**Journal of Language Testing & Assessment (2023)**
**Clausius Scientific Press, Canada**

**DOI: 10.23977/langta.2023.060101**
**ISSN 2616-2938 Vol. 6 Num. 1**

# *The Delivery of Speaking Tests in Traditional or Online Proctored Mode: A Comparability Study*

**Michael Milanovic[a], Tony Lee[b], David Coniam[c,*]**

*Language Cert, London, UK*
*[a]Michael.Milanovic@PeopleCert.org, [b]Tony.Lee@PeopleCert.org,*
*[c]David.Coniam@PeopleCert.org*
**Corresponding author*

***Keywords:*** Test score comparability, English language, Speaking tests, CEFR, online proctoring

***Abstract:*** This paper investigates the comparability of test scores recorded for high-stakes English language Speaking Tests administered face-to-face in either a traditional centre-based mode (TM) or in an online proctored mode (OLP). The data comprise a large sample of test takers taking English language Speaking Tests at four CEFR (the 'Common European Framework of Reference for Languages') levels – B1 to C2 – via TM or OLP. The data were analysed using descriptive statistics, effect size differences and equivalence tests. While a degree of difference in scores obtained between modes was apparent at C2 level, the differences were not found to be statistically significant. The paper concludes that whether Speaking Tests are delivered in online proctored mode or in traditional face-to-face mode, test takers receive similar scores. The study confirms that mode of test delivery does not significantly affect test taker scores.

## 1. Introduction

Since the late 2010s, and more recently due in considerable part to the covid-19 pandemic, many examinations have moved from face-to-face to online delivery. The current study was conducted in order to determine the extent to which mode of delivery might affect performance and in turn, therefore, affect Speaking test scores. Focusing on English language Speaking tests at CEFR levels B2 to C2, this paper examines the comparability of scores achieved by test takers taking examinations administered in traditional face-to-face mode (TM) with those administered by online proctored mode (OLP).

The paper first reviews approaches to the increasingly-common online delivery of learning and teaching. This is followed by a review of the less common online delivery of examinations. A brief consideration of the assessment of speaking and the challenges of conducting communicative speaking tests is then provided. The paper then examines studies which have compared the two modes of delivery.

Following the background, data of a large sample of test takers taking English language Speaking Tests at CEFR levels B1 to C2 via TM and OLP is then presented and analysed for statistical difference.

## 2. Background

This section presents a background to the online delivery of learning and teaching, especially in the face of the Covid-19 pandemic. Issues in the delivery of online assessment – the benefits and drawbacks to taking tests in OLP mode – are then examined. A brief exploration of the assessment of speaking, and the particularly difficult challenges associated with assessing spoken communicative skills is provided. This is followed by a discussion of the increasingly vexed issue of the online assessment of speaking.

### 2.1 Online Delivery of Teaching and Assessment

In the face of the Covid pandemic, the common practice of learning and teaching being conducted by a teacher at the front of an actual class has undergone immense and rapid change (Hodges et al., 2020). Augmented by developments in technology, the acceptance of online learning has grown exponentially over the past two years (Lim & Wang, 2016), with the 'traditional' mode of delivery being rethought (Hodges et al., 2020). Todd (2020), for example, outlines how Covid was a strong mover in the adoption of online teaching.

Nonetheless, while the mindset has changed in terms of teaching content being delivered online, examinations continue to be viewed as an activity which occurs in a more traditional face-to-face situation (Coniam et al., 2021). There has been take-up of technology in the area of assessment, but rather less than has been the case with online teaching (Gardner, 2020; Mays, 2021).

Assessment – and high-stakes assessment in particular outside certain public school systems where online testing is common – is generally viewed as something to be conducted in pen-and-paper mode, in front of an examiner/invigilator, in a physical test centre. While online learning technologies have permitted relatively effective delivery of learning and teaching, the delivery of assessment in online mode has seen a mixture of advantages, problems and challenges: e.g., a reduction in cheating, connectivity issues etc (Sarrayrih & Ilyas, 2013, Berrada et al., 2021).

Khan & Jawaid (2020), reporting on online assessment in Pakistan during the Covid pandemic, discuss how learning, teaching and assessment in particular need to be equally embraced in terms of access and delivery, stressing the need for attitudinal changes in the online delivery of assessment where both administrators and test-takers lose their fear of newly developed technology in economically developing nations.

Garc á-Pe ñalvo et al. (2021), in the context of how Spanish universities responded to the covid pandemic, provide a number of recommendations concerning online assessment. In addition to increased continuous assessment, they also suggest that technologies which support face-to-face teaching – such as teleconferencing – should be used to deliver assessment, in order to develop teacher and student readiness for and confidence in the "new context of online assessment" (p. 87). They stress that any marking schemes must be made known to students before any assessment takes place. Garc á-Pe ñalvo et al. (2021) recommend that specifically designed online assessment methods be developed for the subject or group of students concerned when "complex subjects with a large number of students" (p. 88) are involved.

There are both benefits and drawbacks to taking tests in OLP mode for the test taker and the examining body as noted by Weiner & Henderson (2022). On the positive side, test takers may take an online-proctored exam in the comfort (and safety) of their own home, an important factor in times of a pandemic where movement is restricted or for test takers with a disability who find access to a remote testing centre challenging at the best of times let alone during a pandemic. In addition, the speed of test delivery and issuance of results may represent the benefit of exams taken in an OLP mode.

Online teaching has a rather longer history of accepted practices and expectations than does online

assessment. Online teaching, which now has over a decade's worth of research, stresses collaborative principles, such as discussion, peer support, learning that is tailored to individuals, self-regulated learning, encouraging students to set their own goals, and planning, monitoring and controlling their cognition (Boekaerts & Corno, 2005). In contrast, the online assessment record is shorter. There, expectations of assessment (and in particular high-stakes assessment) remain more traditional and, until relatively recently, have typically been the product of one test taker, working on their own. Furthermore, when it comes to test delivery, traditional views of comparability (and hence reliability), generally require that the same assessment be delivered to all test takers at the same time. However, in an online world, where the traditional approach to large-scale assessment is difficult, such a requirement potentially creates issues around security, honesty and fairness.

Regarding OLP examinations, there has been extensive discussion around security, the "vulnerability" of online tests and academic dishonesty (see Corrigan-Gibbs et al., 2015; Coniam et al., 2021). Such issues are key, especially when examinations are taken in a remote location such as a test taker's home.

Nonetheless, Foster and Layman (2013) describe how levels of security may be put in place which make the online proctoring of examinations viable. Indeed, there have been studies which report how exam security may even be more effective as a result of the technologies associated with monitoring of online examinations rather than in traditional face-to-face settings (Watson & Sottile, 2008; Rose, 2009).

Technical factors may also need some consideration. In their evaluation of OLP examinations, Giller et al. (2021) report a number of problematic issues, such as login failure and other technical issues (pp. 36-37). Such issues are not, however, the focus of the current study.

Despite such concerns, OLP remains a potentially important delivery method going forward. The current study explores the comparability and hence interchangeability of OLP assessment of speaking with traditional methods.

A brief summary of key issues surrounding assessing the speaking skill and assessing the skill remotely will now be provided.

## 2.2 Assessing Speaking

Speaking has long been considered the most complex of the four macro skills to assess. Some 40 years ago, Madsen (1983) outlined some of the reasons why speaking is challenging to assess. Apart from background construct issues such as defining the actual nature of the speaking skill and devising criteria to properly assess speaking in a communicative age, factors such as ability, tone, reasoning etc. as well as the reluctance of some test takers to even speak (p. 147) had to be dealt with.

Luoma (2004) reiterates how speaking is the most difficult language skill to assess reliably. This is especially the case when speaking is assessed by a human assessor in a face-to-face interaction, when assessments can be influenced by a number of factors such as features of spoken language, the test taker's language level, gender, the nature of the interaction, the tasks and topics driving the interactions, as well as the opportunities that the test taker has to demonstrate their ability. (2004: ix-x).

Sujana (2016) echoes many of the above points in their discussion of the complexity of the aspects involved in testing oral proficiency, noting that many teachers almost avoid assessing speaking.

## 2.3 Assessing Speaking Online

Assessing speaking involves various 'complications', as mentioned above. To overcome some of these complexities, various educators and researchers have recommended moving the assessment of speaking to an online mode, which, they argue, affords advantages over a face-to-face mode. Fall et

al. (2007), for example, describe a Simulated Oral Proficiency Interview (SOPI) which renders large-scale assessment of test takers speaking proficiency on the ACTFL Oral Proficiency Scale comparatively easy to administer and rate. However, the process is entirely machine mediated.

Against the backdrop of the covid pandemic, assessment of all forms moved, with differing degrees of success (Ali & Dmour, 2021), to various online modes. As might be expected – following the discussion above of the complexities of assessing speaking – it was indeed assessing students' oral proficiency that emerged as most challenging for many educators. Forrester (2020) elaborates upon the challenges of assessing speaking online in the time of the covid pandemic. These issues apply to all forms of assessing oral proficiency, not just in formal examinations.

## 3. Comparability of Results from Exams Taken via OLP/TM

There has been considerable research into assessment conducted online with and without invigilation, although few studies have directly compared high-stakes tests conducted in OLP versus those conducted in traditional centre-based face-to-face mode. The following section briefly examines the research into these two related, if different, areas.

### 3.1 Examinations Conducted with and without Invigilation

Much of the research conducted on different modes of invigilation has been in higher education settings. Outside higher education and in the field of organisational psychology, Tippins (2015) discusses how new technology has led to "changes in the assumptions made about good testing practices" and the need "to confront new problems that are created by technological enhancements." She also provides examples of how technology is being used in assessments in realistic ways. In general, studies have reported, perhaps unsurprisingly, that students who sat tests without any invigilation – remote or otherwise – recorded higher grades than students who sat remote invigilated tests: Alessio et al., 2017; Goedl & Malla, 2020; Reisenwitz, 2020.

There have, however, been studies which reported no significant differences in the performance of students sitting tests with or without invigilation: Castillo & Doe, 2017; Lee, 2020.

### 3.2 Examinations Conducted using Online Invigilation / in Traditional Centre-based Face-to-face Mode

Despite the increase in high-stakes assessments conducted online following the 2020-2022 covid pandemic, as Weiner & Henderson (2022) observe, there has been little research into comparability of high-stakes test scores obtained from remotely-invigilated tests as opposed to tests invigilated face to face in testing centres. A summary of the limited amount of research in the area is presented below.

Weiner & Hurtz (2017) examined test taker performance in the context of licensing examinations in the USA, exploring the extent to which performance was equivalent regarding test takers sitting examinations in specially prepared computer-equipped 'kiosks' to test takers sitting the same examinations in physical test centres with human invigilators. No significant differences were found between performance in either proctoring mode. Hurtz & Wiener (2022) extended the scope of the above study following extended closures over the covid pandemic. Their study reported no differences in test score due to proctoring mode.

Wuthisatian (2020) examined differences in performance between test takers taking high-stakes economics examinations using remote online proctoring versus those taken in traditional exam centres. Results suggested that test takers performed differently across the two proctoring methods: those who sat the examination at a centre obtained significantly higher scores than those test takers who were proctored online.

Cherry et al. (2021) examined professional licensure examinations in the USA, comparing outcomes for tests administered either using remote online proctoring or in test centres. While statistically significant differences were observed in results obtained between the two modes, no detectable pattern was observed in favour of either mode.

Morin et al. (2022) investigated a high-stakes national medical licensing examination in Canada taken via remote online proctoring or in exam centres. Despite some test takers reporting different examination experiences, Morin et al., report that test scores across the two proctoring modes – despite there being different examination question types – were broadly comparable.

Muckle et al.'s (2022) study explored scores on a study of North American pharmacy licensing examinations taken via the two proctoring modes following the covid pandemic. Muckle et al. reported higher score for examinations taken onsite by examinees. While they attribute some of the differences in results to the makeup of the sample, further research is clearly called for. Research conducted to gauge test taker reactions to LanguageCert's OLP delivery of tests (Coniam et al., 2021; Coniam, 2022) has thus far been generally positive – broadly echoing the results reported by Muckle et al. (2022) in their study

## 4. The Study

The data in the current study are drawn from LanguageCert's International ESOL (IESOL) suite of Speaking tests administered between 2019-2021, with each test in the suite aligned to a CEFR level. The LanguageCert Speaking qualifications involve a comprehensive test of spoken English, with the tasks in the examinations designed to test the use of English in real-life situations. The qualifications are suitable for non-native speakers of English worldwide; young people or adults attending an English course either in the UK or overseas; students learning English as part of their school or college curriculum; people applying to come to the UK for work purposes.

All Speaking tests comprise four tasks – of increasing complexity as test takers move through the test, and last from 12 minutes for the B1 examination to 17 minutes for the C2 examination. There are four rating scales, each of which has four score levels. The Speaking tests are conducted with a live interlocutor (whether face to face or via remote proctoring), with all examinations recorded for later grading and for use in possible appeals. All Speaking tests are scored against four rating scales. The maximum score is 50 with grades awarded being: Fail below 50%, Pass for scores of 50%-74% and High Pass for scores of 75% and above. See https://www.languagecert.org/en/language-exams/english/languagecert-international-esol.

All examinations are assessed by a closed group of markers at LanguageCert, who are regularly standardised through training to ensure consistency and objectivity of assessment that is benchmarked against the CEFR (see Papargyris & Yan, 2022). A number of different test forms are available for each level of test with new test forms continually being added to the test pool.

To enhance security, not only are different test forms used randomly, but the four task types which comprise a test form are also randomised.

Table 1 below presents the number of test forms available for the 2018-2022 tests that were delivered, and the test taker sample for the analysis presented in the current study.

Table 1: Sample size

| CEFR Level | Test Taker Sample Size | Different Test Forms |
|---|---|---|
| B1 | 19,745 | 30 |
| B2 | 21,154 | 30 |
| C1 | 7,943 | 29 |
| C2 | 3,438 | 19 |

LanguageCert operates OLP internationally, with tests delivered in over 70 countries throughout

the world. Consequently, all aspects by which OLP is conducted – logging on, security checks, connections and voice quality checks etc – are administered through the medium of English. In the face of potential English language constraints for lower-level proficiency test takers, the administration of tests in the IESOL suite by OLP principally takes place from B1 upwards. The dataset below for Speaking is therefore presented only for CEFR levels B1 to C2.

In terms of test reliability, since Speaking Test scores are obtained via the four rating scales, reliability cannot be estimated via item- or rater-based estimation methods. It is, however, possible to estimate reliability by uni-dimensional factor analysis calculating McDonald's omega via the raw totals obtained for the four macroskills, i.e., Reading, Listening, Writing and Speaking, together with the CEFR grade awarded. Table 2 presents the reliability estimates, including 95% confidence interval (CI) lower and upper bounds. (For brevity's sake, results are only reported for the Speaking Test.)

Table 2: Reliability estimates via McDonald's omega

| CEFR Level | Omega | 95% CI lower and upper bounds |
|---|---|---|
| B1 | 0.64 | 0.64-0.65 |
| B2 | 0.62 | 0.62-0.63 |
| C1 | 0.65 | 0.64-0.66 |
| C2 | 0.72 | 0.71-0.74 |

McDonald's omega estimates may be interpreted in a similar manner to the Cronbach alpha, with 0.6 being acceptable. Table 3 below reports the McDonald's omega factor loadings for the Speaking Test.

Table 3: Single-factor model standardized loadings

| CEFR Level | Factor | Standardized Loadings |
|---|---|---|
| B1 | Grade | 0.90 |
| | Speaking test | 0.96 |
| | | |
| B2 | Grade | 0.91 |
| | Speaking test | 0.96 |
| | | |
| C1 | Grade | 0.91 |
| | Speaking test | 0.97 |
| | | |
| C2 | Grade | 0.92 |
| | Speaking test | 0.96 |

As can be seen, loadings for Speaking tests and grades awarded at all CEFR levels are 0.90 and above, indicating that the Speaking tests exhibit a high degree of reliability.

Two sets of data are now presented below. One, descriptive statistics: means, standard deviations and effect size differences; two equivalence independent samples t-tests ("equivalence tests").

The equivalence independent samples t-test permit users to test the null hypothesis that the population means of two independent groups fall inside a user-defined interval, i.e., the equivalence region. The procedure of using two-one-sided tests (TOST) permits significance to be observed via specified upper and lower bounds, as opposed to standard t-tests which report a single t score. As Lakens (2017) puts it:

> *Adopting equivalence tests will prevent the common misinterpretations of nonsignificant* p *values as the absence of an effect and nudge researchers toward specifying which effects they find worthwhile (p. 360)*

The upper and lower bounds represent the extent of variation of t values regarding the two

populations of the two samples being tested. If the t value of the equivalence test is within the estimated range, the two populations may be deemed to be equivalent.

### 4.1 Hypotheses

The overarching hypothesis in the current study is that mean scores obtained between the two modes of test delivery – OLP and TM – will not be significantly different. Specifically, the following two hypotheses are pursued:

(1) That, at worst, only small effect size differences between the two modes will be observed.

(2) That on equivalence tests, significance will not emerge against specified upper and lower bounds for any given CEFR level.

### 4.2 Descriptive Statistics

Table 4 presents a summary of the effect size differences between the sets of means for the Speaking Test total score (maximum 50) for each mode using Cohen's d. Cohen's d indicates standardised differences between two means, sharpening comparisons between two means. In general, a small effect is taken as 0.2, a medium effect as 0.5, and a large effect as 0.8 (Glen, 2021).

Table 4: Effect size differences between mode means

| Level | Mode | Number | Mean | Score Difference | SD | Cohen's d |
|---|---|---|---|---|---|---|
| B1 | TM | 17998 | 37.52 | +1.04 (2.08%) | 8.56 | 0.07 |
| | OLP | 1747 | 38.56 | | 9.88 | |
| B2 | TM | 11046 | 37.82 | -0.58 (1.16%) | 8.1 | 0.06 |
| | OLP | 10108 | 37.24 | | 9.38 | |
| C1 | TM | 2284 | 35.18 | +0.14 (0.28%) | 8.92 | 0.01 |
| | OLP | 5659 | 35.32 | | 9.44 | |
| C2 | TM | 1234 | 31.18 | +4.12 (8.24%) | 8.3 | 0.45 |
| | OLP | 2204 | 35.30 | | 9.92 | |

As can be seen from Table 4, effect sizes are negligible for levels, B1 to C1. It is only at C2 level where the score difference between the two modes is greater than 5%, and where there is a notable small-to-medium effect size difference of 0.45.

### 4.3 Equivalence Tests

Tables 5 to 8 below present equivalence test results comparing OLP and TM.

Upper and lower bounds have been set at +/- 0.05 (i.e., the 95% interval) of the raw score (see Lakens, 2017). These bounds may be construed as representing 95% confidence intervals; however, as TOST consists of two one-sided tests, it makes more precise sense to refer to the upper and lower ends of the confidence intervals. The critical decision on equivalence, as stated earlier, is whether the estimated t value (labelled T-Test in the tables below) is between the upper and lower bound. The p values for the t values (Upper bound, T-Test and Lower bound) indicate significant T-Test values where these go beyond the specified bounds.

Table 5: B1 Equivalence test results

| Statistic | t | df | p |
|---|---|---|---|
| Upper bound | -5.26 | 19743 | < .001 |
| T-Test | -4.80 | 19743 | < .001 |
| Lower bound | -4.34 | 19743 | 1.00 |

Table 6: B2 Equivalence test results

| Statistic | t | df | p |
|---|---|---|---|
| Upper bound | 3.97 | 21152 | 1.00 |
| T-Test | 4.80 | 21152 | < .001 |
| Lower bound | 5.63 | 21152 | < .001 |

Table 7: C1 Equivalence test results

| Statistic | t | df | p |
|---|---|---|---|
| Upper bound | -1.07 | 7941 | 0.14 |
| T-Test | -0.63 | 7941 | 0.53 |
| Lower bound | -0.20 | 7941 | 0.58 |

Table 8: C2 Equivalence test results

| Statistic | t | df | p |
|---|---|---|---|
| Upper bound | -12.78 | 3436 | < .001 |
| T-Test | -12.48 | 3436 | < .001 |
| Lower bound | -12.18 | 3436 | 1.00 |

At none of the four levels was significance observed at both lower and upper bounds. This indicates that although there is not a perfect match, the two modes of Speaking Test administration can be considered broadly equivalent for all the CEFR levels in the study. That said, there would appear to be an issue with the C2 level test, where more investigation is clearly called for.

## 5. Discussion and Conclusion

This study has explored the comparability of scores obtained by test takers of LanguageCert's IESOL English language Speaking Tests at CEFR levels B1 to C2 via traditional face-to-face mode (TM) versus online proctored mode (OLP).

The key hypothesis in the study was that mean scores and hence performance obtained in the OLP and TM modes of test delivery would not be significantly different. Specifically, two hypotheses were being investigated.

The first hypothesis was that, at worst, only small effect size differences between the two modes would be observed. While negligible effect sizes were observed for levels B1 to C1, the fact that a small-to-medium effect size was observed for C2 meant that the hypothesis could not be accepted.

The second hypothesis was that, on equivalence tests, significance would not emerge against specified upper and lower bounds for any given CEFR level. As significance was not observed for both bounds in any of the test levels, it was determined that the two modes of test administration may be considered equivalent broadly for the four CEFR levels examined, and the hypothesis was accepted. Nevertheless, at the highest level of ability (CEFR level C2), test takers scored considerably higher in online proctored mode than in face-to-face mode.

There are two possible reasons for such a discrepancy. One relates to the actual makeup of the C2 test taker cohort. C2 level test takers tend to be professionals in their 30s and 40s, whereas at the lower levels, many test takers are younger school children who are more accustomed to traditional face-to-face centre-based assessments. In this light, C2 test takers are also more comfortable with extensive use of technology, a fact which may account for them being perhaps more at ease in the online proctored environment. The second issue is possibly that of malpractice. In this regard, however, stringent security checks to guard against issues such as impersonation are conducted before Speaking Tests take place. Speaking Test materials are, as mentioned, randomised to forestall possible pre-arranged sets of answers. Further, the Speaking Test is an oral performance test

conducted in real time, which makes cheating much more difficult to do from a test taker's point of view.

To conclude, it would appear that results obtained from taking LanguageCert IESOL Speaking Tests at the lower CEFR levels indicate that similar results are obtained irrespective of whether tests are taken in traditional face-to-face mode or in online proctored mode. Nonetheless, the fact that C2 test takers score higher does require further investigations at this level.

One limitation of the current study is that only one skill has been investigated – speaking. The skill of speaking is generally viewed as the most difficult to administer and assess, with difficulties in online delivery exacerbated rather more than with the more 'static' (in the sense that they do not require direct interaction with an interlocutor) skills of listening, reading and writing. A follow-up study analysing the other skills – listening, reading and writing – is underway.

## References


*[1] Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020).The difference between emergency remote teaching and online learning. EDUCAUSE Review.*

*[2] Lim, C. P., & Wang, L. (Eds.). (2016). Blended learning for quality higher education: Selected case studies on implementation from Asia-Pacific. Bangkok: UNESCO Bangkok Office.*

*[3] Todd, R. W. (2020). Teachers' perceptions of the shift from the classroom to online teaching. International Journal of TESOL Studies, 2(2), 4-16.*

*[4] Coniam, D., Lampropoulou, L., & Cheilari, A. (2021). Online proctoring of high-stakes examinations: A survey of past test takers' attitudes and perceptions. English Language Teaching, 14(8), 58-72. doi.org/10.5539/elt.v14n8p58.*

*[5] Gardner, L. (2020). Covid-19 has forced higher ed to pivot to online learning. Here are 7 takeaways so far. The Chronicle of Higher Education, 20(5).*

*[6] Mays, T. J. (2021). Teaching the teachers. In Radical Solutions for Education in a Crisis Context (pp. 163-176). Springer, Singapore. doi.org/10.1007/978-981-15-7869-4_11.*

*[7] Sarrayrih, M. A., & Ilyas, M. (2013). Challenges of online exam, performances and problems for online university exam. International Journal of Computer Science Issues, 10(1), 439.*

*[8] Berrada, K., Ahmad, H. A. S., Margoum, S., EL Kharki, K., Machwate, S., Bendaoud, R., & Burgos, D. (2021). From the paper textbook to the online screen: A smart strategy to survive as an online learner. In Radical Solutions for Education in a Crisis Context (pp. 191-205). Singapore: Springer.*

*[9] Khan, R. A., & Jawaid, M. (2020). Technology enhanced assessment (TEA) in COVID 19 pandemic. Pakistan Journal of Medical Sciences, 36(19), 108-110. doi.org/10.12669/pjms.36.COVID19-S4.2795.*

*[10] García-Peñalvo, F. J., Corell, A., Abella-García, V., & Grande-de-Prado, M. (2021). Recommendations for mandatory online assessment in higher education during the COVID-19 pandemic. In Radical solutions for education in a crisis context (pp. 85-98). Singapore: Springer.*

*[11] Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. Applied Psychology, 54(2), 199-231.*

*[12] Corrigan-Gibbs, H., Gupta, N., Northcutt, C., Cutrell, E., & Thies, W. (2015). Deterring cheating in online environments. ACM Transactions on Computer-Human Interaction, 22(6), 1-23. doi.org/10.1145/2810239.*

*[13] Foster, D., & Layman, H. (2013). Online proctoring systems compared. Webinar. http://www. slideshare.net/ caveonweb/caveon-webinar-series-online-proctoring-best-practicesoct-2013-slideshare-final.*

*[14] Watson, G., & Sottile, J. (2010). Cheating in the digital Age: Do students cheat more in on-line courses? Online Journal of Distance Learning Administration, 13(1).*

*[15] Rose, C. (2009). Virtual proctoring in distance education: An open-source solution. American Journal of Business Education, 2(2), 81-88. doi.org/10.19030/ajbe.v2i2.4039.*

*[16] Giller, P. (2021). E-proctoring in theory and practice: a review. Dublin, Ireland: Quality and Qualifications Ireland.*

*[17] Madsen, H. S. (1983). Techniques in testing. New York: Oxford University Press.*

*[18] Luoma, S. (2004). Assessing speaking. Cambridge University Press.*

*[19] Sujana, I. M. (2016). Assessing oral proficiency: Problems and suggestions for elicitation techniques. https://academia.edu.*

*[20] Fall, T., Adair-Hauck, B., & Glisan, E. (2007). Assessing students' oral proficiency: A case for online testing. Foreign Language Annals, 40(3), 377-406. doi.org/10.1111/j.1944-9720.2007.tb02865.x.*

*[21] Ali, L., & Dmour, N. A. H. H. A. (2021). The shift to online assessment due to COVID-19: An empirical study of university students, behaviour and performance, in the region of UAE. International Journal of Information and*

*Education Technology, 11(5), 220-228. doi.org/10.18178/ijiet. 2021.11.5.1515.*
*[22] Forrester, A. (2020). Addressing the challenges of group speaking assessments in the time of the Coronavirus. International Journal of TESOL Studies, 2(2), 74-88.*
*[23] Tippins, N. T. (2015). Technology and assessment in selection. Annual Review of Organizational Psychology and Organizational Behavior, 2, 551–582. https://doi.org/10.1146/annurev-orgpsych-031413-091317.*
*[24] Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. Online Learning, 21(1), 146-161. doi.org/10.24059/olj.v21i1.885.*
*[25] Goedl, P. A., & Malla, G. B. (2020). A study of grade equivalency between proctored and unproctored exams in distance education. American Journal of Distance Education, 34(4), 280-289. doi.org/10.1080/08923647.2020.1796376.*
*[26] Reisenwitz, T. H. (2020). Examining the necessity of proctoring online exams. Journal of Higher Education Theory and Practice, 20(1), 118-124. doi.org/10.33423/jhetp.v20i1.2782.*
*[27] Castillo, M. S., & Doe, R. (2017). Mobile and nonmobile assessment in organizations: Does proctoring make a difference? Psychology, 8(06), 878. doi.org/10.4236/psych.2017.86057.*
*[28] Lee, J. W. (2020). Impact of proctoring environments on student performance: Online vs offline proctored exams. The Journal of Asian Finance, Economics, and Business, 7(8), 653-660. doi.org/10.13106/jafeb.2020.vol7.no8.653.*
*[29] Weiner, J. A., & Henderson, D. (2022). Online remote proctored delivery of high stakes tests: Issues and research. Journal of Applied Testing Technology, 23, 1-4.*
*[30] 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.*
*[31] Hurtz, G. M., & Weiner, J. A. (2022). Comparability and integrity of online remote vs. onsite proctored credentialing exams. Journal of Applied Testing Technology, 23, 36-45.*
*[32] Wuthisatian, R. (2020). Student exam performance in different proctored environments: Evidence from an online economics course. International Review of Economics Education, 35, 100196. doi.org/10.1016/j.iree.2020.100196.*
*[33] Cherry, G., O'Leary, M., Naumenko, O., Kuan, L. A., & Waters, L. (2021). Do outcomes from high stakes examinations taken in test centres and via live remote proctoring differ? Computers and Education Open, 2, 100061. doi.org/10.1016/j.caeo.2021.100061.*
*[34] Morin, M., Alves, C., & De Champlain, A. (2021). The show must go on: Lessons learned from using remote proctoring in a high-stakes medical licensing exam program in response to severe disruption. Journal of Applied Testing Technology, 23, 15-35.*
*[35] Muckle, T. J., Meng, Y., & Johnson, S. (2022). A quantitative evaluation of a live remote proctoring pilot. Journal of Applied Testing Technology, 23, 46-53.*
*[36] Coniam, D. (2022). Online invigilation of English language examinations: A survey of past China test takers' attitudes and perceptions. International Journal of TESOL Studies, 4(1), 21-31. doi.org/10.46451/ijts.2022.01.03.*
*[37] Papargyris, Y., & Yan, Z. (2022). Examiner quality and consistency across LanguageCert Writing Tests. International Journal of TESOL Studies, 4(1), 203-212. doi.org/10.46451/ijts.2022.01.13.*
*[38] Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. Communication Methods and Measures, 14(1), 1-24.*
*[39] Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science, 8(4), 355-362. doi.org/10.1177/1948550617697177.*
*[40] Glen, S. (2021). Cohen's D: Definition, examples, formulas. https://www.statisticshowto.com/.*