Remote Proctoring in language assessment: Exploring the impact on test-takers' scores and perceptions

Radosveta B. Valkova
Lancaster University, UK

This study investigates test-takers' scores, and their perceptions and experiences of remotely proctored (RP) tests in the context of medium-stakes language testing. Seventy English teachers from a Mexican university took Aptis, a computer-delivered British Council language proficiency test, available in test centre and remote versions. Scores, questionnaires, and interview data analyses revealed test-takers' performance was significantly higher with RP for Reading while for Listening and Grammar & Vocabulary no statistically significant differences were observed. Although test-takers experienced anxiety and expressed mistrust in RP and Artificial Intelligence (AI) systems, they also perceived advantages such as greater comfort and practicality. The study highlights the importance of more research about RP systems and AI algorithms.

Keywords: remote proctoring, record and review proctoring, test-takers' perceptions, artificial intelligence proctoring, test-takers' performance

Introduction

The COVID-19 disruption in the field of education not only affected language proficiency tests at an international level, but it also had a significant effect on local or institutional language tests (Isbell & Kremmel, 2020). Many organisations adopted computer-based tests together with remote proctoring, which is identified as "digital techniques for monitoring and controlling student activities during an exam through webcams and internet connections" (Arnò et al., 2021, p. 55).

Even before the pandemic, advances in technology offered an array of advantages for online assessment and learning with a range of technological functionalities:

Email address for correspondence: radost78@gmail.com

monitoring through a webcam, recording using conference tools, 'lockdown' browser and others (Atoum et al., 2017; Foster & Layman, 2013; Weiner & Hurtz, 2017). However, technology does not come without challenges. Security and e-cheating raise many concerns and, as Dawson (2020) suggests, assessment security needs to incorporate approaches which detect and evidence cheating attempts and design innovative methods to curb cheating.

Some technology might provide answers to the problems of e-cheating and security through remote proctoring by AI algorithms and human proctors. The different types of RP solutions are based on the proctoring spectrum: 1) AI-only proctored (suitable for low-stakes tests as there is no human proctoring at all) 2) Record & Review: AI-proctored and recorded and only reviewed by a human if there are violations flagged (suitable for medium-stakes tests); 3) human proctoring in real-time which could be enhanced by AI algorithms (high-stakes tests) (Arnò et al., 2021; Michel, 2020).

However, these solutions could be viewed as concerns surrounding technology in RP. For both test providers and test-takers, some of the main anxieties and controversies in RP are linked to security (assessment and cybersecurity); use of AI (face checks, face and ID authentication, real-time data forensics, keystroke analytics, and biometrics); pre-test processes (onboarding); technological requirements (software and hardware); implications for privacy and safeguarding; and fairness and inclusion.

Research shows that the use of AI could create discriminatory and unfair conditions and technological bias e.g., facial recognition software discriminates against skin colour and gender (Selwyn et al., 2021; Swauger, 2020). It also disseminates inequality and raises questions about impact on test-takers' experiences and performance.

Test-takers' overall experience also relates to privacy concerns, confidentiality, and safeguarding. Data privacy and the invasive nature of RP have intensified the debate and, even though test-takers' concerns and negative experiences have been researched before (Karim et al., 2014), they have increased during the pandemic (Langenfeld, 2020).

Research in online assessment solutions mainly focuses on metrics and performance in two different delivery modes: at-home and onsite tests. While all onsite tests are human proctored, the at-home ones can be: 1. unproctored 2. AI proctored and 3.

human plus AI proctored. Even though English proficiency tests are concerned mainly with proctored solutions, there are low-stakes language tests which might consider unproctored ones. Studies in that area report higher performance results for unproctored tests (Alessio et al., 2017; Domínguez et al., 2019; Karim et al., 2014; Prince et al., 2009; Rios & Liu, 2017), but some conclude that findings are mixed and biased and more research is needed into malpractices (Dadashzadeh, 2020; Steger et al., 2020).

Research on at-home proctored and onsite tests (medium and high stakes) where the impact of the delivery environment is investigated is even more limited. Findings suggest a strong concordance between the two modes with high reliability and score equivalence (Karim et al., 2014; Weiner & Hurtz, 2017; Woldeab & Brothen, 2019; Zumbo, 2021).

However, these studies focus strictly on statistical measures and test-taker performance. Variations in testing conditions, procedures at home, the use of technology, and related concerns need to be further investigated, not just in relation to test-takers' performance, but their experiences as well. The current exploratory study aimed to identify any construct irrelevant variance which impedes/helps test-takers with their test performance and partially fill the void about test-takers' perceptions and experiences with RP systems and AI algorithms.

The research questions the study aimed to answer were:

1. How does remote proctoring affect test-takers' performance?
2. What are test-takers' perceptions of and attitudes towards remote proctored tests?
3. Is there a significant effect of proctoring mode and test preference on test-takers' performance?

Methods

The current research focused on Aptis Remote, which launched in the summer of 2020 using AI-supported RP to answer the demand for reliable and safe testing from home. It is identical to Aptis taken in a test centre (with human proctoring) in terms of format and structure (British Council, 2020).

Aptis is a large-scale international test and is offered in different combinations of skills (Listening, Reading, Speaking, and Writing: skill components), with the Core Component (Grammar & Vocabulary) being compulsory across all packages. For the study, the test-takers were tested on the official Aptis General (only the Listening and Reading skill components and the Core component).

Aptis is aligned to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2018). Numerical scaled scores for all components are given on a scale of 0 to 50 and for the skill components a "CEFR level is assigned according to the score obtained in each skill" (Dunn, 2020, p. 4). Aptis General targets A1 to B2, reporting an overall C level for performances above B2: it does not discriminate between C1 and C2. For more details on Aptis scoring, see Dunn (2020).

Organisations can administer the test themselves or request the British Council to carry out the administrative functions. For both options, the British Council manages security through its computer delivery system and oversees test creation, live test forms, marking, and finalisation of results (O'Sullivan et al., 2020). Administration of tests, set-up, and proctoring procedures are usually supervised by the organisations themselves.

At the time of the study, Aptis Remote was delivered on the Janison Insights platform and AI monitored through a software provided by ProctorTrack (British Council, 2021). The test used a Record & Review proctoring type where the test was recorded and proctored by AI only. In the Record & review proctoring type, the AI uses security algorithms to detect malpractice and ensure validity. If there are any infractions (red flags) detected by the AI system, the finished test is sent for review by human who watches the recording and decides if the violations might render the test invalid.

A mixed-methods research or concurrent design (Cohen, Manion, & Morrison, 2017) was employed, where both quantitative and qualitative data were collected simultaneously. A counterbalanced within-subjects study design was chosen in order to minimise test order and practice effects through randomisation and to control individual test-taker's differences, so the results are statistically powerful and efficient (Zumbo, 2021).

Participants

The participants in the research were 70 English teachers (L1 Spanish) from Benemérita Universidad Autónoma de Puebla (BUAP) in Mexico, which administers Aptis for its students and teachers. The teachers were from the Faculty of Languages and Faculty of English Teaching. Data collected from surveys (see below) showed that most of them taught English (various subjects from a BA in English Teaching) or used English as a medium of instruction and had various English certifications. The data collection coincided with university exam sessions and strict COVID-19 restrictions; thus, university students, who comprise the main target group for Aptis General, could not be recruited for the study (this is addressed in the Limitations section). In addition, due to administrative constraints, Aptis for Teachers could not be utilised for the study. No additional remote proctoring training was given to the participants, and they received standard test-taker communications for Aptis Remote.

Initially, BUAP recruited 120 teachers for the research, but the sample size was reduced to participants who had consented (N = 96). Once all the data was collected, it became evident that the final data pool consisted of 70 teachers who had taken both tests and answered all questions in the questionnaires (participants who completed only one test were excluded from the data analysis).

The teachers' motivation was intrinsic and extrinsic. Firstly, they were already administering different forms of RP for home assessments and various university tests and were keen on acquiring first-hand experience and better understanding of RP for online assessments in general. Secondly, the study provided them with valid Aptis scores which they could use in their professional development at the university if they wished.

The test-takers were randomly allocated to two groups based on their availability and equivalency was attempted. All pre-testing procedures (test-taker recruitment and allocation) were carried out by BUAP with my remote guidance. There were various dependencies which could not be controlled (e.g., test-takers' availability, COVID-19 changing restrictions, family circumstances, etc.) and a perfect allocation was not achieved.

The participants' ages ranged from 22 to 67 (M = 41.50, SD = 8.83) and most were female (73% vs. 27%). On average, the teachers had been teaching at BUAP for over 10 years (M = 11.03, SD = 7.92). It is unclear if the skewed gender in the sample is representative of BUAP only, most tertiary education in Mexico, or both. Statistics from the OECD or Mexico's Ministry of Education do not report data for female teachers in tertiary education in the country. However, there are more female teachers than male teachers in both primary (69.2%) and secondary (51.1%) levels (OECD, 2018). Most of the participants reported living with a partner/spouse and children (38.6%) or with family members (parents, siblings) (22.9%). This question was included in the questionnaire as interruptions at home (e.g., living with other people versus living alone) can be seen as a considerable factor in creating increased anxiety (as shown in the Results section).

Instruments

Aptis and Aptis Remote

The test was operationalised and used in two different settings with four live forms. There were two forms for Aptis at the test centre (TC) in BUAP with live human proctoring and two forms for Aptis Remote at home (RP). The prescribed test administration procedures were followed, which included initially test-takers' enrolment in each environment and the distribution of free access codes for the remote version via email (each test-taker was given three different access codes for each of the components and a login for the platform). During the test, proctoring in centre and support for the remote version followed all procedures to ensure test validity. Then, immediately after the test was submitted, research questionnaires were distributed (through the proctors at the test centre and through email for the remote version).

The test-takers were not given any extra instructions or clarifications and followed regular test-taker procedures for both RP and TC modes. Aptis administration, communication with test-takers and distribution of access codes were not within the scope of the study but have been considered in the data analysis. A similar approach was taken towards RP onboarding processes: the scope did not include examining RP provider practices and system robustness. Even so, recommendations and further research are introduced based on the results of the data analysis.

Questionnaires

The questionnaires were developed specifically for the study and are not part of Aptis or Aptis Remote administrative procedures. They aimed to gather information about the participants and their test experience, focusing on RP. Sections included: demographics and context, general perceptions of RP, Aptis TC and RP perceptions, technical issues, feelings, and preference. For the complete questionnaires for Group A (RP on Day 1) and Group B (TC on Day 1) see Appendices A and B.

Multiple rounds of edits with an expert researcher were employed. The sections were modified, items were written in a way to avoid complex language and repetition, and considerable effort was made to design the RP and TC questionnaires in a logical way for the test-takers (certain questions had to be presented in different logical order depending on the Testing Day).

The questionnaires comprised open- and close-ended questions. Close-ended questions were dichotomous (Yes/No), multiple-choice, and Likert items. Likert items featured a five-point scale: 5-strongly agree, 4-agree, 3-neutral, 2-disagree, and 1-strongly disagree. The internal consistency of the questionnaires was estimated using Cronbach's alpha and the results are reported further below.

Semi-structured interview sessions

Two interview sessions were planned with three participants in each (N = 6) 24 hours after Testing Day 2. Even though they were limited, they provided an opportunity to probe deeper into the test-takers' experiences and gain valuable insights for the overall conclusions. The goal was to elicit additional information based on the test-takers' RP experience which was directly related to some items in the questionnaires.

All six participants were randomly selected from Group A (RP on Testing Day 1). A perfect number of interviewees and allocation was not achieved due to reasons mentioned above: the procedures were conducted by BUAP with my remote guidance, participants' availability was limited, and there were multiple COVID-19 factors.

Data collection and analysis

Scores

On Testing Day 1, Group A (N = 41) took Aptis RP and Group B (N = 29) Aptis TC. On Testing Day 2 (48 hours later), the participants swapped. To evaluate the test-takers' performance from both delivery modes (TC and RP), their score reports were analysed using SPSS. Descriptive and inferential statistics (t-tests, ANOVA, Wilcoxon test) were calculated to examine the quantitative data. Answering RQ1 involved investigating scores for each component (Grammar & Vocabulary, Reading, Listening) in RP and TC and then comparing the results from both tests.

Questionnaires

On both Testing Days, the participants received Microsoft Form links to each questionnaire (in L1) immediately after the end of each test (RP and TC). The questionnaire responses were then translated into English before the data was analysed. For quantitative data (close-ended items), the questionnaires were coded and labelled in SPSS (e.g., Yes = 1, No = 0) and descriptive and inferential statistics were calculated.

The coding scheme for the qualitative analysis emerged from the textual analysis performed on open-ended questions about remote proctoring likes/dislikes, perceived impact of technical issues on performance, and reasons for test mode preference (RP or TC). A computer assisted qualitative data analysis software (CAQDAS), NVivo, was used, and the identified codes (over 20) were then grouped to categorise themes and patterns, relationships, differences, and similarities between them (Cohen et al., 2017).

Subsequent stages involved recruiting an experienced assessment professional as a second coder for reliability on the content analysis. I randomly allocated 10% of the responses to each of the three open-ended questions (i.e. 21 responses in total) for second coding. After receiving the completed file back, a follow-up discussion was arranged to clarify minor inconsistencies and receive feedback, based on which the list of codes was finalised. They included two main categories (positive and negative experiences) and subcategories such as (dis)comfort, onboarding process, platform,

practicality, onboarding anxiety, etc. (see Appendix 3 for the complete list of codes). Inter-coder reliability analysis through NVivo coding revealed an overall Kappa coefficient of .89, where agreement figures above .90 are "acceptable by all" (O'Connor & Joffe, 2020, p. 9).

Semi-structured interview sessions

On the day following Testing Day 2, two semi-structured interview sessions were conducted using Microsoft Teams. Each session was scheduled for 60 mins and a facilitator (an L1 English speaker with a high proficiency in Spanish) was recruited to conduct the interviews in Spanish. An interview protocol was shared with the facilitator beforehand, and the questions were edited to streamline the process and clarify the aims (for the final protocol, see Appendix 4).

The voice-recorded interviews were transcribed in Spanish and translated into English. The total usable data from both interviews was approximately 100 minutes of speech and about 10,000 words of transcription.

A thematic analysis through an inductive approach was conducted to explore in depth what the participants had shared (Hesse-Biber & Johnson, 2015). The goal of the grounded theory approach was to identify key concepts which would provide additional evidence for the research questions (Kiger & VaRIio, 2020; Roulston, 2018). The process for theme identification and mapping involved manual coding in Word (see Figure 1[t]Figure 1 near here[/t]). The themes that emerged were presented in a way that evidenced the data from the questionnaires and supported the RQs.

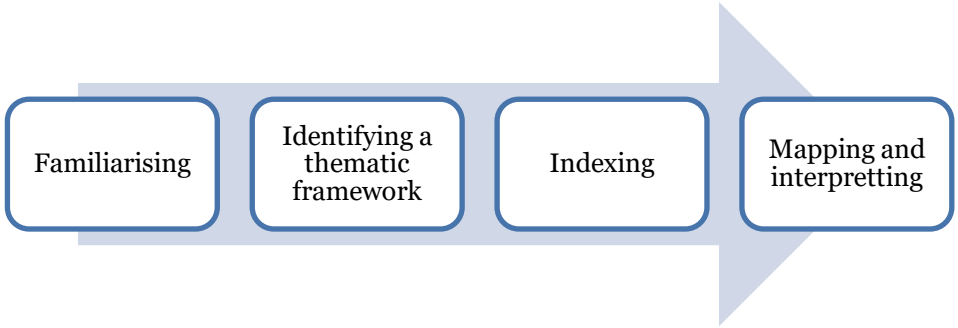

Figure 1. Interviews thematic identification process

Inter-coder reliability was assessed in a similar fashion to the questionnaire, using the same second coder. A ten percent sample of the total recording time (10 minutes) and

the transcription (1006 words) was selected for recoding. This comprised four responses to Q1 from two participants in each interview group. Once the coding data was reviewed and feedback was received, the theme identification scheme was finalised and the whole interview data was coded. The final themes related to participants' RP experience and general views (positive and negative), AI views, and future views and improvement (see Table 1).

Table 1. Final list of themes identified in semi-structured interviews

Themes	Examples
RP experience and general views	
Positive	Comfort
	Practicality
	Use of personal equipment
	Onboarding process
	Good connection
	Ease
	Safety (COVID-19)
	Remote proctoring
	Support
Negative	Anxiety: onboarding, remote proctoring, technical failure
	Onboarding process
	Support
	Discomfort
	Pre-test communication and instructions
	Remote proctoring
AI views	AI: unable to distinguish between natural human gestures and cheating, tech is not developed yet, mistrust, creates anxiety
	Not enough knowledge to understand AI
	Human: immediate support, confidence, trust, social contact
Future views and improvement	RP is here to stay even if there's mistrust; fast development; use in language assessment and other subjects (formative assessment, Social Sciences)
	Need for AI protocols and remote proctoring systems awareness
	Test and RP systems designers should take tests for user experience

Results

Research question 1. How does remote proctoring affect test-takers' performance?

All Aptis tasks were scored automatically by a computer as they consisted of multiple-choice, matching, drop-down, and sentence re-ordering questions (O'Sullivan et al.,

2020). Numerical scores from 0 - 50 were assigned to all three components, and CEFR (A0 - C) levels were also reported for Reading and Listening.

The descriptive statistics in Table 2 show that between the two delivery modes (RP and TC), higher means were reported for all RP components, with Grammar & Vocabulary exhibiting the best test-takers' performance in both delivery modes (for RI: M = 44.73; SD = 3.46; for TC: M = 44.69; SD = 3.29). Within each delivery mode, the lowest performance for RP was observed in Listening (M = 42.03; SD = 5.91) and in TC the lowest was observed in Reading (M = 41.71; SD = 5.67).

Table 2. Descriptive statistics

N = 70	**Remotely Proctored (RP)**			**Test Centre (TC)**		
	Grammar & Vocabulary	**Reading**	**Listening**	**Grammar & Vocabulary**	**Reading**	**Listening**
Mean	44.73	43.86	42.03	44.69	41.71	41.89
Std.Error	.41	.69	.71	.39	.68	.65
Median	45.00	45.00	44.00	45.00	42.00	42.00
Mode	47	50	44	44 & 45[a]	42	44
StDev	3.46	5.76	5.91	3.29	5.67	5.39
Skewness	-2.79	-1.38	-1.71	-3.16	-.65	-1.17
Std.Error	.29	.29	.29	.29	.29	.29
Kurtosis	14.15	2.31	4.23	17.83	.13	1.51
Std.Error	.57	.57	.57	.57	.57	.57
Min	25	24	18	25	26	24
Max	50	50	50	49	50	50

Notes. [a] Bimodal values.

The distribution of CEFR levels for Reading and Listening (see Figure 2) showed that most test-takers achieved C level in both modes, except in Reading TC where most reached B2. Reading RP had more test-takers at C and fewer at B1 levels (N = 44 and N = 3 respectively) compared to Reading TC: C (N = 28) and B1(N = 8).

Listening TC had the highest number of C performances (N = 59), while the corresponding figure for RP was N = 50. The same pattern was observed for B2 level (N = 17 and N = 10 respectively).

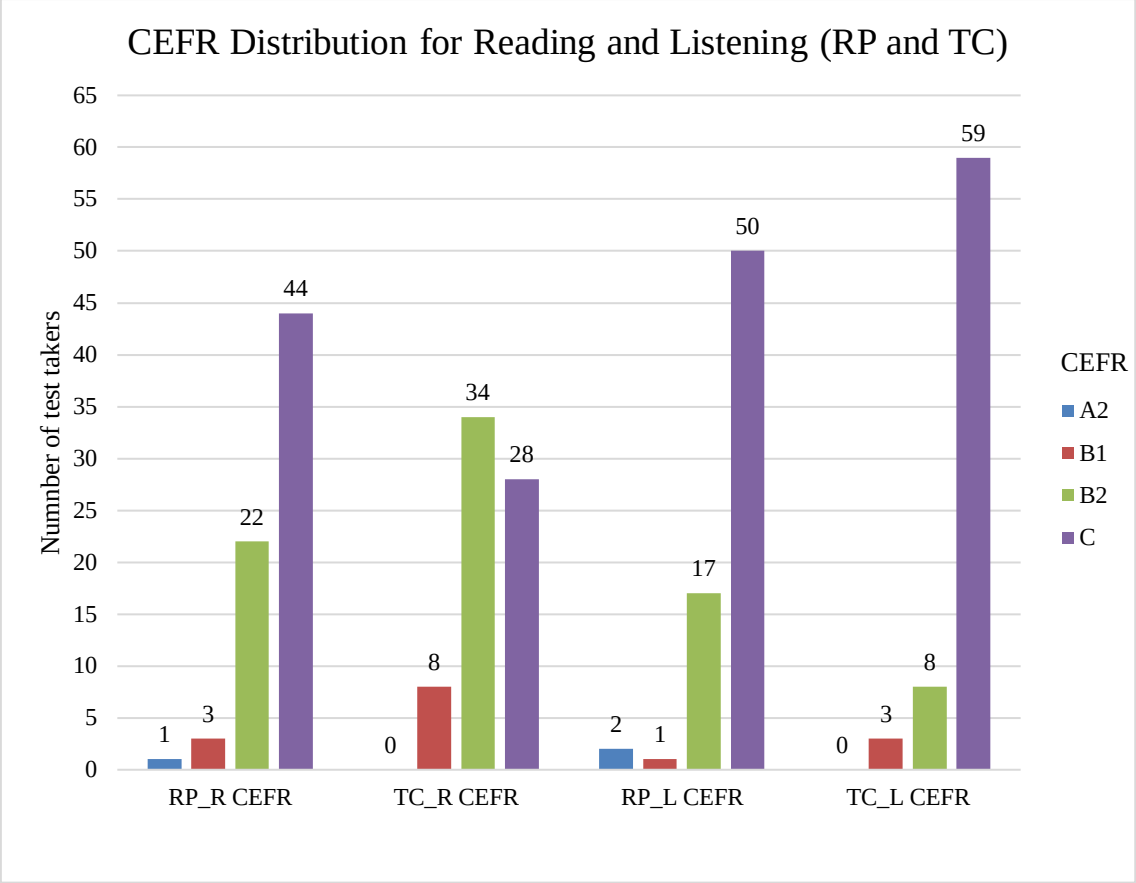

Figure 2. CEFR distribution for Reading and Listening for RP and TC

RP and TC differences

A paired-samples t-test with Bonferroni adjustments was conducted for the Reading_RP and Reading_TC and Listening_RP and Listening_TC pairs to evaluate the impact of the remote proctoring condition on the test-takers' scores. The test showed significantly lower mean scores for Reading_TC (M = 41.71, SD = 5.67) compared to Reading_RP (M = 43.86, SD = 5.76), t (69) = -3.42, p = .001 (two-tailed). The mean decrease in scores was 2.14 with a 95% confidence interval ranging from -.89 to -3.39. Cohen's d (.4) indicated a small effect size (Cohen, 1988).

There was no statistically significant difference in mean scores for Listening_TC (M = 41.89, SD = 5.39) compared to Listening_RP (M = 42.03, SD = 5.91), t (69) = -.27, p = .79 (two-tailed). The mean decrease in scores was .14 with a 95% confidence interval ranging from .90 to -1.19. Cohen's d (.000) indicated no effect size (Cohen, 1988).

A non-parametric significance test (Wilcoxon signed-ranked) was conducted for Grammar & Vocabulary because this component was not normally distributed. The

results revealed a statistically non-significant difference between the two delivery modes (RI and TC), z = -.49, p = .62. The median scores were the same (Mdn = 45).

Research Question 2. What are test-takers' perceptions of and attitudes towards remote proctored tests?

Questionnaire quantitative results

A reliability analysis showed good internal consistency for the twelve items used for the RP questionnaire (α=.82). The same analysis for the nine TC items revealed a lower internal consistency (α=.66), though this was not unexpected for a scale consisting of fewer than ten items (Pallant, 2016).

A Wilcoxon signed-rank test conducted between overlapping items from the RP and TC questionnaires revealed, in general, more positive attitudes towards, and better testing conditions for, TC (see Table 3). The differences in the two delivery modes were statistically significant for all items, with the highest difference in "frustrated" for RP (z = -3.61, p < .001), with a medium effect size (r = .31). Nevertheless, an overview of the results showed largely positive attitudes towards both delivery modes.

Additional examination of RP and TC questions showed that there was a slight preference towards RP (52.9%), even though most test-takers did not have previous remote proctoring experience (88.6%) and more participants needed technical assistance during RP (45.7%) than TC (12.9%). Most also reported positive practicalities such as having reliable internet, aptitude towards learning new technologies, and a home office. However, "good computer knowledge" (M = 3.46, SD = 1.00) and "no disruptions when working at home" (M = 3.40, SD = 1.22) had mixed reactions (see Appendix 5). Disruptions at home and a noisy environment could also be linked to the reported living conditions from the Demographics section of the questionnaire (as mentioned earlier, 61.5% lived with a spouse and children or family).

Table 3. Comparison between TC and RP delivery modes (overlapping questions in both questionnaires)

N = 70	N (%)					M (SD)	z	r
	5	**4**	**3**	**2**	**1**			
Quiet conditions_ TC	41 (58.6)	19 (27.1)	4 (5.7)	5 (7.1)	1 (1.4)	4.34 (.98)		
Quiet conditions_ RP	25 (35.7)	25 (35.7)	9 (12.9)	9 (12.9)	2 (2.9)	3.89 (1.12)	-2.56**	-.22
Confident_ TC	34 (48.6)	29 (41.4)	4 (5.7)	1 (1.4)	2 (2.9)	4.31 (.88)		
Confident_ RP	22 (31.4)	28 (40.0)	15 (21.4)	3 (4.3)	2 (2.9)	3.93 (.98)	-2.99***	-.25
Comfortable _TC	34 (48.6)	27 (38.6)	2 (2.9)	7 (10.0)	0	4.26 (.93)		
Comfortable _RP	24 (34.3)	25 (35.7)	12 (17.1)	6 (8.6)	3 (4.3)	3.87 (1.12)	-2.41*	-.20
Frustrated[a] _ TC	0	4 (5.7)	7 (10.0)	19 (27.1)	40 (57.1)	4.36 (.89)		
Frustrated_ RP	7 (10.0)	7 (10.0)	11 (15.7)	25 (35.7)	20 (28.6)	3.63 (1.28)	**-3.61*** **	**-.31**
Nervous[a]_ TC	3 (4.3)	15 (21.4)	14 (20.0)	16 (22.9)	22 (31.4)	3.56 (.89)		
Nervous_ RP	9 (12.9)	20 (28.6)	13 (18.9)	15 (21.4)	13 (18.6)	3.04 (1.33)	-2.87***	-.24

Notes. Bold denotes largest difference.
[a]Reverse coding for both "frustrated" and "nervous". Higher M denotes positive emotional state, i.e., less frustration and nervousness.
*p < .05, **p < .01, ***p < .001.

Questionnaire qualitative results

The questionnaires also contained open-ended items. These were short answers to the following questions:

(1) Please tell us why you preferred RP or TC.

(2) Please tell us what you liked about RP.

(3) Please tell us what you disliked about RP.

(4) How do you think the technical problem might have affected your performance? (this followed questions about identifying technical problems)

Please tell us why you preferred RP or TC

Those who preferred RI (52.9%) reported comfort (36.71%) and practicality (24.29%) as the two main positive determining factors for their choice. For TC preference, these were practicality (31.75%), support (25.40%), and comfort (23.81%) (see Table 4). The negative comments for RP were higher than TC (8.58% versus 6.36%), but in general were much less than the positive ones.

Table 4. Reasons for delivery mode preference

NVivo Codes	**N (%)**		
	RP	**TC**	**Total**
Negative Experiences			
Discomfort	2 (2.86)	1 (1.59)	3 (2.26)
Others	1 (1.43)	1 (1.59)	2 (1.50)
Remote proctoring	0	0	0
Support	**3 (4.29)**	**1 (1.59)**	**4 (3.01)**
Total (negative)	6 (8.58)	3 (6.36)	9 (6.77)
Positive Experiences			
Comfort	**25 (36.71)**	**15 (23.81)**	**40 (30.08)**
Others	8 (11.43)	7 (11.11)	15 (11.28)
Platform	1 (1.43)	1 (1.59)	2 (1.50)
Practicality	**17 (24.29)**	**20 (31.75)**	**37 (27.82)**
Privacy	2 (2.86)	0	2 (1.50)
Remote proctoring	3 (4.29)	0	3 (2.26)
Safety	2 (2.86)	0	2 (1.50)
Support	**4 (5.71)**	**16 (25.40)**	**20 (15.04)**
General technology	1 (1.43)	1 (1.59)	2 (1.50)
Test interface	1 (1.43)	0	1 (0.77)
Total (positive)	64 (91.42)	60 (93.64)	124 (93.23)

Notes. Bold indicates highest values.

Many of the responses to this question elicited not only comments about the test preference, but also reasons for not choosing the other delivery mode. As an example of how the coding scheme was implemented, consider the quotes in Table 5. Each participant's entry was coded for as many concepts as the participant expressed.

Table 5. Participants' quotes and coding comments

Quotes (preference)	Coding comments
#29 The truth is that I felt very nervous because I was afraid that there would be a problem, that they would knock at home, or that my dog would bark, personally I have a daughter at home and I had to be very careful that she did not come near the room while I was doing my exam. I prefer to go to the place in case there is any inconvenience and above all because I have the peace of mind that nothing can interrupt me. (TC)	Even though the participant mentions discomfort for RP at the beginning, this entry was coded for comfort for TC as the choice was TC and there is a clear mention at the end of the passage.
#32 There were more distractions at the test centre such as outside noise and colleagues coming in and out. (RP)	This is an example where the comments can be reverse-coded: the discomfort of TC is coded as comfort for RP.
#47 Because one does not worry about the equipment and infrastructure that centres generally already have as well as the personalised assistance at hand. (TC)	Clear and unambiguous example of practicality and support for TC.

Please tell us what you liked/disliked about RP

The answers to the questions about likes and dislikes were more straightforward. The analysis revealed that comfort, practicality, test interface, and others (test related, communication with test administrators, equipment, being part of innovation, validity/security perceptions) accounted for almost two-thirds of all positive comments (61.61%), while onboarding process, platform, others (ran out of time, test related, comms with admin, equipment, validity/security perceptions), and remote proctoring made up the same quantity in negative comments (60.25%) (see Table 6).

Table 6. Positive and negative comments about RP

NVivo Codes	***N* (%)**		***f* (%)**
Positive comments		Negative comments	
General	10 (8.93)	General	2 (2.56)
Comfort	**16 (14.29)**	Discomfort	9 (11.54)
Privacy	1 (.89)	Lack of technological experience	4 (5.13)
Onboarding process	4 (3.57)	**Onboarding process**	**11 (14.10)**
Safety	2 (1.79)	Onboarding anxiety	7 (8.97)
Others	**19 (16.96)**	**Others**	**15 (19.23)**
Platform	12 (10.71)	**Platform**	**11 (14.10)**
Practicality	**18 (16.07)**	Tech failure anxiety	3 (3.85)
Remote proctoring	2 (1.79)	**Remote proctoring**	**10 (12.82)**
Support	11 (9.82)	Support	3 (3.85)
General technical issues	1 (.89)	General technical issues	1 (1.28)
Test interface	**16 (14.29)**	Test interface	2 (2.56)
Total	112		78

Notes. Bold indicates highest values.

Many test-takers reported unclear communication and confusion about the free access keys they were given to start each test component. On the day of the RP test, they had problems locating them in the emails and using the right ones to start the test. These issues, when clearly mentioned by the test-takers, were coded under other categories as they were not specific to the RP protocols, platform, or technical issues. As the research was focusing on remote proctoring experiences, the NVivo coding tried to extract specific comments related to the platform, provider, and overall experience, including anxiety and technical issues. Most comments under others were related to pre-test communication and the test in general (e.g., scheduling, good test, etc.). Coding samples are presented in Table 7.

Table 7. Participants' quotes about RP (dis)likes and coding comments

Quotes(dis/like)	Coding comments
#13 The questions are easy. (like)	This was coded for others as it relates to the test itself.
#19 The convenience of doing it at home. Clear instructions and easy to use software. (like)	The codes were comfort and platform.
#68 The number of e-mails. (dislike)	An example of others (comms with admin).
#1 I felt quite pressured as to what to do if something went wrong or if I had any doubts. I wish they would have let me wear headphones because it is very noisy here at home. The rules that I had to keep my eyes straight ahead and that I could be disqualified caused me tension. It made me nervous to know that we had to do a scan of the room because it was not clear to me when it would be done. I appreciated that they sent several e-mails reminding me of the dates of the exam; however, that itself caused me some confusion when trying to remember where the keys were. (dislike)	An example of tech failure anxiety, discomfort, remote proctoring, and others (comms with admin)(negative due to confusion caused).

Even though pre-test communication and access codes were not part of the research focus, a close examination was performed as the questionnaires revealed gaps. Most of the communication was done in English, sometimes there was confusing information, and it was easy for some test-takers to lose track of the email communications. In the interviews that followed, test-takers shared views about the strong language in the email communications, which exhibited mistrust in the candidates (related to cheating and test invalidation).

An attempt was made to distinguish between feelings of anxiety about technology and any technical failures, and actual technical issues and disruptions during the test. As the coding results show (Table 6), anxiety was reported in 12.82% of comments (both onboarding and technical failure) while actual technical issues were only 1.28%. Onboarding process comments were related to the length and instructions; actual technical issues with it were coded under technical issues.

It is also worth mentioning that discomfort at home (11.54%) was also mentioned in the interviews. This combined with the quantitative results leads to the conclusion that people who lived alone (14.3% from the research pool) had less anxiety and disruptions.

How do you think the technical problem might have affected your performance?

There were 32 participants who reported technical issues (some had more than one problem) during RP and 9 during TC. The participants were asked to describe the nature of the problem and how it might have affected their performance. Results show that 'Others' (mostly issues with the access keys as mentioned above) accounted for 54.85% of all comments (N = 39) and the platform was a problem in 25.64% (see Table 8). It was interesting to observe that most reported issues did not seem to have a perceived negative impact on the test-takers' performance. For TC, six participants reported no perceived negative impact (3 said yes, and 1was unclear).

Table 8. Technical problems reported for RP and perceived negative impact on test performance

Code	N [a] (%)	Performance impact		
		Yes	No	Unclear
		N (%)		
Others	**21 (54.85)**	**2 (9.52)**	**12 (57.14)**	**7 (33.33)**
Platform	**10 (25.64)**	**2 (20.0)**	**6 (60.0)**	**2 (20.0)**
Remote proctoring	1 (2.56)	0	0	1 (100.0)
Support	2 (5.13)	1 (50.0)	0	1 (50.0)
General technical issues	3 (7.69)	0	1 (33.33)	2 (66.67)
Test interface	2 (5.13)	0	1 (50.0)	1 (50.0)

Note. [a] Some reported more than one problem.
N = 39.

The coding for perceived negative impact was conservative: if the participant did not specify explicitly yes or no or was confused, the response was coded unclear. Examples of coding are presented in Table 9.

Even though a Pearson correlation analysis through SPSS confirmed no significant correlation between technical problems and test preference ($r = -.06$, $p = .61$), the data was examined in more detail, looking at the participants who reported technical problems during RP (N = 32). Most issues were coded as others (N = 12) and most of the participants preferred RP (N = 18) (see Table 10. I also examined these participants' positive and negative comments (see Table 11).

Table 9. Participants' quotes about tech problems during RP and perceived negative impact

Quotes		Coding comments
Problem	**Perceived neg impact**	
#10: The first problem was that I couldn't enter the platform and then I had problems entering the keys. I thought I needed an extra code as a student but I didn't. I just had to insert the code for each exam. It almost took me 45 minutes to be able to start the exam.	I was nervous and I feel that I answered it too quickly because I was so nervous, I thought I would not have time.	This was coded for platform and others as it lists them separately. The impact was unclear as it was not explicitly stated ("feeling nervous" does not explicitly mean negative impact on performance).
#13: Locate the icon to start the test and enter the keys.	Not at all	Clear coding for platform and no.
#18: I couldn't access the exam and it took me an hour to solve it.	Nerves and desperation to start and finish solving the assessments. I think I was fed up and discouraged to continue. I might not do so well in my results because of that.	Coding for platform and yes as it clearly states not only feelings, but also perceived negative impact.

The results which emerged suggested:

(1) Comments from those who preferred RP and reported technical issues were coded mostly as others (60%; access key issues) and most were unclear or no about any negative impact. Comments from test-takers who preferred TC were almost split between others (47%; access keys issues) and platform (37%) as technical issues and most were coded for yes or unclear.

(2) More positive comments (related to Q2 and Q3) were coded for RP preference (26 versus 13), most of which were about comfort, practicality, others, and support, while platform and support dominated in TC preference.

(3) Finally, the total negative comments for each preference were almost equal (RP 17; TC 18), but they differed in choice: platform and others dominated RP, while for TC they were remote proctoring, others, and discomfort.

This analysis falls under the qualitative side of the study and no statistical analyses were conducted to investigate the significance of the relationships. However, the data suggests a need for further research and tentative observations.

Table 10. Reported technical issues and perceived impact

NVivo Codes	**Test Preference**	
	RP (*N* = 18)	**TC (*N* = 14)**
	F	
Issue		
Platform	3	7
Others	**12**	**9**
Test interface	2	0
General technical issues	2	1
Support	1	1
RP	0	1
Total	20	19
Impact		
Yes	0	4
No	8	**5**
Unclear	**10**	**5**

Table 11. Reported technical issues and positive/negative comments across test preference

	Test Preference	
	RP (*N* = 18)	**TC (*N* = 14)**
	f	
Positive comments		
Platform	0	**4**
Others	**5**	2
Remote proctoring	1	0
Support	**5**	**3**
Onboarding process	0	1
Practicality	**5**	2
Comfort	**5**	0
Safety	2	0
Test interface	3	1
Total	26	13
Negative comments		
Platform	**5**	3
Others	**5**	**4**
Remote proctoring	1	**4**
Onboarding process	2	1
Onboarding anxiety	1	0
Technical failure anxiety	0	1
Lack of tech exp	1	1
Discomfort	2	**4**
Test interface	0	0
Total	17	18

Semi-structured interview sessions results

Results from the data revealed that the interview participants were between 30 and 53 years old, had been teaching at BUAP from 3 to 18 years, and were equally distributed in terms of test delivery preference (interviewees in session 1 preferred RP while those in session 2 preferred TC). Even though the interviewees represented only 9% of all test takers, the analysis of their responses revealed further insights into the test-takers' RP experiences and general views on remote proctoring. For their positive experiences, they highlighted the flexibility and comfort, appreciated the convenience of taking the test at home (especially during COVID-19 times), the ability to use their personal equipment, and the opportunity to get familiar with the systems during the onboarding process. For some, this was also combined with their confidence in the remote proctoring systems and support. One participant said: "As far as the other advantage is concentration. Obviously by choosing the space, so I didn't have any distractions or other people's nerves."

Some had felt nervous during their RP test, but in general had liked the idea of the online assessment and the flexibility. One of them commented:

I feel that I am a bit nervous because of connectivity issues. Normally the internet that I have around here is not very good. So I had that kind of, well, I thought if it's really going to work for me at that precise moment.

Information scarcity and insufficient understanding of AI remote proctoring algorithms contributed to the negative sides of their experience, together with anxiety about technical failure, the long onboarding process and anxiety about technical requirements. Some test-takers gave examples of needing to go to the toilet and the urge to wave their hands at mosquitoes but did neither in fear of invalidating the test: they did not know if these were considered acceptable movements and gestures by the system. Most kept staring at the screen due to lack of information about the effect of eye movements on test invalidation. Some people also expressed anxiety about their living conditions (e.g., a noisy outside environment; living with small children who could burst in the room) while others viewed their homes in a positive way (e.g., live alone and thus have no distractions; can control the environment).

Their views on AI expressed mistrust and lack of knowledge about the RP systems and algorithms. They discussed the use of AI, and their preference for human proctoring was dominated by their own lack of knowledge, anxiety, and doubts about AI's ability to execute human-like decisions. Two interviewees commented: "I think that artificial intelligence does not yet have the ability to identify all the gestures that a person can make during an exam and separate these gestures that are natural from those that are cheating"; and "I completely agree that it is better for the human being to do the monitoring, because there are factors that in a certain way only humans can say what it feels like to take an oral exam, for example."

Their recommendations were for more and improved information about how the online learning protocols work in their assessment context and what kinds of body movement cause the AI algorithm to generate a red flag. By having this information, their levels of anxiety would be sufficiently reduced. Two of them said: "So yes, obviously I don't think that in five years' time, but in two years' time, that this type of application will be fully considered as a higher priority"; and

I think it would actually be very important that the people who are in charge of creating this test and being able to facilitate it in this remote or face-to-face way, is that they realise that it would be good for them to take the test as well, to be more aware of what we are going through.

Research question 3. Is there a significant effect of proctoring mode and test preference on test-takers' performance?

This research question was formulated to further explore the data and understand the results. A mixed between-within subjects analysis of variance (a mixed two-way ANOVA) was conducted as the distribution in each test preference group was fairly balanced. The analyses assessed the effect of the delivery mode and test preference on the participants' scores and examined any interaction effects (Laerd, 2021; Pallant, 2016).

Reading scores

There was homogeneity of variances ($p > .05$) and covariance ($p > .05$), as assessed respectively by Levene's test ($p = .59$ and $.93$) and Box's M test ($p = .68$) (see Table 12).

Maulchy's test of sphericity was not considered as the factor was only 2 (Pallant, 2016). There was no statistically significant interaction between test preference and delivery mode: Wilk's Lambda = 1.00, F (1, 68) = .001, p = .97, Cohen's d = .000. However, there was a statistically significant main effect for delivery mode: Wilk's Lambda = .86, F (1,68) = 11.46, p < .001, Cohen's d = .03 (small effect). Both groups showed lower scores for TC delivery mode (see Table 13). The main effect for test preference was not significant, F (1, 68) = .05, p = .82, Cohen's d = .000, suggesting no difference in scores between the two groups (Laerd, 2021; Pallant, 2016).

The analyses presented in Tables 12 and 13 indicate no difference in scores between the two groups for Reading.

Table 12. ANOVA interpretation READING

Interpretations	**Measures**	**Results**	**Values**
Homogeneity of variances[a]	Levene's test	Yes	p = .59 and .93
Covariance[a]	Box's M test	Yes	p = .68
Interaction between test preference and delivery mode[b]	Wilk's Lambda	No	1.00, F (1, 68) = .001, p = .97
Main effect for delivery mode[c]	Wilk's Lambda	Yes	.86, F (1,68) = 11.46, p < .001
Main effect for test preference[d]		No	F (1, 68) = .05, p = .82

Notes. [a] p > .05.
[b] Cohen's d = .000.
[c] Cohen's $_d$ = .03 (small to medium effect).
[d] Cohen's d = .000.

Table 13. Reading scores for test preference across the two delivery modes

Delivery Mode	**Test Preference**	
	RP[a]	**TC[b]**
	M (SD)	***M (SD)***
RP	44.00 (5.38)	43.70 (6.25)
TC	41.84 (5.93)	41.58 (5.45)

Notes. [a] N = 37.
[b] N = 33.

Listening scores

There was homogeneity of variances ($p > .05$) as assessed respectively by Levene's test ($p = .75$ and $.83$) and a violation of covariance ($p = .05$), as assessed by Box's M test. Maulchy's test of sphericity was not considered as the factor was only 2 (see Table 14). There was no statistically significant interaction between test preference and delivery mode: Wilk's Lambda = .99, F (1, 68) = .70, p = .41, Cohen's d = .000. In addition, there was no significant main effect for delivery mode: Wilk's Lambda = 1.00, F (1,68) = .10, p = .75, Cohen's d = .000. Both groups showed slightly different scores for the TC delivery mode: an increase in the RP test preference group; a decrease in the TC preference group (see Table 15). The main effect for test preference was not significant, F (1, 68) = .58, p = .45, Cohen's d = .000, suggesting no difference in the scores between the groups (Laerd, 2021; Pallant, 2016).

The analyses from Tables 14 and 15 indicate no difference in scores between the two groups for Listening.

Table 14. ANOVA interpretation LISTENING

Interpretations	Measures	Results	Values
Homogeneity of variances[a]	Levene's test	Yes	p = .75 and .83
Covariance	Box's M test	No	p = .05
Interaction between test preference and delivery mode[b]	Wilk's Lambda	No	.99, F (1, 68) = .70, p = .41
Main effect for delivery mode[c]	Wilk's Lambda	No	1.00, F (1,68) = .10, p = .75
Main effect for test preference[d]		No	F (1, 68) = .58, p = .45

Notes. [a] $p > .05$.
[b, c, d] Cohen's d = .000.

Table 15. Listening scores for test preference across the two delivery modes

Delivery Mode	**Test Preference**	
	RP[a]	**TC[b]**
	M (SD)	**M (SD)**
RP	42.27 (6.15)	41.76 (5.72)
TC	42.54 (5.39)	41.15 (5.39)

Notes. [a] N = 37.
[b] N = 33.

Grammar & vocabulary scores

There was homogeneity of variances ($p > .05$), as assessed respectively by Levene's test, and a violation of covariance ($p < .01$), as assessed by Box's M test. Maulchy's test of sphericity was not considered as the factor was only 2 (see Table 16). There was no statistically significant interaction between test preference and delivery mode: Wilk's Lambda = .95, F (1, 68) = 3.51, p = .65, Cohen's d = .000. In addition, there was no significant main effect for delivery mode: Wilk's Lambda = 1.00, F (1,68) = .008, p = .93, Cohen's d = .000. Both groups showed slightly different scores for the TC delivery mode: a decrease in the RP test preference group; an increase in the TC preference group (see Table 17). The main effect for test preferences was not significant, F (1, 68) = .64, p = .43, Cohen's d = .000, suggesting no difference in the scores between the groups (Laerd, 2021; Pallant, 2016).

Table 16. ANOVA interpretation GRAMMAR & VOCABULARY

Interpretations	Measures	Results	Values
Homogeneity of variances[a]	Levene's test	Yes	p = .20 and .17
Covariance	Box's M test	No	p = .003
Interaction between test preference and delivery mode[b]	Wilk's Lambda	No	.95, F (1, 68) = 3.51, p = .65
Main effect for delivery mode[c]	Wilk's Lambda	No	1.00, F (1,68) = .008, p = .93
Main effect for test preference[d]		No	F (1, 68) = .64, p = .43

Note. [a] p > .05
[b, c, d] Cohen's d = .000.

Table 17. Grammar & Vocabulary scores for test preference across the two delivery modes

Delivery Mode	Test Preference	
	RP[a]	TC[b]
	M (SD)	M (SD)
RP	45.22 (2.46)	44.18 (4.30)
TC	44.78 (2.20)	44.58 (4.22)

Note. [a] N = 37.
[b] N = 33.

The analyses from Tables 16 and 17 indicate no difference in scores between the two groups for Grammar & Vocabulary.

The only statistically significant main effect for delivery mode with a small to medium effect was observed for Reading. The main effect for test preferences was not significant for either component, suggesting no difference in scores between the two groups. Certain assumptions were violated in the analyses for Listening and Grammar & Vocabulary, so more detailed analyses are needed to further explore the interactions.

Discussion and conclusion

Research question 1

Overall, the performance results of the research align with, to an extent, concordance studies discussed in the Literature Review section (Karim et al., 2014; Weiner & Hurtz, 2017; Woldeab & Brothen, 2019; Zumbo, 2021). No significant differences were found in Listening and Grammar & Vocabulary and the delivery mode did not affect test-takers' scores. Their performance in Reading was higher for RP, with a large significant difference (a mean increase of 2.14 points from TC). For this skill, RP had a positive impact on the scores.

It was interesting to observe these diverging results for Reading. In Reading RP, test-takers who achieved the highest level C on the CEFR scale were double than those who got B2 (44 vs 22). For Reading TC, the results were the reverse: more achieved B2 than C, with less of a difference (34 vs 28). Certain assumptions can be made after analysing the survey results about possible reasons, one of them being comfort at home: no distractions, a quiet environment, and less anxiety.

One possible explanation that could impact the interpretation of these results is the order effect of the components during both delivery modes: first, Grammar & Vocabulary, second, Listening, and Reading at the end. This could not be controlled as it is part of the test structure and is administered by the Aptis team. It could be argued that for RP, by the time the test-takers logged in for their reading tasks, they had located the access keys, had become familiar with the platform, AI and proctoring systems, and their anxiety levels had been reduced. This could have led to the higher scores observed in the remote session. However, the results showed that highest scores were achieved for Grammar & Vocabulary, which was the first component. Another supposition could be that the unfamiliarity of the test centre environment was detrimental to them demonstrating their ability. Further research might clarify these seemingly opposite explanations.

Research question 2

The results from the survey (questionnaires and interviews) demonstrated that test-takers felt slightly more comfortable and confident, and less frustrated and nervous, during the test at the test centre than at home. They expressed positive views about remote proctoring related to home comfort, practicality, and safety during the pandemic. However, they disliked the onboarding process for its length and technical requirements; the platform (even though this could have been partially related to access keys and pre-test communications with the British Council team); and remote proctoring for its unclear AI protocols about red flags, test violations, and lack of a human touch.

Further insights revealed mistrust towards AI which, for some, was based on lack of understanding and information about the specific algorithms and systems. This contributed to test-takers' general anxiety in combination with anxiety from technical failure. The human technical support provided during the RP session was comforting and efficient to some, but others experienced delays and confusion which also contributed to their anxiety.

Test-takers' negative experiences and the idea of mistrust link back to Karim et al.'s (2014) and Dawson's (2020) views earlier and their suggestions about a more positive, awareness-raising approach which will lead to creating trust and better experiences

for the test-takers. This could be achieved through a more flexible attitude towards online learning protocols and AI algorithms.

Pre-test communication seemed to represent a big portion of the anxiety, confusion, and platform access problems on the day of RP session. It raises a question about the possibility of a more positive test-takers' overall experience had this issue been eliminated. Even though technical problems were reported by 32 participants, most did not think these affected their performance negatively and still preferred RP (again, most technical problems were related to pre-test communications and access keys).

The qualitative data provided support for tentative conclusions: the affective factor of anxiety could be reduced through the convenience and comfort of taking the exam at home. However, it could be argued that for some people the home environment is not ideal and could potentially add to the anxiety the test-takers are experiencing.

On balance, the lack of knowledge and anxiety about the red flags and test invalidations due to AI's inability to make decisions like a human were problematic for the test-takers. This reinforces Dawson's (2020) multi-layered approach: the tension between test reliability and solutions to assessment security needs which are innovative, progressive, and flexible. AI and the proctoring systems should not be singled out and portrayed negatively, but examples of societal monitoring such as "the balance between policing and crime prevention" could be used to develop better understanding and acceptance (p. 142).

For the participants in the study, privacy, confidentiality, accessibility, and technological bias did not appear to be major concerns, as found in research by Langenfeld (2020), Selwyn et al. (2021), and Swauger (2020). However, safeguarding and privacy still merit more research, especially in environments which are unsafe or where privacy and confidentiality systems are not as robust.

While cheating and assessment security were not in the scope of the study, some test-takers expressed positive views towards the remote proctoring system and noted that they had much better awareness of what their students have to go through while doing a remotely proctored exam. Others remarked on the strong view taken by Aptis and the RP provider about detecting cheating patterns and the threat of test invalidation.

Research question 3

Combining and extrapolating the different data sources to validate assumptions revealed a statistically significant effect of delivery mode on test-takers' scores for Reading (lower for TC) when their test preference was considered. However, no main effects for Grammar & Vocabulary and Listening, and no interaction effects between test preference and delivery mode for any of the three components, were observed. This suggests that test preference did not make a difference in the scores and there was no interaction between the test choice and delivery mode. This result suggests that preferences might not be sufficiently strong to influence test performance. However, the test-takers in the research were a high performing group and, even though some exhibited anxiety, it did not impact on their performance as they were able to follow the RP procedures. Questions about lower-performing test-takers could be further investigated to see how the results from the current research might differ.

Overall, even though some test-takers experienced technical problems and anxiety, reported mistrust in AI algorithms and RP protocols, and found certain proctor-system specific processes cumbersome (onboarding) during their Aptis RP session, most still preferred the remote session and believed that the advantages outweighed the disadvantages.

Implications

There are several implications from the study for the language testing community. First, test developers have to design remotely proctored tests with the test-takers' experience at the core of validity and reliability considerations. Remote proctoring offers new solutions which need to be further explored, and flexibility in applying these solutions will lead to increased positive perceptions and experiences. Test providers need to ensure timely, precise, and unambiguous pre-test communication with the test-takers, which will solve many of the problems they experience and reduce their anxiety and confusion.

Secondly, developers of online testing protocols need to improve their systems continuously so that test-takers feel properly supported during the test and know exactly what to expect from the AI detection protocols. Information sharing about dos and don'ts related to gestures and movements, toilet breaks and drinking water during

a remotely proctored test should be clearly indicated. Online proctoring providers have a responsibility and should work collaboratively with test providers to ensure improved flexible options for test-takers and good systems which can correct any faulty AI decisions. Since test-takers are mostly on their own in the RP testing environment, seamless and unambiguous procedures encompassing pre-test, during-the-test, and post-test requirements will greatly benefit everyone involved.

Thirdly, test-takers need the tools to familiarise themselves better with remote proctoring as they recognise the fact that these systems are here to stay. Their mistrust in the systems needs to be mitigated so that they can have positive experiences which will in turn increase their acceptance of remote proctoring. However, for many the impact from digital divide, imperfect home testing conditions, and lack of technical skills will have to be considered by test developers and remote proctoring providers in the high-stakes context.

Limitations

While robust in its nature, the study has several limitations. The context was relatively narrow, not only geographically (Mexico), but also in terms of test-takers who were university teachers. This led to a lack of range in the scores and a score distribution which was predominantly at the higher end of the CEFR, in keeping with initial expectations. The perfect test-taker group might have given insights into perceptions by lower-performing and younger test-takers, which might have shown other types of concerns. Secondly, exploring only receptive skills and Grammar & Vocabulary raises questions about productive skills, and especially Speaking which, when done in the comfort of one's own home environment, could potentially be more advantageous to test-takers.

Further research

More research is also needed on productive skills, especially Speaking to validate future remote proctoring findings. Some researchers have already provided evidence of positive test-takers' views on conducting Speaking tests remotely, where the test-takers were located at a test centre and proctored by humans on site while the examiner was administering the test remotely via a camera (Lee et al., 2021). Future

studies should also explore other contexts, environments, and test-taker representation to broaden the scope. It would also be valuable to conduct observational research at test-takers' homes to better understand variability in home conditions versus the traditional test centre. Additional research on RP providers and RP systems and algorithms will also be beneficial as improvements might alleviate test-takers' anxiety. Although the study did not directly investigate lower-proficiency test-takers, questions about providing levels of support for different test-taker profiles (including digital literacy) arise. Finally, the digital divide (access to reliable internet and devices), even though not in scope for this research, could provide further insight into test-takers' anxiety about online assessments (Moore et al., 2018).

Remote proctoring in the medium-stakes language assessment context is an area which merits ongoing and timely research, perhaps mostly concerning test-takers' experiences and perceptions. The same questions remain in the high-stakes context where security should be carefully considered and not tip the balance between validity and test-takers' experiences and concerns.

Acknowledgements

I would like to thank the late Dr Jamie Dunlea for his tremendous help and unwavering trust in this research and me. I would also like to thank Professor Luke Harding from Lancaster University for his support, supervision, and belief in the study and Dr Emma Bruce from the British Council for her expertise in helping me with this manuscript. Special thanks to Fiona Mitchell from the British Council and Andrew Mullooly from Cambridge University Press & Assessment who played an important role in the data collection and analysis. Last, but not least I would like to thank all colleagues at the British Council who were involved for their support.

Author disclosures

The author is employed by the British Council, the publisher of the Aptis tests, but has not been involved in the development or management of the tests.

The research was carried out while the author was a postgraduate student at Lancaster University and her studies were funded by a professional development grant from the

British Council. Apart from the fact that it was the location of the research, the author has no association with Benemérita Universidad Autónoma de Puebla.

ORCID iD

Radosveta B. Valkova https://orcid.org/0000-0003-0578-1945

References

Alessio, H. M., Malay, N. J., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning*, 21(1). https://doi.org/10.24059/olj.v21i1.885

Arnò, S., Galassi, A., Tommasi, M., Saggino, A., & Vittorini, P. (2021). State-of-the-art of commercial proctoring systems and their use in academic online exams. *International Journal of Distance Education Technologies*, 19(2), 55–76. https://doi.org/10.4018/IJDET.20210401.oa3

Atoum, Y., Chen, L., Liu, A. X., Hsu, S. D. H., & Liu, X. (2017). Automated online exam proctoring. *IEEE Transactions on Multimedia*, 19(7), 1609–1624. https://doi.org/10.1109/TMM.2017.2656064

British Council (2020). *Aptis candidate guide*. https://www.britishcouncil.org/sites/default/files/aptis_candidate_guide_2020_a4_final.pdf

British Council (2021). *Aptis Remote: Practice materials*. https://www.britishcouncil.org/exam/aptis/practice-materials/remote

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge Academic.

Cohen, L., Manion, L., & Morrison, K. (2017). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Council of Europe Language Policy Division, Strasbourg. (2018). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A manual*. https://www.coe.int/en/web/common-european-framework-reference-languages/relating-examinations-to-the-cefr

Dadashzadeh, M. (2020). The online examination dilemma: To proctor or not to proctor? *Journal of Instructional Pedagogies*, *25*, 1–11. https://files.eric.ed.gov/fulltext/EJ1294386.pdf

Dawson, P. (2020). *Defending assessment security in a digital world: Preventing e-cheating and supporting academic integrity in higher education*. Routledge. https://doi.org/10.4324/9780429324178

Domínguez, C., López-Cuadrado, J., Armendariz, A., Jaime, A., Heras, J., & Pérez, T. A. (2019). Exploring the differences between low-stakes proctored and unproctored language testing using an Internet-based application. *Computer Assisted Language Learning*, 32(5–6), 483–509. https://doi.org/10.1080/09588221.2018.1527360

Dunn, K. (2020). *Aptis scoring system* (Technical Report No. 2.0). British Council. https://www.britishcouncil.org/sites/default/files/aptis_scoring_system_v2.0.pdf

Foster, D., & Layman, H. (2013). *Online proctoring systems compared*. Retrieved from https://www.caveon.com/wp-content/uploads/2014/03/Online-Proctoring-Systems-Compared-Mar-13-2013.pdf

Hesse-Biber, S. N., & Johnson, R. B. (Eds.). (2015). *The Oxford handbook of multimethod and mixed methods research inquiry*. Oxford University Press.

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Karim, M. N., Kaminsky, S. E., & Behrend, T. S. (2014). Cheating, reactions, and performance in remotely proctored testing: An exploratory experimental study. *Journal of Business Psychology*, 29, 555–572. https://doi.org/10.1007/s10869-014-9343-z

Kiger, M. E., & VaRIio, L. (2020). Thematic analysis of qualitative data: AMEE Guide No. 131. *Medical Teacher*, 42(8), 846–854. https://doi.org/10.1080/0142159X.2020.1755030

Laerd Statistics (2021). SPSS. https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php

Langenfeld, T. (2020). Internet-based proctored assessment: Security and fairness issues. *Educational Measurement: Issues and Practice*, 39(3), 24–27. https://doi.org/10.1111/emip.12359

Lee, H., Patel, M., Lynch, J., and Galaczi, E. (2021). Development of the IELTS video call speaking test: Phase 4 operational research trial and overall summary of a four-phase test development cycle. *IELTS Partnership Research Papers*, 2021/1. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. https://www.ielts.org/teaching-and-research/research-reports

Moore, R., Vitale, D. , & Stawinoga, N. (2018). The digital divide and educational equity: A look at students with very limited access to electronic devices at home. ACT Center for Equity in Learning. https://www.act.org/content/dam/act/unsecured/documents/R1698-digital-divide-2018-08.pdf.

Michel, R. S. (2020). Remotely proctored K-12 high stakes standardized testing during COVID- 19: Will it last? *Educational Measurement*: Issues and Practice, 39(3), 28–30. https://pesquisa.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/covidwho-787760

Organisation for Economic Co-operation and Development (OECD) (2018). Distribution of teachers by age and gender. https://stats.oecd.org/Index.aspx?datasetcode=EAG_PERS_SHARE_AGE

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines. *International Journal of Qualitative Methods*, 19, 1–13. https://doi.org/10.1177/1609406919899220

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., & Dunn, K. (2020). Aptis General technical manual (Technical Report No. 2.2). British Council. https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Pallant, J. (2016). SPSS survival manual: A step by step guide to data analysis using IBM SPSS (6th ed.). *McGraw-Hill Education*. https://doi.org/10.4324/9781003117407

Prince, D. J., Fulton, R. A., & Garsombke, T. W. (2009). Comparisons of proctored versus non-proctored testing strategies in graduate distance education

curriculum. *Journal of College Teaching & Learning (TLC)*, 6(7), 51–62. https://doi.org/10.19030/tlc.v6i7.1125

Rios, J. A., & Liu, O. L. (2017). Online proctored versus unproctored low-stakes internet test administration: Is there differential test-taking behavior and performance? *American Journal of Distance Education*, 31(4), 226-241. https://doi.org/10.1080/08923647.2017.1258628

Roulston, K. (2018). Interviews in qualitative research. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). https://doi:10.1002/9781405198431.wbeal0572.pub2

Selwyn, N., O'Neill, C., Smith, G., Andrejevic, M., & Gu, X. (2021). A necessary evil? The rise of online exam proctoring in Australian universities. *Media International Australia*. 186(1), 149-164. https://doi.org/10.1177/1329878X211005862

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, 36(1), 174–184. https://doi.org/10.1027/1015-5759/a000494

Swauger, S. (2020). Software that monitors students during tests perpetuates inequality and violates their privacy. *MIT Technology Review*. www.technologyreview.com/2020/08/07/1006132/ software-algorithms-proctoring-online-tests-ai-ethics/

Weiner, J. A., & Hurtz, G. M. (2017). A comparative study of online remote proctored versus onsite proctored high-stakes exams. *Journal of Applied Testing Technology*, 18(1), 13–20. http://www.jattjournal.com/index.php/atp/article/view/113061

Woldeab, D., & Brothen, T. (2019). 21st century assessment: Online proctoring, test anxiety, and student performance. I*nternational Journal of E-Learning & Distance Education*, 34(1), 1–10. https://www.ijede.ca/index.php/jde/article/view/1106

Zumbo, B. D. (2021). A novel multimethod approach to investigate whether tests delivered at a test centre are concordant with those delivered remotely online: An investigation of the concordance of the CAEL [Research Monograph B]. https://doi.org/10.14288/1.0400581

Appendix 1. Group A complete questionnaire (Microsoft Forms version)

Aptis Test Perceptions (Aptis Remote, Group A, Day 1)

This is the first of two surveys which support a research project looking into the use of remote proctoring systems for Aptis and assessing any differences in test takers' performance, their perceptions and experience, and suggesting further solutions.

Thank you for your participation! Please, take your time to answer all questions carefully.

* Required

Demographics and Context

1. Please tell us your name: *

2. What is your gender? *

- 2.1 Woman
- 2.2 Man
- 2.3 Other
- 2.4 Prefer not to disclose

3. How old are you? *

4. What subject(s) do you teach at BUAP university? *

5. Which subjects do you teach in English? *

6. How many years have you been teaching at BUAP university? *

7. What is/are your first language(s)? *

8. Who do you live with (most of the time)? *

- 8.1 Live alone
- 8.2 Live with a partner/spouse
- 8.3 Live with a partner/spouse and child(ren)
- 8.4 Live with family (parents, siblings)
- 8.5 Live with flatmate(s)
- 8.6 Other

9. What teaching qualifications do you have? *

10. What is the highest English qualification/exam certificate level that you have? *

11. Have you taken remote proctored tests before? *

- 11.1 Yes
- 11.2 No

General Perceptions of Remote Proctoring

12. Please indicate your level of agreement (disagreement) with the following statements. Technology and Computers:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
12.1 My internet connection at home is reliable (most of the time)	○	○	○	○	○
12.2 I have good computer knowledge	○	○	○	○	○
12.3 It's easy for me to learn new technologies	○	○	○	○	○

13. Please indicate your level of agreement (disagreement) with the following statements. Home Environment:

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
13.1 My home environment is calm	○	○	○	○	○
13.2 I don't get disturbed when I work at home	○	○	○	○	○
13.3 I have my own room when I work at home	○	○	○	○	○

14. Please choose the best answer. Ways of Remote Proctoring:

- ○ 14.1 I prefer fully automated (only AI (Artificial Intelligence) invigilated)
- ○ 14.2 I prefer record & review (AI invigilated and Human reviewed after the test if there're red flags)
- ○ 14.3 I prefer live proctoring (a combination of both Human and AI real time invigilation)

Aptis Remote Test Perceptions

15. Which of these devices do you own at home? Choose all that apply.

- ○ 15.1 Laptop
- ○ 15.2 Desktop Computer
- ○ 15.3 Tablet
- ○ 15.4 None of the above

16. Which device did you use for the Aptis Remote test? *

17. Please indicate your level of agreement (disagreement) with the following statements. Aptis Remote: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
17.1 The remote proctoring affected my performance negatively	○	○	○	○	○
17.2 I liked the remote proctored test	○	○	○	○	○

18. Please indicate your level of agreement (disagreement) with the following statements. During the test I felt mostly: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
18.1 Nervous	○	○	○	○	○
18.2 Confident	○	○	○	○	○
18.3 Frustrated	○	○	○	○	○
18.4 Comfortable	○	○	○	○	○

19. Please tell us what you liked about Aptis Remote Proctored Test: *

20. Please tell us what you disliked about Aptis Remote Proctored Test: *

21. Did you need any technical assistance during the test? *

○ 21.1 Yes

○ 21.2 No

22. What was the problem and how long did it take to resolve it? *

23. How do you think the problem might have affected your performance? *

Aptis Test Perceptions (Test Centre, Group A, Day 2)

This is the second of two surveys which support a research project looking into the use of remote proctoring systems for Aptis and assessing any differences in test takers' performance, their perceptions and experience, and suggesting further solutions.

Thank you for your participation! Please, take your time to answer all questions carefully.

* Required

Aptis Test Perceptions

1. Please tell us your name: *

2. Please indicate your level of agreement (disagreement) with the following statements. Test Center Environment: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
2.1 The test centre environment was quiet	○	○	○	○	○
2.2 I felt calm knowing the staff would help me if I had problems	○	○	○	○	○

3. Please indicate your level of agreement (disagreement) with the following statements. Technical Aspects: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
3.1 I understood what I had to do	○	○	○	○	○
3.2 It was easy to navigate the test	○	○	○	○	○
3.3 I was comfortable with the technology	○	○	○	○	○

4. Please indicate your level of agreement (disagreement) with the following statements. During the test I felt mostly: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
4.1 Nervous	○	○	○	○	○
4.2 Confident	○	○	○	○	○
4.3 Frustrated	○	○	○	○	○
4.4 Comfortable	○	○	○	○	○

5. Did you need any technical assistance during the test? *

- ○ 5.1 Yes
- ○ 5.2 No

6. What was the problem and how long did it take to resolve it? *

7. How do you think the problem might have affected your performance? *

8. Which of the two tests did you prefer? *

- ○ 8.1 Aptis at Test Centre
- ○ 8.2 Aptis Remote (at home)

9. Please, tell us why (in reference to Q8). *

Appendix 2. Group B complete questionnaire (Microsoft Forms version)

Aptis Test Perceptions (Test Centre, Group B, Day 1)

This is the first of two surveys which support a research project looking into the use of remote proctoring systems for Aptis and assessing any differences in test takers' performance, their perceptions and experience, and suggesting further solutions.

Thank you for your participation! Please, take your time to answer all questions carefully.

* Required

Demographics and Context

1. Please tell us your name: *

2. What is your gender? *

- 2.1 Woman
- 2.2 Man
- 2.3 Other
- 2.4 Prefer not to disclose

3. How old are you? *

4. What subject(s) do you teach at BUAP university? *

5. Which subjects do you teach in English? *

6. How many years have you been teaching at BUAP university? *

7. What is/are your first language(s)? *

8. Who do you live with (most of the time)? *

- 8.1 Live alone
- 8.2 Live with a partner/spouse
- 8.3 Live with a partner/spouse and child(ren)
- 8.4 Live with family (parents, siblings)
- 8.5 Live with flatmate(s)
- 8.6 Other

9. What teaching qualifications do you have? *

10. What is the highest English qualification/exam certificate level that you have? *

Aptis Test Centre Perceptions

11. Please indicate your level of agreement (disagreement) with the following statements. Test Centre Environment: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
11.1 The test centre environment was quiet	○	○	○	○	○
11.2 I felt calm knowing the staff would help me if I had problems	○	○	○	○	○

12. Please indicate your level of agreement (disagreement) with the following statements. Technical Aspects: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
12.1 I understood what I had to do	○	○	○	○	○
12.2 It was easy to navigate the test	○	○	○	○	○
12.3 I was comfortable with the technology	○	○	○	○	○

13. Please indicate your level of agreement (disagreement) with the following statements. During the test I felt mostly: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
13.1 Nervous	○	○	○	○	○
13.2 Confident	○	○	○	○	○
13.3 Frustrated	○	○	○	○	○
13.4 Comfortable	○	○	○	○	○

14. Did you need any technical assistance during the test? *

- ○ 14.1 Yes
- ○ 14.2 No

15. What was the problem and how long did it take to resolve it? *

16. How do you think the problem might have affected your performance? *

Aptis Test Perceptions (Aptis Remote, Group B, Day 2)

This is the second of two surveys which support a research project looking into the use of remote proctoring systems for Aptis and assessing any differences in test takers' performance, their perceptions and experience, and suggesting further solutions.

Thank you for your participation! Please, take your time to answer all questions carefully.

* Required

General Perceptions of Remote Proctoring

1. Please tell us your name: *

2. Have you taken remote proctored tests before? *

- 2.1 Yes
- 2.2 No

3. Please indicate your level of agreement (disagreement) with the following statements:. Technology and Computers: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
3.1 My internet connection at home is reliable (most of the time)	○	○	○	○	○
3.2 I have good computer knowledge	○	○	○	○	○
3.3 It's easy for me to learn new technologies	○	○	○	○	○

4. Please indicate your level of agreement (disagreement) with the following statements. Home Environment: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
4.1 My home environment is calm	○	○	○	○	○
4.2 I don't get disturbed when I work at home	○	○	○	○	○
4.3 I have my own room when I work at home	○	○	○	○	○

5. Please choose the best answer. Ways of Remote Proctoring: *

- ○ 5.1 I prefer fully automated (only AI (Artificial Intelligence) invigilated)
- ○ 5.2 I prefer Record & Review (AI invigilated and Human reviewed after the test if there're red flags)
- ○ 5.3 I prefer live proctoring (a combination of both Human and AI in real time invigilation)

Aptis Remote Test Perceptions

6. Which of these devices do you own at home? Choose all that apply. *

- [ ] 6.1 Laptop
- [ ] 6.2 Desktop Computer
- [ ] 6.3 Tablet
- [ ] 6.4 None of the above

7. Which device did you use for the Aptis Remote test? *

8. Please indicate your level of agreement (disagreement) with the following statements. Aptis Remote: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
8.1 The remote proctoring affected my performance negatively	○	○	○	○	○
8.2 I liked the remote proctored test	○	○	○	○	○

9. Please indicate your level of agreement (disagreement) with the following statements. During the test I felt mostly: *

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
9.1 Nervous	○	○	○	○	○
9.2 Confident	○	○	○	○	○
9.3 Frustrated	○	○	○	○	○
9.4 Comfortable	○	○	○	○	○

10. Please tell us what you liked about Aptis Remote Proctored Test: *

11. Please tell us what you disliked about Aptis Remote Proctored Test: *

12. Did you need any technical assistance during the test? *

- ○ 12.1 Yes
- ○ 12.2 No

13. What was the problem and how long did it take to resolve it? *

14. How do you think the problem might have affected your performance? *

15. Which of the two tests did you prefer? *

- 15.1 Aptis at Test Centre
- 15.2 Aptis Remote (at home)

16. Please, tell us why (in reference to Q15). *

Appendix 3. Final list of NVivo codes used in questionnaires

NVivo Codes	Description
Negative Experience	
Discomfort	Distractions (people), environment noises (indoor/outdoor), COVID-19 (masks/face shields)
Lack of tech experience	
Onboarding anxiety	Technical requirements
Onboarding process	Length, instructions
Others	Test related (e.g., scheduling), comms with admin, access keys, ran out of time, equipment, validity/security perceptions
Platform	Accessibility, log in
Remote proctoring	Anxiety, RP protocols (eyes on screen, room scan, cheating detection, verification)
Support	
Technical failure anxiety	Computer failure, connection drop
Technical issues	Computer failure, connection drop
Test interface	Navigation, audios
Positive Experience	
Comfort	Lack of distractions, min noise, better concentration, calmness, environment, COVID-19 (masks/face shields)
Onboarding process	
Others	Test related, (e.g., scheduling), comms with admin, equipment, being part of innovation, validity/security perceptions
Platform	Accessibility, log in
Practicality	RP: save time (no travelling, no waiting at TC), convenience, managing own time; TC: device and connection control/check, infrastructure, convenience
Privacy	
Remote proctoring	
Safety	
Support	
Technology	Good connection
Test interface	Navigation, audios, ease

Appendix 4. Interview protocol

Welcome! My name is [name] and I'm your facilitator for this interview group. This is [name]: the researcher in this study.

This interview group will focus on questions related to the questionnaires you completed after each test. Please answer and share only information that you are comfortable sharing. Your confidentiality is our priority – your real names will not be used at **any time** during this research and will not be made public. To keep your anonymity, I'd like to propose the following: [name] you'll be *Participant 1*, [name 2] you'll be *Participant 2*, and [name 3] you'll be *Participant 3*. Is that fine with everyone? I'll write this in the chat now. Please do not use real names during the meeting.

Alright, do you have any questions before we start? I will start the recording now.

Q1. Can you tell us a bit more about your RP experience?

Q2. What are the two biggest advantages and the two biggest disadvantages of remote proctoring?

Q3. Do you think it's easier to use remote invigilation for testing certain subjects (e.g., languages) than others?

Q4. There was a question in the survey about Artificial Intelligence and Human invigilation. Why do you think most people might prefer to have a human invigilation in real time with the help of AI instead of only AI, for example?

Follow-up Q (Group 2 only)

If you knew there was immediate technical help when you needed it, would that change your opinion?

Q5. Do you think remote invigilation will become the norm in language assessment in 5 years?

Q6. How would you compare both test experiences: TC and RP? Is there anything else you'd like to add?

Follow-up Q (both Groups)

> Would it be helpful if we gave more information about the proctoring system before you do the test to see exactly how it works? The process about the video system, recording, and the red flags specifically.

This is the end of the interview. Thank you very much for your answers. Your opinions and ideas are extremely valuable, and we appreciate your participation. I hope that the discussion has been interesting for you as well. Please contact [name] the researcher if you have any additional questions. As we said at the beginning, all your comments will be anonymised and kept strictly confidential. Thank you very much. Goodbye.

Appendix 5. Results from additional questions (RP and TC questionnaires)

N = 70	*f*n(%)					*M* (*SD*)
	5	4	3	2	1	
Reliable internet_RP	27 (38.6)	28 (40.0)	6 (8.6)	8 (11.4)	1 (1.4)	4.03 (1.04)
Good computer knowledge_ RP	7 (10.0)	34 (48.6)	16 (22.9)	10 (14.3)	3 (4.3)	3.46 (1.00)
Easy to learn new tech_TC	25 (35.7)	31 (44.3)	10 (14.3)	3 (4.3)	1 (1.4)	4.09 (.90)
Liked_RP	25 (35.7)	28 (40.0)	10 (14.3)	6 (8.6)	1 (1.4)	4.00 (.99)
Home office_RP	28 (40.0)	23 (32.9)	8 (11.4)	7 (10.0)	4 (5.7)	3.91 (1.20)
No disruptions when working from home_RP	14 (20.0)	25 (35.7)	10 (14.3)	17 (24.3)	4 (5.7)	3.40 (1.22)
Feeling calm knowing there's staff to help_ TC	60 (85.7)	9 (12.9)	1 (1.4)	0	0	4.84 (.40)
Understood what I had to do_ TC	66 (94.3)	4 (5.7)	0	0	0	4.94 (.23)
Easy test navigation_ TC	67 (95.7)	3 (4.3)	0	0	0	4.96 (.20)
Comfortable with tech_TC	62 (88.6)	6 (8.6)	2 (2.9)	0	0	4.86 (.43)