did not answer the specific instructions accompanying each Writing Task. When suspicious answers were found, it was practice to Google a portion of the answer to try to recover the source. To our recollection this procedure was followed on half a dozen or so cases, all of which were annulled after the student had been informed, and given a chance to explain any similarity (which typically they could not). However, as the results below will show, this informal method did not prove fool proof, and *Ithenticate*™ identified ten examples of tests taken in OS contexts where 50 or more words matched internet resources and 23 for OL contexts.

Ithenticate™ works by comparing a given text to sources on the internet looking for matches and calculating a percentage.[17] In Figure 3, we give the results of the analysis of all the Writing Tasks submitted by the students by *Ithenticate*™, that is 266 (OS), 597 (OL), or

[16] Calculating standard deviation involves a complex formula. Fortunately, it can be done automatically in Excel using a special command.

[17] "When a document is scanned *iThenticate*™ calculates a percentage score of the number of matching words from all identified sources compared to the total number of words in the document." *Interpreting Ithenticate™ Reports: A Guide For Researchers*, p.2, latrobe.edu.au/researchers/grs/red.

863 in all (see Table 15), focussing on the average size in number of words (max. 3%)[18] of sections of the Writing Task submitted by the students that were also found in internet resources.

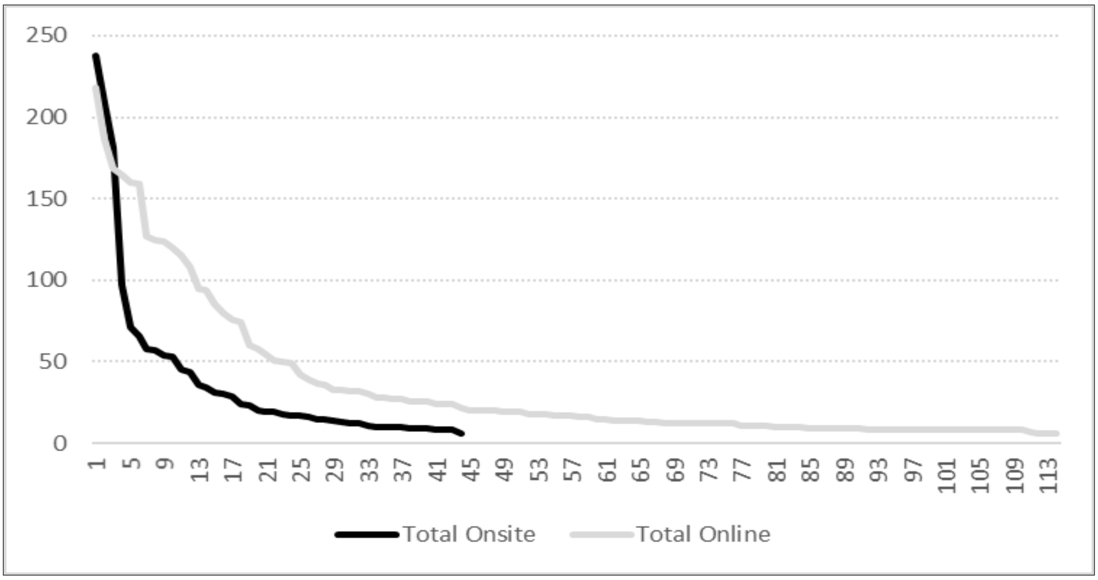

Figure 3
Size in words of similar sections to internet resources (*iThenticate*™).

The y (vertical) axis shows number of words and the x axis the progressive number of the texts analysed in order from those containing largest sections being matched to sections of texts found on the internet to least. For example, for OS the text with the highest amount of text which matched internet sources was text number 1 (238 words) and towards the bottom end 3 texts each had an eight-word block of text that matched internet sources (texts 41-43) and one (44) that had a match of six words. By contrast, the highest score for OL was 218 words.

Looking at the areas enclosed by the two lines on Figure 3, it is evident even to the naked eye that that of OL is larger than that of OS, but it should be taken into account that number of OL Writing Tasks we are examining is over twice that of OS (597 vs. 266).

Overall the average number of words that match internet resources, in texts where such matches were found, for the two text contexts is as shown in Table 18:

	Average matching words OS	Average matching words OL
All texts in corpus	6.37	6.73
Texts where matches above 3% were found	38.52	35.25

Table 18
Average size in words of similar sections to internet resources (*Ithenticate*™).

According to this measure (which is calculated including all texts, even those with less than 3% matches), there is slightly more evidence of plagiarism in OL contexts than in OS but

[18] Below 3%, it is a matter of strings of five words or less. This could include many set constructions and formulaic expressions that are a common part of language, and do not necessarily indicate copying, e.g. “It would not have been”; “I look forward to”; “on the one hand”.

the difference is, on average, only by 0.36 words. If we confine our calculations to texts where there were more than 3% matches, then the average for OS is 38.52 words and OL, 35.25: a difference this time in favour of OS, but again by only a small number: 3.27 words, under 10% of the averages just quoted.

If we look at the effect of such matching on marks, we can see from Table 19, how the presence of similar sections of language to that found on the internet is related to marks. To do this, we take those answers where matches over 3% were found (like row 2 in Table 18) dividing them into 3 categories: fail (<60%), low pass (60%-76.67%) and high pass (80%-100%). In this way, we can see how far the presence of matching correlates with higher marks[19] (Table 19):

	Fail (<60%)	Low pass (60%-76.67%)	High pass (80%-100%)	Correlation coefficient *r*
OS	23.20	24.18	51.85	0.88
OL	36.36	31.73	30.11	-0.96

Table 19
Average size in words of similar sections to internet resources (*Ithenticate*™) according to test mark.

As can be seen, in both contexts, OS and OL, there are strong correlations (i.e. above ±0.70), namely 0.88 for OS and -0.96 for OL, between amount of matching and test mark. However, they are in completely opposite directions, for OS, there is a positive correlation (i.e. the more one parameter increases, the more the other one does too) and for OL a negative one (i.e. one parameter decreases in value as the other rises). In plain language, for OS, texts that contain more matches with internet sources do indeed get higher marks, but with OL, they actually get lower ones. This shows that although there may have been slightly more copying from the internet going on in OL contexts (Table 18), it did not lead systematically to higher marks in confirmation of survey respondents' comments about cheating being counterproductive (see for example Table 6 or Option 2 on Table 5).

Why there should be this difference is difficult to ascertain from the data that we have here alone, but one may suspect that, in OS contexts, assessors were more complacent and assumed that students would not be able to copy from the internet while being supervised in a computer laboratory – a complacency that analysis with *Ithenticate*™ shows was ill-advised. By contrast, it is feasible that in OL contexts, following the sudden move to testing at a distance, assessors were aware that copying would be considerably easier and thus were more vigilant while evaluating the Written Task paper. This would go to show that tests in OL contexts, even those with a low level of supervision, can actually be more secure than those taken OS if the invigilation in the latter is not up to a high enough standard and if proper test conditions are not ensured.

5. Conclusions

From the survey results (§3), one can estimate that between 9 and 12-16% of students admitted (directly or indirectly) to having been involved in some kind of malpractice while taking one or more tests. It is interesting that the students surveyed did not seem entirely clear about what malpractice constituted, which indicates that, for them, there are definite

[19] To do this we simply ascribed to the categories a progressive numerical value: 1 (Fail), 2 (Low pass) and 3 (High pass) and calculated the coefficient *r* using Excel, as before in Table 16.

grey areas as to what is and is not acceptable. For example, while a majority considered that using the internet to look up answers during a test was malpractice, only a minority took a similar view about using a dictionary or online translator during a test: (the latter was indeed, the method that most of the confessed cheaters cited using) – see Table 4.

However, the analysis of the results indicated that, even though malpractice was undoubtedly going on, both in OS and OL tests, it seemed both from the survey, and the common sense conclusion that malpractice would occur in the environments where it was easiest, that it was more frequent in OL test situations. Notwithstanding this, marks did not seem to differ excessively between the two (only by 1.01 marks out of 30, or 3.37%) looking at the population of the test takers as a whole. Of course, it bears repeating that, from this study, we have no direct way of knowing, in individual cases, how much the difference between marks taken with and without malpractice varies.

It is however interesting to note that, with the exception of the Writing Task, there was an inverse relationship between the difference in marks in the various tests and the specific tests that students confessed to having cheated in. For example, the second highest number of students said that they had cheated in the BLUE tests but this was, in fact, the test where the difference in marks, OS vs OL, was smallest (0.58%) – see Fig. 15. Furthermore, by examining the matching of content of the Writing Task with material found online using *Ithenticate*™, we see a positive correlation with higher marks only in OS contexts, but the reverse in OL (see Table 19). It is also revealing to consider that respondents to the survey were divided on the effectiveness of cheating, 30.6% agreeing that "It's not a problem because most cheaters do not get as good marks as they would have if they had studied." (Table 5).

The question that poses itself is that, if someone has to cheat in such tests, do they know enough to do so in such a way that it significantly effects their mark? Table 13 provides some interesting data on this point; only 2 of the 24 confessed cheaters (8.34% or 1.32% of the total population of the survey) believed that their marks had increased by "a lot" as result of malpractice. 16 out of 24 either believed their marks had only increased by a little or not at all. The remaining four could not say.

Despite these doubts over the amount and effect of malpractice in these tests, it is undeniable that the option to administer and take tests online is a welcome innovation for testers and students; one that they may not have considered seriously before the COVID lockdown. Online testing has notable advantages in many different areas (e.g. costs, accessibility, the lack of a need to find and secure suitable premises, or to travel to take tests). The fear of malpractice should not be allowed to outweigh all these. The fact that, in this specific case, the written tests were only a preliminary to an oral exam meant that the stakes were lower than they would have been if they had been the only way of assessing student performance on the course. That said, one should not be too complacent. As Karim *et al.* (2014) note, if one has to rely on the cooperation and honesty of students not to cheat or on further checks, then OL testing is not the simple solution that it at first seems:

> Principled methods, such as using honor codes and decreasing the high-stakes nature of the test through in-person test score verification, do not directly address the issue of cheating or verifying applicant identity. Further, multi-hurdle approaches with in-person testing as a later hurdle may negate some of the cost efficiency of Internet-based tests.

Malpractice can be limited using increasingly sophisticated measures, which as computing power rises may become less cumbersome, and less expensive (as more players enter the market): e.g. safe exam browsers, eye-trackers, screen recorders, key loggers. As in every other sector, Artificial Intelligence (AI) will no doubt offer opportunities not just to cheaters but also to invigilators, for example with the creation of software that, at very least, will be able to able to carry out the kinds of statistical and pattern analyses that can currently be used to identify anomalous or suspicious results of the kind discussed by, among others, D'Souza and Siegfeldt (2017).

However, in such an "arms race" between invigilators and cheaters, effort should also be directed to design tests where malpractice does not significantly alter results. Indeed, it could be argued, as many are doing with the advent of generative AI (and user-friendly applications like Open AI's ChatGPT), that perhaps it would be more appropriate, and more realistic, to allow test takers access to such tools during tests, in the same way that they may use dictionaries or calculators in some traditional exams. In that way, it could be argued that one would, at very least, be creating a level playing field for all candidates, and removing many of the ways that malpractice would otherwise occur (although making sure the test taker is who they are supposed to be would still remain a priority).

How, in such contexts, one can reliably assess a test-taker's linguistic level (and not, for example, their expertise at writing prompts for generative AI applications) will however be a challenge, but there is certainly an argument for considering such a radical move also because tests should reflect what happens in the classroom and in the workplace.[20] The world being what it is, it is difficult to conceive of a situation in the near future where a language user will be reading or writing without the assistance of translators or AI, or other kinds of tool not yet invented or even imagined.

Through the use of learning platforms that make use of more and more sophisticated AI, one may even foresee the situation where tests are no longer necessary, [21] in that one will be able to keep track of a learners in real time, and at any point give a report of their progress and current level. Even on a platform like Duolinguo, which in the near future will no doubt seem relatively simple and unsophisticated, periodic reports on performance are generated. In the case of a habitual user who clocks up many hours of exercises over an extended period, their level of performance will be hard to fake. More technologically advanced applications and platforms could be made to recognise the individual's quirks, e.g. their voice or their particular speed and way of typing, meaning that impersonation by an imposter will be all but impossible.[22]

However, such solutions will only be possible for those taking courses on the same platform for a period of time, and there will still presumably be a demand for traditional stand-alone tests for learners who, for whatever reason, have not been following one particular course on a specific platform.

[20] See the recent guidelines on AI and teaching and assessment drawn up by the Russell Group research intensive universities in the UK, which "will help universities to capitalise on the opportunities of AI while simultaneously protecting academic rigour and integrity in higher education (www.theguardian.com/technology/2023/jul/04/uk-universities-draw-up-guiding-principles-on-generative-ai), or the Danish Broadcasting Corporation report / slide show at "Kunstig intelligens / Gymnasiernes eksamensform matcher ikke undervisningen" ("Artificial intelligence / The secondary schools exam format does not match teaching"): www.dr.dk/stories/1288510966/gymnasiernes-eksamen-matcher-ikke-undervisningen.

[21] A similar view has also been expressed by Prof. Barry O'Sullivan OBE, Head of Assessment Research & Development at the British Council in his talk entitled "The Comprehensive Learning System and integrated education systems" at the British Council New Directions event, in Rome (Italy) 10/05/23.

[22] Incredible as it made seem, even with something as basic as Morse code, individual operators have a personal transmitting style which can allow others to identify them, and which is known as a "fist" – see for example: "Morse Code Hams Let Their Fingers Do the Talking", *New York Times*, By Catherine Greenman, 12/08/1999: www.nytimes.com/1999/08/12/technology/morse-code-hams-let-their-fingers-do-the-talking.html.

The debate about the future of testing is still very much in progress and the fact that we are standing on the cusp of radical advances that will no doubt be brought about by AI makes it very difficult to look too far ahead with any certainty.

Surveys like the one that we conducted (see §3) show that dialogue and negotiation between test givers and test takers can be fruitful. Particularly in Table 5, there is evidence that students do not all approve of malpractice (46.36% believe "It's a serious problem because cheating is widespread and mostly goes unpunished"); more reassuringly on Table 8, only 2 out of the confessed cheaters (1.32% of the total population of the survey) used the excuse that they did so "Because everyone else was cheating". There is thus some common ground between students and educators, and thus there is every reason to hope that open and frank discussion of the issue of malpractice may also lead to greater cooperation and collaboration on all sides, and ideally a reduction in cases of what both sides agree is malpractice.

Finally, such a dialogue, regarding how students are tested and what constitutes malpractice, may also help to reduce exam stress among a generation that worldwide, for whatever reason (and this trend precedes the COVID pandemic but was no doubt aggravated by it), appears to feel the pressure of succeeding more acutely than previous ones,[23] with all the negative effects that this has on their general wellbeing and mental health.

Bionote: Thomas Wulstan Christiansen is an associate professor in English Language and Translation at the Università del Salento (Italy) and Director of the University Language Centre. He has taught in various positions at various universities in Apulia (Italy), the UK, and Poland. He completed his PhD in textual linguistics at Salford (UK). He has researched into various areas of linguistics including systemic linguistics and functional grammar, varieties of English, ELF, teaching English, language testing, and analysis of different corpora, including spoken discourse. He is a qualified teacher trainer and has also worked as an expert consultant for Cambridge Assessment English for many years.

Author's address: thomas.christiansen@unisalento.it

[23] See, for example, Bouchrika (2023).

References

Association of Language Testers in Europe (ALTE). 2011. “Manual for Language Test Development and Examining”. © Council of Europe 2011. https://www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf.

Berkey D. and Halfond J. 2015, *Cheating, Student Authentication and Proctoring in Online Programs*, “New England Journal of Higher Education”. http://www.nebhe.org/thejournal/cheating-student-authentication-and-proctoring-in-online-programs/.

Bouchrika I. 2023, *50 Current Student Stress Statistics: 2023 Data, Analysis & Predictions*. “Research.com” 17/03/2023. https://research.com/education/student-stress-statistics.

Chalhoub-Deville M. 2002, *Technology in Standardized Language Assessment*, in Kaplan R.B. (ed.), *Handbook of Applied Linguistics*, Oxford University Press, Oxford, pp. 471-484.

Dias M.S., Boehmer S., Johnston-Walsh L. and Levi B.H. 2015, *Defining ‘reasonable medical certainty’ in court: What does it mean to medical experts in child abuse cases?*, in “Child Abuse & Neglect” 50, pp. 218-227.

D’Souza K.A. and Siegfeldt D.V. 2017, *A Conceptual Framework for Detecting Cheating in Online and Take-Home Exams*, in “Decision Sciences Journal of Innovative Education” 15, pp. 370-391.

Groves R.M., Fowler F.J. Jr., Couper M.P., Lepkowski J.M., Singer E. and Tourangeau R. 2009, *Survey Methodology*, Wiley, Hoboken.

Johnson R. and Molfetas M. 2018, *Online Proctoring: A key feature of secure high stakes tests in the future?*, presentation given at “ALTE 52nd Conference Day”, 23 November 2018, Salamanca, Spain. https://www.alte.org/resources/Documents/ALTE%20Presentation%20Nov%2018%20Salamanca%20-%20final%20(RJ%20and%20MM).pptx.

Karim M., Kaminsky S. and Behrend T. 2014, *Cheating, Reactions, and Performance in Remotely Proctored Testing: An Exploratory Experimental Study*, in “Journal of Business and Psychology” 29, pp. 1-18.

Ketab S.S., Clarke N. and Haskell-Dowland P. 2015, *E-Invigilation of E-Assessments*, in *Proceedings of the 9th International Technology, Education and Development Conference*, IATED Academy, Madrid, pp. 1582-1591.

Levitt S.D. and Dubner S.J. 2005, *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*, William Morrow, New York.

Suvorov R. and Hegelheimer V. 2013, *Computer-assisted Language Testing*, in Kunnan A.J. (ed.), *Companion to Language Assessment*, Wiley-Blackwell, Malden, pp. 593-613.

Usher E.L., Golding J.M., Han J., Griffiths C.S., McGavran M.B., Brown C.S. and Sheehan E.A. 2021, *Psychology Students’ Motivation and Learning in Response to the Shift to Remote Instruction During COVID-19*, in “Scholarship of Teaching and Learning in Psychology”, pp. 1-8. https://psycnet.apa.org/fulltext/2021-56701-001.html.